

The Next Turn

Recommendations, Textual Anthropomorphic Cues, and Subsequent Self-Disclosure in Mental-Health Conversations with ChatGPT



Recommendations, Textual Anthropomorphic Cues, and Self-Disclosure in Mental-Health Conversations with ChatGPT: A Turn-Level Analysis Across Conversation Positions

Master thesis submitted by

Martin Han Carel Meyer

To Delft University of Technology
In partial fulfilment of the requirements for the degree of

Master of Science

Management of Technology

Student Number: 4817087

To be defended in public on 4 November 2026

Graduation Committee

First Supervisor: **Savvas Zannettou**, Assistant Professor,
TPM, TU Delft

Chair and Second Supervisor: **Aaron Ding**, Senior Associate Professor,
TPM, TU Delft

Faculty: Technology, Policy and Management



Acknowledgements

This thesis marks the completion of my MSc in Management of Technology at Delft University of Technology. The process has been challenging, but also an important learning experience. It gave me the opportunity to work on a topic that combines my interest in emerging technologies with questions about privacy, responsibility, and human interaction with conversational AI. Throughout the project, I learned not only how to conduct an empirical research project, but also how difficult it can be to translate research choices, analyses, and findings into a clear and coherent academic argument.

I would like to thank Dr. Savvas Zannettou for his guidance during the research process. Our discussions helped me develop the research design and make important decisions as the study progressed. I would also like to thank Dr. Aaron Ding for his feedback during the later stages of the thesis, which contributed to strengthening the final work, and for serving as chair of my graduation committee.

I am especially grateful to my friends who repeatedly took the time to read drafts of this thesis and provide thoughtful and critical feedback, helping me to critically evaluate methodological choices and their limitations. Several of them also shared their experience in academic research and thesis writing, which helped me improve the structure, clarity, and academic presentation of the thesis. Their comments often challenged me to reconsider how I explained my reasoning, methodological choices, findings, and contribution. These discussions helped me see where the writing was unclear, where assumptions needed to be made explicit, and where the argument could be strengthened. Their involvement during the final writing, revision, and fine-tuning stages made a meaningful contribution to how this research is presented in its final form.

Finally, I would like to thank my family and the people close to me for their patience, encouragement, and support throughout the thesis process. Completing this thesis required persistence over a long period, and their support helped me continue through the more demanding stages of the project. I am grateful to everyone who contributed to this process, whether academically, critically, or personally.

Martin Meyer
Delft, October 2026

Executive Summary

Context and Research Aim

Generative conversational AI is increasingly used for personal and mental-health-related conversations in which users may share sensitive experiences, thoughts, and feelings (Karnam et al., 2026; Siddals et al., 2024). In these conversations, self-disclosure occurs as part of an ongoing message-by-message exchange between the user and ChatGPT. However, existing research provides limited insight into how self-disclosure unfolds at the level of individual exchanges in naturally occurring conversations, both in relation to characteristics of the preceding ChatGPT response and to where the exchange occurs within the conversation (Greilich et al., 2025; Karnam et al., 2026; Li et al., 2023; Skjuve et al., 2023). This thesis therefore examines how ChatGPT recommendations, human-like wording, and where an exchange occurs within a conversation relate to subsequent user self-disclosure in mental-health conversations.

Approach

To investigate these message-to-message relationships, the study analysed 69 donated English-language mental-health conversations comprising 2,639 ChatGPT-user exchanges. It focused on two characteristics of ChatGPT responses, recommendations and human-like wording, and examined how they related to self-disclosure in the user response that immediately followed. The exchanges were qualitatively coded to identify these characteristics and subsequent self-disclosure. The resulting codes were then analysed quantitatively, while also considering where each exchange occurred within the conversation.

Key Findings

Recommendations were not consistently related to subsequent self-disclosure, including when users continued in the direction suggested by the recommendation. Human-like wording showed the same pattern, with no consistent relationship to subsequent self-disclosure. The clearest pattern concerned where an exchange occurred within the conversation: the sensitivity and amount of personal information varied across conversation position, with several of the observed differences concentrated in the earlier part of conversations.

Implications

For product teams, the findings suggest evaluating recommendations, human-like wording, and self-disclosure as separate aspects of the interaction. When a conversational feature is used for a particular purpose, teams should evaluate whether it supports that purpose while assessing the personal information users disclose separately. Because users tended to share more, and more sensitive, information near the beginning of conversations, privacy checks should pay particular attention to these early exchanges. This helps teams assess whether a conversational feature is useful while also considering whether the amount and sensitivity of the information users share match what is actually needed for the purpose of the interaction.

Table of Contents

Acknowledgements	A
Executive Summary	B
Table of Contents	i
List of Figures	v
List of Appendix Figures	vi
List of Tables	vii
List of Appendix Tables	viii
List of Appendix Equations	ix
List of Abbreviations	x
Glossary	xi
1 Introduction	1
1.1 Background and Context	1
1.2 Research Problem	1
1.2.1 Research Gap	2
1.3 Research Significance	2
1.3.1 Academic Significance	3
1.3.2 Practical Relevance	3
1.3.3 Management of Technology Relevance	3
1.4 Objective and Research Questions	4
1.5 Scope	4
1.6 Structural Outline	5
2 Literature Review	6
2.1 Mental-Health Conversations with ChatGPT as a Disclosure-Sensitive Context	6
2.2 Self-Disclosure and Its Dimensions	7
2.2.1 Conceptualising Self-Disclosure	7
2.2.2 Depth and Breadth	7
2.2.3 Disclosure Amount (Volume)	9
2.2.4 Self-Disclosure Evidence in Conversational AI	9
2.3 Recommendations and Conversational Steering	10
2.3.1 Recommendations within Conversational Steering	10
2.3.2 Conditions Shaping Responses to Recommendations	10
2.3.3 Recommendation Presence, Conversational Uptake, and Self-Disclosure	11
2.4 Textual Anthropomorphic Cues and Self-Disclosure	12
2.4.1 Textual Cues and Perceived Anthropomorphism	12
2.4.2 Evidence Linking Anthropomorphic Cues to Self-Disclosure	12
2.4.3 Interpreting Mixed Evidence: Social Connection and Privacy Concerns	13
2.5 Conversation Position and Disclosure Dynamics	14
2.5.1 Self-Disclosure Across Time and Conversation Position	14
2.5.2 Conversation Position and the Relevance of Conversational Features	15
2.6 Integrated Synthesis, Research Gap, and Conceptual Framework	16
2.6.1 Integrated Synthesis	16
2.6.2 Remaining Research Gap	17
2.6.3 Conceptual Framework	17
3 Methodology	19
3.1 Research Design	19

3.2 Data Source and Access.....	21
3.3 Construction of the Sampling Frame	21
3.3.1 Data Restructuring	22
3.3.2 Language Eligibility.....	22
3.4 Sampling Strategy	22
3.4.1 Conversation-Length Stratification.....	23
3.4.2 Candidate Selection and Screening	24
3.4.3 Candidate Eligibility Screening	26
3.4.4 Final Sample	27
3.5 Construction of the <AI, user> Turn Pair	27
3.6 Qualitative Coding	29
3.6.1 Code Creation	29
3.6.2 Coding Sequence of an <AI, user> Turn	34
3.6.3 Coding Consistency and Quality Assurance	35
3.7 Quantitative Analysis	36
3.7.1 Descriptive Statistics	36
3.7.2 Exploratory Analysis of Self-Disclosure Dimensions	37
3.7.3 RQ1: Recommendation Presence and Observable Uptake Analysis	37
3.7.4 RQ2: Textual Anthropomorphic Cue Analysis	39
3.7.5 RQ3: Conversation Position and Cue-Context Analysis	40
3.7.6 Statistical Procedures and Reporting Conventions	42
3.8 Research Quality	43
3.9 Ethical Considerations	45
4 Results	46
4.1 Descriptive Overview and Exploratory Analysis	46
4.1.1 Self-Disclosure Outcomes	46
4.1.2 Recommendations and Observable Uptake.....	48
4.1.3 Textual Anthropomorphic Cues	49
4.1.4 Breadth and Volume Across Depth Levels.....	49
4.2 RQ1: Recommendation Presence, Observable Uptake, and Subsequent Self-Disclosure.....	51
4.2.1 Primary Three-Group Comparison	52
4.2.2 Pairwise Group Comparisons	53
4.2.3 Adjusted Analysis	55
4.2.4 Descriptive Differences Across Recommendation Types.....	56
4.3 RQ2: Textual Anthropomorphic Cues and Subsequent Self-Disclosure	57
4.3.1 Primary Two-Group Comparison	57
4.3.2 Adjusted Analysis	59
4.3.3 Descriptive Variation Across Textual Anthropomorphic Cue Types	60
4.4 RQ3: Conversation Position, Cue Context, and Self-Disclosure	60
4.4.1 Self-Disclosure Across Conversation Quartiles	61
4.4.2 Conversation-Length Sensitivity Analysis.....	63
4.4.3 Self-Disclosure by Cue Group and Conversation Quartile	64
4.4.4 Conversation Quartile × Cue-Group Interaction	69
4.4.5 Conversation-Level Cluster Analysis	69
5 Interpretation of the Findings	71
5.1 RQ1: Recommendation Presence, Observable Uptake, and Subsequent User Self-Disclosure.....	71
5.2 RQ2: Textual Anthropomorphic Cues and Subsequent User Self-Disclosure	72
5.3 RQ3: Conversation Position, Cue Context, and Subsequent User Self-Disclosure....	72
6 Discussion	74

6.1 Academic Contributions to Self-Disclosure in Conversational AI	74
6.2 Practical Implications	75
6.3 Limitations and Generalisability	77
6.3.1 Observational Design and Secondary Data	78
6.3.2 Dataset and Sampling	79
6.3.3 Measurement and Coding	80
6.3.4 Unit of Analysis and Conversation Position	81
6.3.5 Statistical Uncertainty	82
6.3.6 Literature Review Coverage and Reproducibility	83
6.4 Future Research	83
7 Conclusion	86
7.1 Answers to the Research Questions	86
7.1.1 RQ1: Recommendation Presence and Observable Uptake	86
7.1.2 RQ2: Textual Anthropomorphic Cues	86
7.1.3 RQ3: Conversation Position and Cue Context	87
7.2 Answer to the Main Research Question	87
8 Bibliography	89
Appendix A: Literature-Search and Review Process	92
A.1 Review Approach and Scope	92
A.2 Search Strategy	93
A.3 Search Development and Supplementary Searching	94
A.4 Source Evaluation and Selection	94
A.5 Literature Documentation and Synthesis	96
Appendix B: Codebook	98
B.1 Recommendation Coding	98
B.2 Textual Anthropomorphic Cue Coding	103
B.3 User Self-Disclosure Coding	107
B.4 Coding Procedure	120
B.4.1 Unit of Analysis and Coding Context	120
B.4.2 Coding Sequence and Field Conventions	120
B.4.3 Coding the ChatGPT Message	121
B.4.4 Coding the User Response	121
B.4.5 Contextual and Ambiguous Cases	122
B.4.6 Coding Completion and Consistency Checks	122
Appendix C: Statistical Implementation Details	123
C.1 Statistical Procedures by Analysis	123
C.2 Outcome Representations, Rank Handling, and Effect Sizes	124
C.3 Multiple-Testing Corrections	126
C.4 RQ1 and RQ2 Statistical Implementation	127
C.4.1 Primary and Pairwise Comparisons	127
C.4.2 Adjusted Regression Models	128
C.4.3 Conversation-Level Bootstrap	129
C.5 RQ3 Statistical Implementation	129
C.5.1 Permutation-Based Group Comparisons	130
C.5.2 Conversation Quartile × Cue-Group Interaction Analysis	130
C.5.3 Conversation-Level Clustered Analysis	132
C.6 Resampling and Reproducibility Settings	133
C.7 Computational Environment and Reporting Conventions	134
Appendix D: Complete Results Tables and Figures	136
D.1 Cumulative Distribution of Conversation Lengths in the Sampling Frame	136

D.2 Breadth Type Frequencies	137
D.3 Pairwise Comparisons Across Depth Levels	139
D.4 Pairwise Comparisons Between Recommendation Groups	139
D.5 Supplementary Recommendation-Type Analysis	141
D.6 Descriptive Textual Anthropomorphic-Cue Analysis	145
D.7 Pairwise Comparisons Between Conversation Quartiles	149
D.8 Conversation-Length Sensitivity Analysis	151
D.9 Self-Disclosure Depth Distributions by Cue Group and Conversation Quartile	157
D.10 Complete Secondary Cue-Group Comparisons	158
Appendix E: Use of AI and Other Digital Tools	160
E.1 Overview of Generative AI and Digital Tool Use	160
E.2 Writing and Revision Support	162
E.3 Literature and Conceptual Support	163
E.4 Coding and Operationalisation Support	164
E.5 Programming and Statistical Analysis Support	164
E.6 Figures, Tables, and Visual Materials	165
E.7 Data Handling, Verification, and Researcher Responsibility	165
E.8 Reflection on AI Use	166

List of Figures

Figure 1	Breadth and Depth of Self-Disclosure in Social Penetration Theory	8
Figure 2	Conceptual Framework of ChatGPT Conversational Features, Conversation Position, and Subsequent User Self-Disclosure	18
Figure 3	Overview of the Research Design and Analytical Workflow	20
Figure 4	Overview of the Sampling Strategy	23
Figure 5	Distribution of English-Language Conversations Across Conversation- Length Strata	24
Figure 6	Distribution of Conversations Across Conversation-Length Strata in the Final Sample	27
Figure 7	Construction of the <AI, user> Turn Pair	28
Figure 8	Overview of the Final Coding Scheme and Operationalisation	30
Figure 9	Coding Sequence for an <AI, user> Turn Pair	35
Figure 10	Analysis Roadmap for RQ1: Recommendation Presence and Observable Uptake	38
Figure 11	Analysis Roadmap for RQ2: Textual Anthropomorphic Cues	39
Figure 12	Analysis Roadmap for RQ3: Conversation Position and Cue Context	41
Figure 13	Distribution of Self-Disclosure Depth Levels Across User Responses	46
Figure 14	Distribution of Breadth Amount Across User Responses	47
Figure 15	Breadth Types Occurring in at Least 5% of User Responses	48
Figure 16	Distribution of Recommendation Groups Across Analysed Turn Pairs	49
Figure 17	Presence of Textual Anthropomorphic Cues Across ChatGPT Messages .	49
Figure 18	Breadth Amount Across Depth Levels	50
Figure 19	Volume Across Depth Levels	50
Figure 20	Distribution of Self-Disclosure Depth Levels by Recommendation Group	53
Figure 21	Distribution of Depth Levels by Textual Anthropomorphic-Cue Presence	58
Figure 22	Depth Distribution Across Conversation Quartiles	62
Figure 23	Cue-Group Distribution Across Conversation Quartiles	65
Figure 24	Depth Distribution Across Conversation Quartiles by Cue Group	66

List of Appendix Figures

Figure D1	Cumulative Distribution of Conversation Lengths in the Sampling Frame	137
Figure D2	Complete Distribution of Coded Breadth Types Across User Responses	138
Figure D3	Prevalence of Recommendation Types Across ChatGPT Messages	141
Figure D4	Frequency of Recommendation Types Across ChatGPT Messages	142
Figure D5	Depth Distribution by Recommendation Type and Observable Uptake	143
Figure D6	Mean Breadth Amount by Recommendation Type and Observable Uptake	144
Figure D7	Mean Volume by Recommendation Type and Observable Uptake	145
Figure D8	Frequency of Textual Anthropomorphic-Cue Types Across ChatGPT Messages	146
Figure D9	Depth Distribution by Textual Anthropomorphic-Cue Type	147
Figure D10	Mean Breadth Amount by Textual Anthropomorphic-Cue Type	148
Figure D11	Mean Volume by Textual Anthropomorphic-Cue Type	149

List of Tables

Table 1	Candidate Selection and Retention by Conversation-Length Stratum	25
Table 2	Eligibility Criteria for Candidate Conversations	26
Table 3	Overview of Statistical Procedures Across Analyses	43
Table 4	Breadth Amount and Volume Across Depth Levels	51
Table 5	Breadth Amount and Volume by Recommendation Group	52
Table 6	Overall Differences Across Recommendation Groups	53
Table 7	Pairwise Comparisons of Self-Disclosure Across Recommendation Groups	54
Table 8	Adjusted Recommendation-Group Contrasts	55
Table 9	Breadth Amount and Volume by Textual Anthropomorphic-Cue Presence	57
Table 10	Primary Two-Group Comparisons by Textual Anthropomorphic-Cue Presence	58
Table 11	Adjusted Associations of Textual Anthropomorphic-Cue Presence	59
Table 12	Breadth Amount and Volume by Conversation Quartile	61
Table 13	Overall Differences Across Conversation Quartiles	62
Table 14	Statistically Supported Pairwise Comparisons Between Conversation Quartiles	63
Table 15	Conversation-Length Sensitivity Analysis of Quartile Differences	64
Table 16	Mean Breadth Amount and Volume Across Conversation Quartiles by Cue Group	67
Table 17	Statistically Supported Cue-Context Comparisons Across Quartiles and Cue Groups	68
Table 18	Conversation Quartile × Cue-Group Interaction Tests	69
Table 19	Conversation-Level Clustered Analysis of Conversation Position and Cue-Group Interactions	70
Table 20	Overview of Study Limitations and Interpretation Boundaries	77

List of Appendix Tables

Table A1	Main Search Concepts and Representative Search Terms.....	93
Table A2	Criteria for Source Evaluation, Selection, and Evidential Weighting	95
Table B1	Recommendation Variables and Coding Rules.....	99
Table B2	Recommendation-Type Decision Rules	100
Table B3	Recommendation Counting and Uptake Rules.....	102
Table B4	Textual Anthropomorphism Variables and Coding Rules	104
Table B5	Textual Anthropomorphic Cue-Type Decision Rules	105
Table B6	Counting Rules for Textual Anthropomorphic Cues	107
Table B7	User Self-Disclosure Variables and Coding Rules	108
Table B8	Self-Disclosure Depth Decision Rules	110
Table B9	Self-Disclosure Breadth-Type Decision Rules	112
Table B10	Self-Disclosure Breadth Counting Rules	116
Table B11	Self-Disclosure Volume Decision Rules	118
Table C1	Statistical Procedures Used Across the Analyses	123
Table C2	Resampling Settings Used in the Statistical Analyses	133
Table C3	Principal Software and Python Packages Used in the Computational Analysis	134
Table D1	Pairwise Comparisons of Breadth Amount and Volume Across Depth Levels	139
Table D2	Complete Pairwise Comparisons Between Recommendation Groups ...	140
Table D3	Complete Pairwise Comparisons Between Conversation Quartiles	150
Table D4	Breadth Amount and Volume by Conversation-Length Stratum and Conversation Quartile	151
Table D5	Depth Distributions by Conversation-Length Stratum and Conversation Quartile	152
Table D6	Overall Conversation-Quartile Tests Within Conversation-Length Strata	154
Table D7	Complete Pairwise Conversation-Quartile Comparisons for Supported Conversation-Length Stratum × Outcome Combinations	155
Table D8	Complete Depth Distributions by Cue Group and Conversation Quartile	157
Table D9	Conversation-Quartile Differences Within Cue Groups.....	158
Table D10	Cue-Group Differences Within Conversation Quartiles	159
Table E1	Overview of Generative AI and Digital Tool Use	161

List of Appendix Equations

Equation C1	Log1p Transformation of Self-Disclosure Volume	125
Equation C2	Rank-Biserial Correlation for Mann–Whitney U Comparisons	125
Equation C3	Cramér’s V for Categorical Comparisons	126
Equation C5	Additional Variance Explained (ΔR^2)	126
Equation C6	Adjusted Direct Contrast Between Recommendation Groups	128
Equation C7	Empirical Upper-Tail Bootstrap Probability	129
Equation C8	Empirical Lower-Tail Bootstrap Probability	129
Equation C9	Two-Sided Bootstrap p-Value	129
Equation C10	Two-Sided Mann–Whitney U Permutation Statistic	130
Equation C11	Monte Carlo Permutation p-Value With +1 Correction	130
Equation C12	Freedman-Lane Permuted Outcome.....	131
Equation C13	Freedman-Lane Model-Comparison Statistic	131
Equation C14	Freedman-Lane Monte Carlo Permutation p-Value.....	131
Equation C15	Cluster-Robust Finite-Sample Covariance Correction.....	132
Equation C16	Cluster-Based Inferential Degrees of Freedom	133

List of Abbreviations

Abbreviation	Full term
AI	artificial intelligence
APA	American Psychological Association
CI	confidence interval
CSV	comma-separated values
GDPR	General Data Protection Regulation
GIMP	GNU Image Manipulation Program
GPT	generative pre-trained transformer
HREC	Human Research Ethics Committee
ID	identifier
IQR	interquartile range
JSON	JavaScript Object Notation
LLM	large language model
MOT	Management of Technology
MSc	Master of Science
NA	not applicable
RQ	research question
SSH	Secure Shell
TPM	Technology, Policy and Management

Glossary

Term	Definition	Page(s)
Anthropomorphism	The attribution of human characteristics to a non-human entity (Greilich et al., 2025; Li et al., 2023).	12–13, 31–32, 103–106
Breadth amount	The number of distinct personal content areas included in a disclosure.	33–34, 36–42, 46–70, 108–121, 123–134, 138–158
Breadth type	A category describing the type of personal content included in a disclosure.	33–34, 36, 47–48, 108–116, 136–137
Conversation length	The number of user-AI turns that make up a conversation.	22–27, 41, 63–64, 82, 135–136, 150–155
Conversation position	The relative location of an exchange between the beginning and end of a conversation.	14–18, 40–42, 60–70, 73–88, 129–132, 148–158
Conversation quartile	One of four intervals representing relative position within a conversation.	40–42, 61–70, 129–132, 148–158
Conversation-length stratum	A category used to group conversations according to their length.	22–27, 41, 63–64, 150–155
Conversational AI	Artificial-intelligence systems that interact with users through natural-language conversation (Papneja & Yadav, 2025).	1, 6–9
Conversational steering	The process through which one conversational participant influences how an interaction proceeds (Karnam et al., 2026).	10–11, 15–18, 30–31
Cue group	A grouping used to distinguish different combinations of conversational cues.	42, 64–70, 130–132, 156–158
Disclosure-sensitive context	An interaction context in which users may reveal intimate, emotionally sensitive, or private information (Kelly et al., 2025; Meadi et al., 2025).	6, 16
Generative AI	Artificial intelligence capable of generating new content in response to user input (Siddals et al., 2024).	1, 6

Term	Definition	Page(s)
Mental-health conversation	A conversation concerning mental health, emotional well-being, psychological difficulties, or related support (Karnam et al., 2026; Siddals et al., 2024).	1, 6, 19–28
Observable recommendation uptake	An observable response indicating that a conversational participant (the user) continues in the direction suggested by a preceding recommendation (from ChatGPT).	17–18, 29–31, 37–38, 51–56, 71, 98–102, 138–143
Original turn (<user, AI>)	A conversational exchange consisting of a user message followed by the AI system's response.	22, 27–29
Perceived anthropomorphism	The extent to which a user perceives a system as possessing human-like characteristics (Li et al., 2023).	12–13, 31
Psychological reactance	A motivational state of resistance that can arise when individuals perceive their freedom or autonomy as threatened (Fitzsimons & Lehmann, 2004; Kim & Wang, 2025).	11, 13
Recommendation	Advice or a suggestion that proposes, favours, or discourages a possible course of action (Feng, 2009; Fitzsimons & Lehmann, 2004).	10–11, 15–18, 30–31, 37–42, 51–56, 71–72, 98–102, 138–143
Recommendation group	A grouping that distinguishes conversational exchanges according to recommendation presence and observable uptake.	37–38, 52–56, 71, 138–143
Recommendation presence	Whether a conversational message contains at least one recommendation.	11, 17–18, 31, 37–42, 51–59, 65–72, 98–102, 138–143, 156–158
Self-disclosure	The communication of personal information about oneself (Omarzu, 2000; Papneja & Yadav, 2025).	1, 6–18, 29–89, 107–122
Self-disclosure breadth	The range of distinct personal content areas covered by a disclosure (Carpenter & Greene, 2015; Omarzu, 2000).	7–9, 33–44, 46–89, 108–116, 138–158

Term	Definition	Page(s)
Self-disclosure depth	The degree of intimacy or sensitivity of disclosed information (Carpenter & Greene, 2015; Omarzu, 2000).	7–9, 32–44, 46–89, 109–110, 138–158
Self-disclosure volume	The amount of self-disclosing content communicated (Harmsen et al., 2023; Omarzu, 2000).	9, 33–44, 46–89, 117–119, 138–158
Social Penetration Theory	A theory of relationship development in which self-disclosure is characterised by breadth and depth (Carpenter & Greene, 2015).	8, 109–110
Social presence	The perception that another social actor is present and socially accessible within an interaction (Li et al., 2023).	12–14, 18
Textual anthropomorphic cue	An observable linguistic feature through which a non-human system is presented as possessing human-like characteristics (Greilich et al., 2025; Li et al., 2023).	12–18, 31–32, 39–42, 57–60, 72–75, 103–106, 144–147, 156–158
Textual anthropomorphic-cue presence	Whether a message contains at least one textual anthropomorphic cue.	31–32, 39–42, 57–59, 72, 103–106, 144–147, 156–158
Turn pair (<AI, user>)	A conversational exchange consisting of an AI-system message followed by the user response that immediately follows it.	27–29, 34–45, 120–134

1 Introduction

This chapter introduces the context, research problem, and research questions that guide this thesis.

1.1 Background and Context

Generative artificial intelligence has become increasingly embedded in everyday digital interaction. Conversational AI systems such as ChatGPT, which generate responses using large language models (LLMs), are used not only for information seeking and practical assistance, but also for personal and socially oriented purposes, including mental-health support (Siddals et al., 2024). Karnam et al. (2026) show that users increasingly engage with ChatGPT across diverse domains, including sensitive areas such as health and mental health. Compared with conversations using GPT-3.5, those using GPT-4o showed a 33% increase in Health topics and a 19% increase in Mental Health topics. At the same time, users increasingly engaged with ChatGPT in social roles such as a companion, while personal-data disclosure became both more common and more diverse.

This development is particularly relevant in mental-health conversations, which are defined in this thesis as interactions in which users discuss emotional or personal difficulties such as anxiety, depression, stress, grief, relationship problems, or coping with difficult experiences (Karnam et al., 2026; Siddals et al., 2024). In interviews with users of generative-AI chatbots, participants described using these systems to process difficult emotions and gain perspective on personal and relationship problems (Siddals et al., 2024). This type of interaction is also available without the real-time involvement of a human professional and is therefore not subject to the same patient waiting periods as human care (Meadi et al., 2025).

Using conversational AI for personal and emotional concerns can involve sharing experiences, circumstances, thoughts, and feelings with the system (Siddals et al., 2024). This communication of personal information is referred to as self-disclosure (Omarzu, 2000). In conversational-AI interactions, self-disclosure is therefore an important part of how users communicate personal context to an AI system (Papneja & Yadav, 2025).

1.2 Research Problem

Self-disclosure in conversational AI does not occur in isolation: each user response forms part of an ongoing exchange with the system. Previous research indicates that characteristics of conversational-AI communication can be associated with users' self-disclosure. However, the available evidence is methodologically diverse, and many studies rely on experimental, hypothetical, self-reported, or otherwise structured interactions (Harmsen et al., 2023; Kelly et al., 2025; Y. C. Lee et al., 2020; Li et al., 2023; Papneja & Yadav, 2025). Consequently, these findings provide only

limited insight into how self-disclosure occurs within naturally occurring exchanges between users and ChatGPT.

Research based on donated ChatGPT histories has begun to address real-world interaction by examining broader patterns in conversational purpose, framing, disclosure, and steering over time (Karnam et al., 2026). However, such broader patterns provide different evidence from an analysis of individual ChatGPT-user exchanges. They provide less insight into whether observable characteristics of a particular ChatGPT message are associated with the self-disclosure contained in the user response that immediately follows. The research problem addressed in this thesis is therefore the limited empirical basis for understanding self-disclosure at the level of individual exchanges within naturally occurring mental-health conversations with ChatGPT.

1.2.1 Research Gap

Within the message-to-message interaction identified in the research problem, direct evidence remains limited in two areas. The first concerns characteristics of the ChatGPT message that immediately precedes a user response. This thesis focuses on recommendations and textual anthropomorphic cues because both are observable features of ChatGPT communication that have been examined in related conversational-AI research (Karnam et al., 2026; Li et al., 2023). However, existing research provides limited direct evidence on whether these features are associated with the self-disclosure contained in the immediately following user response in naturally occurring ChatGPT conversations (Greilich et al., 2025; Karnam et al., 2026; Li et al., 2023).

To examine this association meaningfully, self-disclosure must be distinguished across different dimensions. Depth concerns the intimacy or sensitivity of that information, whereas breadth concerns the range of personal areas disclosed (Carpenter & Greene, 2015; Omarzu, 2000). Disclosure can also differ in amount, which Omarzu (2000) describes in terms of the amount or persistence of disclosure; this dimension is referred to as volume in this thesis. These dimensions capture different properties of self-disclosure and need not change together, which makes it important to examine them separately (Omarzu, 2000; Skjuve et al., 2023).

The second area concerns where an exchange occurs within the conversation. Longitudinal research shows that breadth and depth can change across repeated interactions as human-chatbot relationships develop (Skjuve et al., 2023). However, this research examines changes across interactions occurring over an extended period, rather than differences between earlier and later exchanges within a single conversation (Skjuve et al., 2023). Direct evidence therefore remains limited on whether depth, breadth, and volume differ across positions within a naturally occurring ChatGPT conversation, or whether the associations between message characteristics and subsequent self-disclosure differ across those positions.

Taken together, the literature provides limited understanding of how self-disclosure unfolds at the level of individual message-to-message exchanges in naturally occurring ChatGPT conversations. In particular, less is known about how disclosure relates to characteristics of the preceding ChatGPT message and to where the exchange occurs within the conversation.

1.3 Research Significance

The significance of this study is considered from three complementary perspectives: academic, practical, and Management of Technology. These perspectives address the study's contribution to the literature, its relevance for conversational-AI development and evaluation, and its positioning within the Management of Technology field.

1.3.1 Academic Significance

The academic significance of this study lies in extending existing research on self-disclosure to naturally occurring mental-health conversations with ChatGPT. Previous evidence has largely relied on controlled interactions, disclosure intentions, or broader conversation-level patterns, leaving less insight into self-disclosure as it occurs in actual message exchanges (Karnam et al., 2026; Papneja & Yadav, 2025). By examining observable characteristics of ChatGPT messages together with multiple dimensions of subsequent self-disclosure and conversation position, this study adds evidence at the level of individual exchanges. This provides more detailed insight into how self-disclosure varies in relation to the preceding ChatGPT message and the position of the exchange within the conversation.

1.3.2 Practical Relevance

This more detailed understanding also has practical relevance for developers and providers of conversational-AI systems used in mental-health-related interactions. These conversations can involve highly sensitive personal information, while privacy and user autonomy are recurring concerns in conversational AI for mental health (Meadi et al., 2025). Evidence about how specific conversational features are associated with subsequent disclosure can therefore help developers evaluate design choices in privacy-sensitive interactions. Together, these findings provide a stronger empirical basis for assessing how such conversational features are associated with self-disclosure in mental-health-related interactions.

1.3.3 Management of Technology Relevance

From a Management of Technology perspective, this thesis treats conversational AI as a corporate technological resource whose interaction characteristics create organisational choices concerning product design, evaluation, and governance. The study connects observable properties of the technology with user behaviour and the organisational implications of deploying conversational-AI services. This positioning is reflected in the integration of human-computer interaction and communication

perspectives for examining user-system interaction, self-disclosure theory for conceptualising user responses, and privacy research for interpreting disclosure-sensitive outcomes. These perspectives are applied across the research design, operationalisation, and interpretation of the findings rather than being treated as separate theoretical lenses. In doing so, the study brings together technological, behavioural, and organisational perspectives within a management-relevant problem concerning the development and responsible deployment of conversational AI.

1.4 Objective and Research Questions

Building on the identified research problem and gap, this study examines self-disclosure at the message-to-message level in naturally occurring mental-health conversations. It focuses on how ChatGPT recommendations, textual anthropomorphic cues, and conversation position relate to the depth, breadth, and volume of self-disclosure in the immediately following user response.

Based on this objective, the following main research question guides the study:

How do ChatGPT recommendations, textual anthropomorphic cues, and conversation position relate to the depth, breadth, and volume of subsequent user self-disclosure in naturally occurring mental-health conversations?

The main research question is addressed through three sub-questions:

RQ1: How do the depth, breadth, and volume of user self-disclosure differ between responses following no recommendation, a recommendation without observable uptake, and a recommendation with observable uptake?

RQ2: How do the depth, breadth, and volume of user self-disclosure differ between responses following ChatGPT messages with and without textual anthropomorphic cues?

RQ3: How does conversation position relate to the depth, breadth, and volume of user self-disclosure across different recommendation and textual-anthropomorphic-cue contexts?

1.5 Scope

This thesis examines naturally occurring, English-language mental-health conversations with ChatGPT, using information observable in the conversation text. The analysis examines associations between two characteristics of ChatGPT messages and the self-disclosure in the user response that immediately follows, while also considering where these exchanges occur within the conversation.

The study does not assess the clinical effectiveness of ChatGPT, the quality or appropriateness of its recommendations, or whether greater self-disclosure is

beneficial or harmful. These outcomes fall outside the study because the analysis examines observable associations in conversation text rather than clinical outcomes, recommendation accuracy, or the consequences of disclosure. User perceptions and underlying psychological mechanisms, such as perceived anthropomorphism, trust, social presence, or reactance, are also outside the scope because they cannot be directly determined from the available conversation text.

1.6 Structural Outline

The remainder of this thesis is organised as follows. Chapter 2 develops the theoretical background, reviews the relevant literature, and integrates the reviewed concepts and evidence into the conceptual framework guiding the study. Chapter 3 describes and justifies the research methodology, including the data, sampling, coding, analytical approach, and ethical considerations. Chapter 4 presents the empirical results for the three research questions. Chapter 5 interprets these findings in relation to the theoretical expectations and prior empirical evidence. Chapter 6 brings the findings together to discuss their broader academic contributions, practical implications, limitations and generalisability, and directions for future research. Chapter 7 concludes the thesis by answering the three research questions and the main research question. Chapter 8 provides the bibliography.

2 Literature Review

This chapter reviews the theoretical and empirical literature on self-disclosure in mental-health conversations with ChatGPT, recommendations, textual anthropomorphic cues, and conversation position. Self-disclosure is examined through its breadth, depth, and volume dimensions, after which the literature on recommendations and anthropomorphic cues is synthesized in relation to disclosure. Finally, conversation position is considered as the temporal (change over time) context of these relationships, and the reviewed evidence is integrated to establish the research gap and conceptual framework guiding the study. The literature-search process is described in Appendix A.

2.1 Mental-Health Conversations with ChatGPT as a Disclosure-Sensitive Context

Mental-health conversations with ChatGPT provide a relevant setting for studying self-disclosure because such conversations occur in actual use, rather than only in hypothetical or experimental settings. User interviews document people turning to generative-AI chatbots for mental-health concerns, while donated ChatGPT histories provide evidence of these conversations in practice (Karnam et al., 2026; Siddals et al., 2024). Together, these findings establish mental-health conversations with generative AI as an observed interaction context in which personal disclosure can occur.

What makes this setting particularly relevant to self-disclosure is the type of information users may share. Mental-health conversations can involve intimate, emotionally sensitive, or otherwise private information. Privacy and confidentiality are consequently prominent concerns in the literature on conversational AI for mental health (Meady et al., 2025). Therefore, the importance of disclosure in this context lies not only in whether users reveal personal information, but also in the sensitivity of what they reveal.

Sensitivity also appears to matter for people's willingness to disclose. In hypothetical health scenarios, participants reported greater willingness to disclose less sensitive than more sensitive information and greater willingness to disclose in physical-health than mental-health situations (Kelly et al., 2025). This suggests that both the sensitivity of the information and the health context are related to disclosure willingness. However, willingness is not the same as actual disclosure behaviour, and these findings do not show whether the same pattern occurs during naturally occurring ChatGPT conversations (Kelly et al., 2025).

Evidence from donated ChatGPT conversations shows that users disclose personal information during actual interactions with ChatGPT (Karnam et al., 2026). Karnam et al. (2026) found that such disclosure became more common and diverse over the period covered by their data. This creates an important distinction in the literature: hypothetical research shows caution around sensitive disclosure, while donated

ChatGPT conversations show that personal disclosure also occurs in actual use (Karnam et al., 2026; Kelly et al., 2025). What remains less clear is how these disclosures differ in their range, sensitivity, and amount within individual exchanges. Understanding those differences first requires a clearer account of what self-disclosure consists of and how its different dimensions are understood in the literature.

2.2 Self-Disclosure and Its Dimensions

The previous section showed that users disclose personal information in mental-health conversations with ChatGPT. Such disclosure, however, can differ in its range, intimacy, and amount. The literature therefore distinguishes between breadth, depth, and amount or duration as separate properties of self-disclosure (Carpenter & Greene, 2015; Omarzu, 2000). These dimensions capture different aspects of disclosure and should not be treated as a single measure of “more” or “less” self-disclosure (Omarzu, 2000). The following sections examine these distinctions and how self-disclosure has been studied in conversational-AI research (Papneja & Yadav, 2025).

2.2.1 Conceptualising Self-Disclosure

Self-disclosure broadly refers to communicating personal information about oneself to another party. This information can concern personal facts and experiences as well as thoughts, opinions, and emotions (Omarzu, 2000). The literature does not define the construct identically. Some definitions focus broadly on the communication of personal information, whereas others additionally emphasise that disclosure is voluntary or deliberate (Harmsen et al., 2023; Omarzu, 2000; Papneja & Yadav, 2025). Self-disclosure therefore has a common conceptual core, but its precise boundaries differ across studies.

Definitions also differ in the type of personal information they emphasise. Broad definitions include ordinary personal facts, experiences, opinions, and emotions, whereas privacy- and health-oriented research often concentrates more specifically on sensitive or identifiable information (Kelly et al., 2025; Omarzu, 2000). Studies using the same term self-disclosure may therefore still examine different parts of the construct.

Self-disclosure is also studied in different ways. Some studies analyse information that users actually communicate, whereas others measure willingness or intentions to disclose (Kelly et al., 2025; Papneja & Yadav, 2025). These are different, because willingness does not show what a user actually reveals during a conversation (Kelly et al., 2025). When disclosure is observed directly, it can still differ in its range, intimacy, and amount, which leads to the dimensions discussed next.

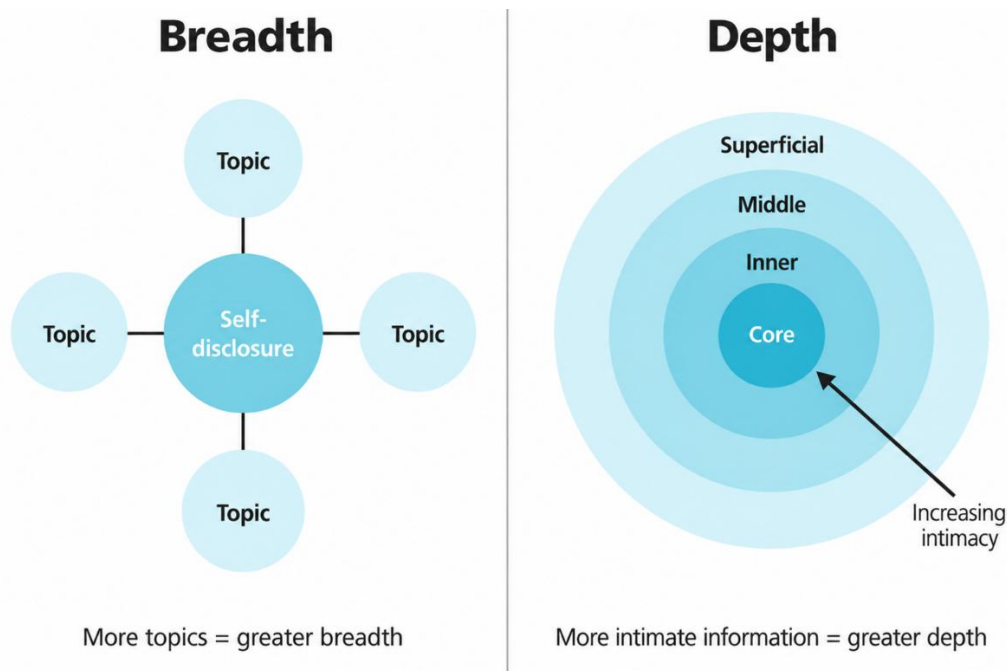
2.2.2 Depth and Breadth

Self-disclosure can differ in both depth and breadth. Depth refers to how intimate or sensitive the disclosed information is, whereas breadth refers to the range of

personal areas a person discusses. These dimensions can vary independently: a person may discuss several personal topics without revealing highly intimate information, while another may disclose deeply personal information about a single issue (Carpenter & Greene, 2015; Omarzu, 2000). Figure 1 illustrates this distinction.

Figure 1

Breadth and Depth of Self-Disclosure in Social Penetration Theory



Note. Author's conceptual illustration of breadth and depth based on Social Penetration Theory (Carpenter & Greene, 2015). The figure does not represent the coding procedure used in this study.

Social Penetration Theory illustrates disclosure depth through the onion model, in which personal information ranges from relatively superficial outer layers to increasingly private inner and core layers (Carpenter & Greene, 2015). Depth therefore concerns the intimacy of the disclosed information rather than simply the topic being discussed.

The distinction between depth and breadth is central to Social Penetration Theory, which describes interpersonal disclosure in terms of the range and intimacy of information exchanged as relationships develop (Carpenter & Greene, 2015). Although the theory was developed for human relationships, depth and breadth have also been applied to self-disclosure in human-chatbot interactions (Skjuve et al., 2023). Research with social chatbots shows, however, that disclosure does not necessarily develop in the same way as expected in human relationships. Depth and breadth are therefore useful for describing disclosure without assuming that human-chatbot relationships follow the same developmental process (Skjuve et al., 2023).

This distinction is particularly important in human-chatbot interactions. Seemingly ordinary topics, such as everyday activities, can still be experienced by users as personal or intimate, meaning that topic content alone provides only an approximation of disclosure depth (Skjuve et al., 2023).

2.2.3 Disclosure Amount (Volume)

Depth and breadth describe the range and intimacy of self-disclosure, but disclosure can also differ in the amount of personal information communicated. Omarzu (2000) labels this dimension *duration*, referring to the sheer amount or persistence of disclosure, which may be represented by the time spent disclosing or the number of words contained in disclosing statements. A disclosure can therefore contain more personal content without necessarily becoming broader or more intimate (Omarzu, 2000). Because this thesis focuses on the amount of self-disclosing text rather than the time spent disclosing, this dimension is referred to as *volume*. This term more directly reflects the amount of self-disclosing content expressed in words.

An important distinction is that general response length is not equivalent to disclosure volume. Harmsen et al. (2023), for example, found that disclosure-containing voicebot questions elicited longer responses and more linguistic markers of reciprocity. However, the number of manually coded self-disclosure statements did not significantly increase. Their findings show that longer responses may contain more text without necessarily containing more self-disclosing content (Harmsen et al., 2023). Volume therefore captures the amount of self-disclosing content communicated, whereas depth captures the level of intimacy and breadth captures the range of disclosed areas (Omarzu, 2000). This distinction is particularly relevant when comparing conversational-AI studies because different studies operationalise self-disclosure in different ways.

2.2.4 Self-Disclosure Evidence in Conversational AI

Research on self-disclosure to conversational AI is conducted across substantially different research settings. These include structured voicebot interactions, controlled chatbot experiments, hypothetical scenarios, and longer-term relationships with social chatbots (Harmsen et al., 2023; Kelly et al., 2025; Y. C. Lee et al., 2020; Skjuve et al., 2023). These settings differ in how interactions are structured and in the extent to which users determine what they disclose. Because self-disclosure is influenced by both interaction modality and context, findings should therefore be interpreted within the setting in which they were obtained rather than assumed to generalise directly to naturally occurring ChatGPT conversations (Papneja & Yadav, 2025).

The studies also use different indicators of self-disclosure, including willingness to disclose, coded disclosure statements, response length, and self-reported disclosure (Harmsen et al., 2023; Kelly et al., 2025; Papneja & Yadav, 2025). Because these measures capture different aspects of disclosure, a relationship found with one measure does not necessarily apply to depth, breadth, or volume in another setting. This distinction is important when examining how specific features of ChatGPT's responses relate to the user disclosure that follows.

2.3 Recommendations and Conversational Steering

With these dimensions of self-disclosure established, attention can turn to characteristics of the ChatGPT message that precedes the user response. The first concerns recommendations and the broader conversational steering through which a chatbot can direct or invite what happens next in the interaction (Karnam et al., 2026; Xiao et al., 2020).

2.3.1 Recommendations within Conversational Steering

Recommendations are generally understood as advice or suggestions that favour, discourage, or propose a possible course of action. They are therefore different from other supportive moves such as acknowledging emotions or asking questions to better understand a problem (Feng, 2009; Fitzsimons & Lehmann, 2004). Feng (2009), for example, treats emotional support, problem inquiry, and advice as distinct parts of a supportive interaction. A recommendation thus does more than respond to what the user has said: it introduces a possible direction for what the user could do next.

Conversational systems can also guide an interaction without recommending a course of action. Follow-up questions, prompts, and requests for elaboration can direct or invite the user's next response while remaining different from advice in the narrower sense (Karnam et al., 2026; Xiao et al., 2020). This makes it useful to distinguish recommendations from the broader idea of conversational steering. Recommendations direct the user toward a possible action, whereas other steering moves may instead invite the user to continue, clarify, or elaborate on the conversation. The broader operationalisation used in this thesis is described in Chapter 3.

2.3.2 Conditions Shaping Responses to Recommendations

A recommendation is not likely to be received in the same way in every interaction. Its reception can depend on what has happened before the advice is given. Feng (2009) found that advice was evaluated more positively when emotional support preceded it, and when problem inquiry occurred before rather than after the advice. This suggests that timing and conversational sequence matter: advice may be received differently depending on how well it follows the recipient's situation and the preceding interaction.

Recommendations can also be received negatively when they appear unwanted or conflict with the recipient's own direction. Fitzsimons and Lehmann (2004) found that recommendations given without being requested could trigger resistance when they conflicted with recipients' existing preferences, sometimes leading them to choose against the recommendation. Similar concerns appear in generative-AI use, where users have described advice as premature when the chatbot moved toward solutions before they felt sufficiently heard (Siddals et al., 2024). Related mental-health chatbot research also found greater psychological reactance to be associated with lower self-reported disclosure, although that study manipulated emotional support

and chatbot profile rather than recommendations themselves (Kim & Wang, 2025). These findings connect resistance to disclosure, but only indirectly for recommendations.

Taken together, the literature suggests that responses to recommendations are conditional rather than uniform. Advice may fit an interaction when it follows sufficient exploration and support, while poorly timed or preference-conflicting advice may instead meet resistance (Feng, 2009; Fitzsimons & Lehmann, 2004; Siddals et al., 2024). Supportive sequencing and reactance provide possible explanations for these different responses, but neither establishes how recommendations are associated with the depth, breadth, or volume of subsequent self-disclosure in naturally occurring ChatGPT conversations. Understanding that relationship also requires distinguishing the presence of a recommendation from how the user responds to it.

2.3.3 Recommendation Presence, Conversational Uptake, and Self-Disclosure

The presence of a recommendation and the user's conversational response to it are separate parts of the interaction. Karnam et al. (2026) analysed donated records of actual ChatGPT conversations. They distinguish steering according to whether the subsequent interaction entails, contradicts, or remains neutral toward the model-initiated direction. Such conversational uptake is observable in the interaction, but it does not establish whether the user agreed with the suggestion, evaluated it positively, or acted on it outside the conversation (Karnam et al., 2026).

Research using both donated ChatGPT conversations and structured conversational-AI interactions shows that users can follow model-initiated directions within the conversation. Karnam et al. (2026) observe users continuing along model-initiated conversational directions, while Xiao et al. (2020) show that users frequently provide additional information following chatbot prompts and follow-up questions. These findings demonstrate observable conversational responsiveness, but they do not establish that a particular steering move causes the information disclosed in the subsequent response (Karnam et al., 2026; Xiao et al., 2020).

Conversational uptake is also different from self-disclosure. A user may follow the direction proposed by ChatGPT and disclose personal information while doing so, but these behaviours do not necessarily occur together. Existing studies mainly examine whether users follow model-initiated steering or provide additional information, rather than whether the subsequent disclosure becomes deeper, broader, or greater in volume (Karnam et al., 2026; Xiao et al., 2020). It therefore remains unclear how recommendation presence and conversational uptake are associated with the different dimensions of self-disclosure in naturally occurring mental-health conversations with ChatGPT.

2.4 Textual Anthropomorphic Cues and Self-Disclosure

Whereas recommendations concern how ChatGPT may direct the interaction, textual anthropomorphic cues concern observable language features that express human-like social or personal qualities (Greilich et al., 2025; Karnam et al., 2026). This section first distinguishes such observable cues from users' perceived anthropomorphism and then examines how anthropomorphic interaction has been associated with self-disclosure in prior conversational-AI research.

2.4.1 Textual Cues and Perceived Anthropomorphism

Anthropomorphism refers to the attribution of human characteristics to a nonhuman entity (Greilich et al., 2025). In conversational AI, this human-likeness can be considered at two different levels. At the level of the chatbot message, textual anthropomorphic cues are observable linguistic features through which the system presents human-like characteristics. Examples include self-introduction, greetings, and other human-like verbal formulations. At the level of the user, these cues may contribute to perceived anthropomorphism: the extent to which the user perceives the system as possessing human-like qualities (Greilich et al., 2025; Li et al., 2023). The two are therefore related but should not be treated as the same construct. Li et al. (2023) demonstrate this distinction experimentally by manipulating human-like verbal cues and measuring perceived anthropomorphism separately. Participants reported higher perceived anthropomorphism when human-like verbal cues were present (Li et al., 2023). A textual anthropomorphic cue can therefore be identified in a chatbot message without assuming that the user actually perceived ChatGPT as human-like.

Textual cues are only one way in which human-likeness can be introduced into a conversational agent. Prior studies have also examined visual appearance and human-like profiles, sometimes in combination with verbal cues. These features represent different observable forms of anthropomorphic design and should therefore be distinguished when interpreting their effects (Greilich et al., 2025; Li et al., 2023). Studies that combine multiple cue types can inform the broader relationship between anthropomorphic design and self-disclosure. However, such studies cannot isolate the specific relationship between textual anthropomorphic cues and subsequent user self-disclosure.

2.4.2 Evidence Linking Anthropomorphic Cues to Self-Disclosure

The clearest evidence connecting anthropomorphic design with disclosure-related outcomes comes through users' perceptions of the chatbot. Li et al. (2023) found that human-like verbal cues increased perceived anthropomorphism and social presence in a healthcare conversational agent. Social presence refers to the extent to which a user experiences the chatbot as a present social interaction partner, rather than merely as a technical system (Greilich et al., 2025; Li et al., 2023). Greater social presence was associated with stronger trust, while trust was associated with greater

intention to disclose information (Li et al., 2023). The study therefore supports a possible pathway between human-like language and disclosure intention.

Other chatbot research provides related, but less direct, evidence. In an AI psychotherapy-chatbot study, greater social presence was associated with higher self-reported disclosure, although anthropomorphic textual cues were not isolated as the source of that social presence (J. Lee et al., 2024). Kim and Wang (2025) similarly found greater self-reported disclosure when psychological reactance was lower in a GPT-4 mental-health chatbot experiment, but their design combined emotional support with a human-like profile. Together, these studies show that socially oriented chatbot interaction can be related to disclosure while leaving the specific role of textual anthropomorphic cues unresolved.

The available evidence does not yet establish a simple relationship between more anthropomorphic interaction and greater self-disclosure. Studies differ in the cues they examine, the conversational settings in which those cues appear, and whether disclosure is measured through intentions or post-interaction self-report (Kim & Wang, 2025; J. Lee et al., 2024; Li et al., 2023). The specific relationship between textual anthropomorphic cues and self-disclosure therefore remains unclear. Two possible pathways identified in prior research may help explain these different disclosure responses. Human-like interaction may facilitate disclosure through greater social presence and trust (Greilich et al., 2025; Li et al., 2023), while privacy concerns, social-judgment concerns, and AI anxiety may work in the opposite direction (Greilich et al., 2025).

2.4.3 Interpreting Mixed Evidence: Social Connection and Privacy Concerns

Li et al. (2023) provide empirical support for this proposed pathway. Human-like verbal cues increased social presence, greater social presence was associated with stronger trust, and trust predicted a greater intention to disclose information. These findings suggest that social presence and trust may help explain how anthropomorphic interaction relates to disclosure-related intentions (Li et al., 2023). However, social presence and trust are user perceptions rather than direct consequences of textual cue presence. They should therefore be treated as possible explanatory mechanisms rather than assumed outcomes of textual anthropomorphic cues (Li et al., 2023).

The social-connection pathway is not the only possible response to more human-like interaction. Research on anthropomorphic chatbots also identifies privacy concerns, social-judgment concerns, and AI anxiety as factors that can negatively shape responses to human-like design (Greilich et al., 2025). These concerns are particularly relevant in mental-health contexts, where privacy and confidentiality, anthropomorphisation and deception, and user autonomy recur as ethical concerns in conversational-AI research (Meady et al., 2025). This evidence does not show that textual anthropomorphic cues themselves cause privacy harm or reduce autonomy. It does, however, show that greater trust or willingness to disclose should not

automatically be interpreted as indicating that disclosure is less privacy-sensitive or less consequential.

The mixed evidence can therefore be understood as a tension rather than as a single directional effect. Human-like interaction may lower barriers to disclosure when it strengthens social presence or trust, while in other settings it may increase concerns related to privacy, judgment, or the nature of the interaction (Greilich et al., 2025; Li et al., 2023). However, existing studies provide limited direct evidence about how observable textual anthropomorphic cues relate to self-disclosure. This leaves the relationship between textual cue presence and actual next-turn disclosure insufficiently understood.

2.5 Conversation Position and Disclosure Dynamics

The relationships discussed above unfold within conversations that develop over time. Self-disclosure can change as interactions continue (Skjuve et al., 2023), while the way later conversational moves are received can depend on what has occurred beforehand (Feng, 2009). Conversation position may therefore be relevant both to the disclosure users provide and to the relationships between recommendations or anthropomorphic cues and the response that follows. This section therefore examines what existing research establishes about self-disclosure across conversational time. It also considers whether conversation position may provide relevant context for the relationship between conversational features and subsequent self-disclosure.

2.5.1 Self-Disclosure Across Time and Conversation Position

Self-disclosure in human-chatbot interaction does not follow one fixed pattern over time. In longitudinal research with the social chatbot Replika, conversational breadth generally became narrower as relationships developed, while depth could increase, decrease, remain stable, or fluctuate (Skjuve et al., 2023). Lee et al. (2020) similarly found that disclosure changed across repeated chatbot interactions, but that the pattern depended on the conversational task and chatbot condition. Together, these findings show that disclosure can develop in different directions and that its dimensions do not necessarily change together (Y. C. Lee et al., 2020; Skjuve et al., 2023).

These findings concern change across repeated interactions rather than progression within a single conversation. Skjuve et al. (2023) examined relationships developing over several weeks, while Lee et al. (2020) compared interactions across multiple experiment days. A user disclosing differently after weeks of chatbot use is therefore not equivalent to disclosure changing between an earlier and later exchange within the same conversation. The two temporal levels should be kept separate when considering conversation position.

Evidence closer to within-conversation progression remains indirect. Xiao et al. (2020) found that the quality of participants' first responses was strongly related to

the quality of their responses across the conversational survey. Participants also frequently complied with prompts for additional information and responded to follow-up questions. However, the study did not compare self-disclosure systematically across earlier and later conversation positions, and the questions and adaptive prompts changed as the interaction progressed (Xiao et al., 2020). More broadly, reviews of conversational-AI self-disclosure synthesize a wide range of conversational and contextual factors, but do not establish relative position within a single conversation as a well-understood determinant of disclosure (Papneja & Yadav, 2025).

2.5.2 Conversation Position and the Relevance of Conversational Features

Conversation position may also matter because recommendations occur within an interaction that has already developed to some extent. Research on supportive advice shows that advice evaluations can differ depending on the preceding conversational sequence. Feng (2009) found stronger evaluations when emotional support preceded advice and when problem inquiry occurred before rather than after the advice. These findings show that advice is not received independently of the preceding interaction. However, the study concerned hypothetical human conversations and perceived advice quality rather than ChatGPT interactions or subsequent self-disclosure. It therefore supports the relevance of preceding conversational context, but does not establish whether recommendation-disclosure associations differ between earlier and later positions within a ChatGPT conversation.

Comparable evidence for textual anthropomorphic cues is even more limited. Existing anthropomorphism research mainly examines the presence or type of human-like cues and their relationships with user perceptions and responses (Greilich et al., 2025). It provides little evidence on whether the relevance of such cues changes depending on where they occur within a conversation. Research using donated records of actual ChatGPT conversations also shows that anthropomorphic behaviour and conversational steering change over calendar time and across model periods (Karnam et al., 2026). These findings concern changes in ChatGPT use and system behaviour across longer periods. They do not establish whether the same type of cue relates differently to subsequent user disclosure when it occurs earlier or later within an individual conversation.

Taken together, the literature establishes two relevant patterns over time: self-disclosure can change across sustained interaction, and the way conversational features are received can depend on what has occurred beforehand (Feng, 2009; Y. C. Lee et al., 2020; Skjuve et al., 2023). Neither body of evidence, however, establishes how disclosure varies across relative positions within a single ChatGPT conversation. It also does not establish whether the relationships involving recommendations and textual anthropomorphic cues differ across those positions. It therefore remains unclear how the position of an exchange within a conversation relates to self-disclosure and to the relevance of conversational features.

2.6 Integrated Synthesis, Research Gap, and Conceptual Framework

The preceding sections established the disclosure-sensitive context of mental-health conversations with ChatGPT and the multidimensional nature of self-disclosure. They also identified recommendations, textual anthropomorphic cues, and conversation position as relevant features for understanding how disclosure may vary across interactions. This section integrates these findings, identifies the relationships that remain unresolved in the existing literature, and brings them together in the conceptual framework guiding this study.

2.6.1 Integrated Synthesis

Taken together, the reviewed literature suggests that neither recommendations nor anthropomorphic cues have a uniform relationship with self-disclosure. Responses to recommendations appear to depend on how advice fits the preceding interaction and the user's existing preferences or direction (Feng, 2009; Fitzsimons & Lehmann, 2004; Siddals et al., 2024). Anthropomorphic interaction, meanwhile, can be associated with socially supportive perceptions as well as concerns about privacy or the nature of the interaction (Greilich et al., 2025; Li et al., 2023). These findings therefore point to context-dependent responses rather than a uniformly positive or negative pattern. The mechanisms proposed in prior work help interpret why these relationships may differ across contexts, but they do not establish a single direction in which either conversational feature relates to subsequent self-disclosure.

Self-disclosure also needs to be understood as multidimensional and as something that can change over time. Depth, breadth, and volume capture different aspects of disclosure (Omarzu, 2000). Longitudinal chatbot research further shows that depth and breadth can develop differently over time (Skjuve et al., 2023). Evidence about change over time must also be interpreted according to how time is measured. Studies of repeated interactions or developing human-chatbot relationships examine change across multiple interactions, which is different from comparing earlier and later positions within a single conversation (Y. C. Lee et al., 2020; Skjuve et al., 2023).

The existing evidence has also been produced through substantially different research approaches. Studies variously examine hypothetical willingness or disclosure intentions, self-reported disclosure, structured chatbot interactions, repeated interaction over time, and naturally occurring conversation histories (Harmsen et al., 2023; Karnam et al., 2026; Kelly et al., 2025; Li et al., 2023; Skjuve et al., 2023). These approaches operate at different analytical levels, including individual responses, conversational sequences, repeated sessions, developing relationships, and calendar-time trends (Karnam et al., 2026; Y. C. Lee et al., 2020; Skjuve et al., 2023; Xiao et al., 2020). Their findings are therefore complementary but not directly interchangeable.

2.6.2 Remaining Research Gap

The reviewed literature leaves a specific gap in the existing evidence. Existing studies provide limited direct evidence on how recommendations and textual anthropomorphic cues in ChatGPT messages relate to subsequent user self-disclosure. More specifically, it remains unclear whether these features are associated with the depth, breadth, or volume of self-disclosure in the immediately following user response. For recommendations, it remains unclear whether recommendation presence or observable uptake is associated with these disclosure dimensions in actual ChatGPT exchanges (Karnam et al., 2026; Xiao et al., 2020). For anthropomorphic cues, existing research primarily examines human-like design, user perceptions, and disclosure intentions or self-reports rather than the relationship between observable textual cues and multidimensional subsequent disclosure (Greilich et al., 2025; Li et al., 2023).

A further unresolved question concerns where these exchanges occur within the conversation. Existing studies show that disclosure can change across repeated interactions and that preceding conversational sequence can matter (Feng, 2009; Y. C. Lee et al., 2020; Skjuve et al., 2023). However, they do not establish how depth, breadth, and volume vary between earlier and later positions within a single ChatGPT conversation. They also do not establish whether associations involving recommendations and textual anthropomorphic cues differ across those positions. The research gap therefore concerns how these three elements relate to one another: observable features in ChatGPT messages, multidimensional self-disclosure in the immediately following user response, and the position of that exchange within a mental-health conversation. Addressing this gap allows these previously separate aspects of conversational-AI research to be examined within the same analysis.

2.6.3 Conceptual Framework

Figure 2 integrates the constructs derived from the reviewed literature into the conceptual framework guiding this study. The framework is situated within naturally occurring mental-health conversations with ChatGPT and focuses on individual message-response exchanges. Recommendation presence and textual anthropomorphic cue presence are characteristics of the preceding ChatGPT message, while depth, breadth, and volume represent distinct dimensions of self-disclosure in the immediately subsequent user response (Omarzu, 2000). For recommendation-containing exchanges, observable recommendation uptake is distinguished from recommendation presence. Observable uptake is conceptually distinct from recommendation presence and does not imply psychological acceptance or real-world compliance (Karnam et al., 2026).

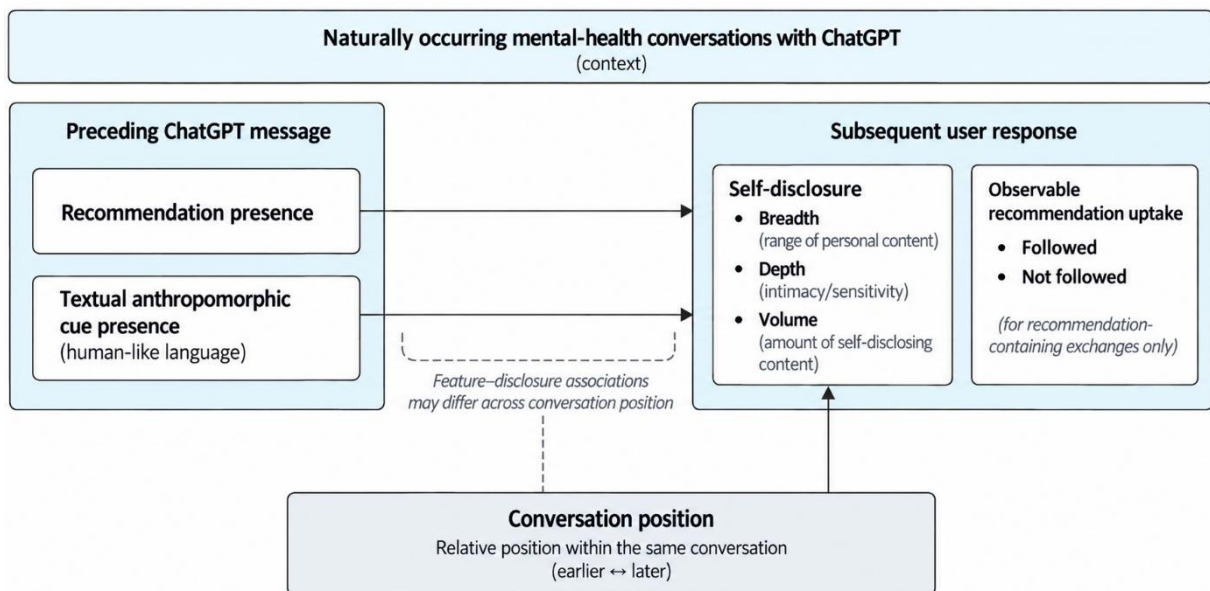
Conversation position indicates where the exchange occurs within the progression of the conversation. The framework allows depth, breadth, and volume to vary across conversation position and allows the associations involving recommendations and textual anthropomorphic cues to differ between earlier and later parts of the conversation. These relationships are represented as unresolved associations

because existing research does not directly establish such effects within individual ChatGPT conversations (Y. C. Lee et al., 2020; Skjuve et al., 2023).

Prior literature provides possible explanations for the relationships represented in the framework. These include supportive sequencing and reactance in response to recommendations and social presence, trust, and privacy-related perceptions in response to anthropomorphic interaction (Feng, 2009; Fitzsimons & Lehmann, 2004; Greilich et al., 2025; Li et al., 2023). These mechanisms are not included as tested pathways because they cannot be observed directly in the conversation data and are not measured in this thesis. Figure 2 therefore represents associations among observable conversational features, subsequent multidimensional self-disclosure, and conversation position. The framework organises the relationships examined by the research questions and provides the conceptual basis for their operationalisation in Chapter 3.

Figure 2

Conceptual Framework of ChatGPT Conversational Features, Conversation Position, and Subsequent User Self-Disclosure



Note. Solid arrows indicate associations examined in this study and do not imply causation. Observable recommendation uptake is coded only for recommendation-containing exchanges and is treated as distinct from psychological acceptance or real-world compliance. The dashed bracket indicates that associations between preceding conversational features and subsequent self-disclosure may differ across conversation position.

3 Methodology

This chapter describes the research design and procedures used to examine associations between user self-disclosure and recommendations, textual anthropomorphic cues, and relative conversation position in mental-health conversations with ChatGPT. It explains the data source and sampling procedure, the construction of the unit of analysis, the qualitative coding process, the quantitative analyses, and the study's quality and ethical considerations.

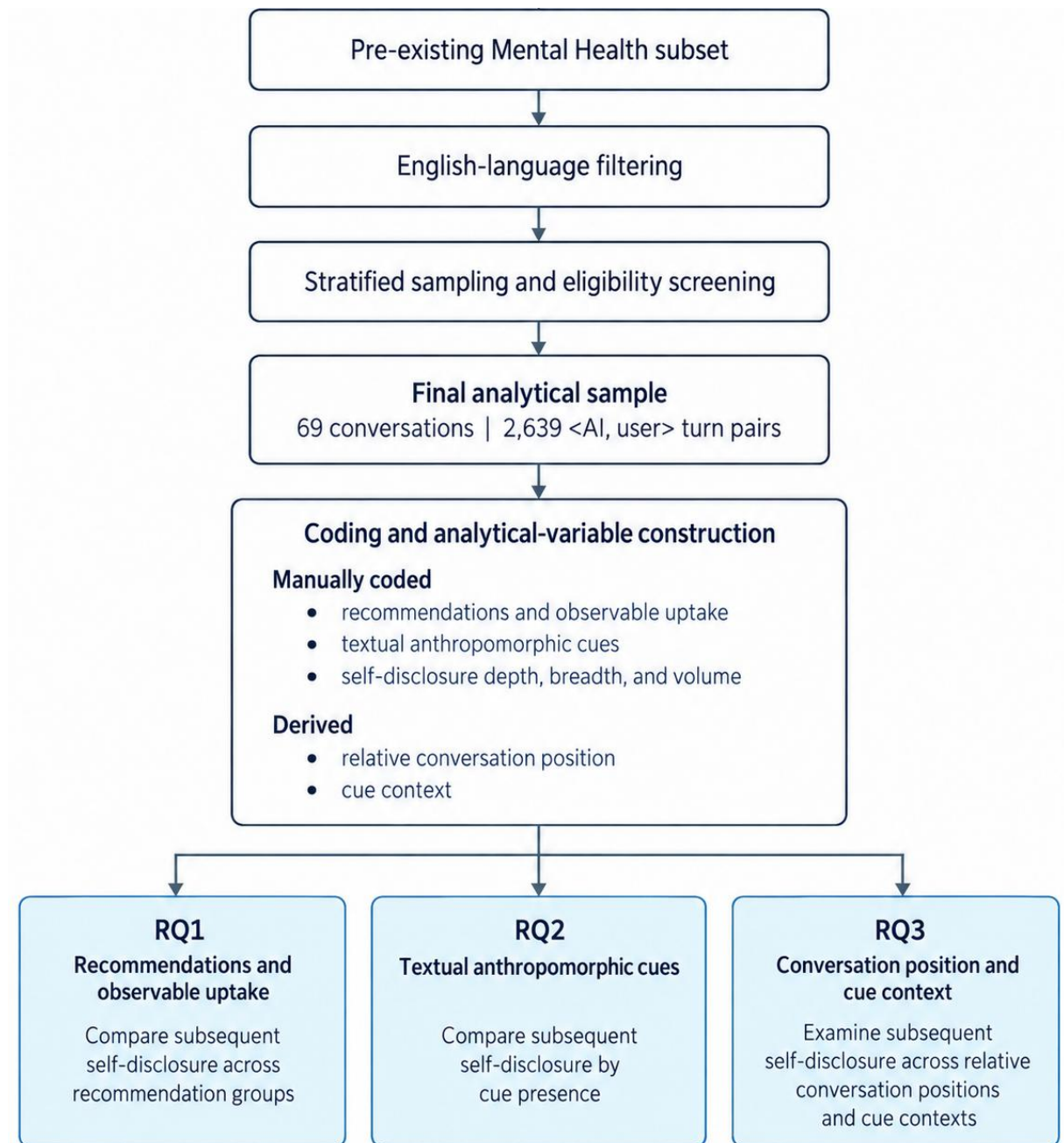
3.1 Research Design

This study used a non-experimental, observational research design based on secondary conversation data from the Karnam et al. (2026) dataset. Because the conversational features occurred naturally rather than being experimentally manipulated or randomly assigned, this design supports analysis of associations.

The study was predominantly deductive: the theoretical constructs and dimensions developed in Chapter 2 provided the initial analytical structure. These constructs were operationalised using directed qualitative content analysis, in which existing theory and prior research guide the initial coding framework (Hsieh & Shannon, 2005). Limited inductive refinement was used to develop specific coding categories for the ways in which the predefined theoretical dimensions appeared in the conversations. The qualitative coding was used to convert the conversation text into coded variables that could subsequently be analysed quantitatively.

This design was selected because RQ1 and RQ2 examine naturally occurring associations between characteristics of preceding ChatGPT messages and self-disclosure in the immediately subsequent user response. RQ3 instead examines how self-disclosure varies across relative positions within the same conversations. The dataset is suited to these analyses because it preserves donated ChatGPT conversations in their original conversational sequence. This allows adjacent ChatGPT-user turn pairs to be identified and each user response to be located relative to the progression of its conversation (Karnam et al., 2026). This makes the dataset suitable for analysing the turn-level patterns targeted by the research questions without reconstructing the interaction retrospectively or generating researcher-controlled conversation scenarios.

Figure 3 provides an overview of the research design, showing how the source data were transformed into the final analytical dataset and how the shared coding and variable-construction workflow led to the three research-question analyses. The corresponding procedures are detailed in Sections 3.2–3.7.

Figure 3*Overview of the Research Design and Analytical Workflow*

3.2 Data Source and Access

This study used secondary conversational data obtained from the research project reported by Karnam et al. (2026). In the source study, ChatGPT conversation histories were donated by 300 participants recruited through Prolific. Access for the present thesis was restricted to a pre-existing subset of conversations classified as Mental Health. The researcher did not have access to the complete dataset described in the source study. The subset available at the start of this thesis contained 6,482 conversations comprising 39,158 original user-ChatGPT turns. The records contained the conversational text and the information required to preserve conversation identity and message order. This Mental Health subset constituted the starting dataset for the present study and is hereafter referred to as the dataset.

The Mental Health classification had been established before the present study. In the source study, GPT-4o was used to generate a concise topic description for each conversation. Then the descriptions were clustered using BERTopic and qualitatively grouped by two researchers into 40 higher-level topic categories, including Mental Health (Grootendorst, 2022). The present study did not independently determine which conversations constituted mental-health interactions but used the supplied Mental Health category as its initial empirical boundary. Consequently, the composition of the starting dataset partly reflects the topic-classification procedure applied in the source study (Karnam et al., 2026).

This dataset was suitable for this study because the donated histories preserve ChatGPT messages and user responses in their original conversational sequence. This made it possible to analyse successive turns within naturally occurring conversations. The conversations had already occurred before donation and were therefore not generated specifically for a researcher-directed interaction task. These characteristics made the dataset suitable for examining the observable conversational patterns addressed by the research questions.

3.3 Construction of the Sampling Frame

The sampling frame was constructed from the pre-existing Mental Health category, which contained 6,482 conversations. The complete conversation was used as the sampling unit because the study examines variation across positions within a conversation and therefore required the original sequence and length of each conversation to remain intact. Individual turns were consequently not sampled independently.

The source conversations were provided as nested JSON records. Before candidate conversations could be selected, these records were prepared in two steps. First, they were restructured into a consistent format while preserving conversation identity and message order. Second, language eligibility was determined so that only English-language conversations were retained in the sampling frame.

3.3.1 Data Restructuring

The mental-health conversations contained 39,158 original user-ChatGPT turns and were stored as nested JSON records. An original turn consisted of one user message followed by the corresponding ChatGPT response (Karnam et al., 2026). Each JSON file represented one conversation. For language classification, the nested records were restructured into a table in which each row represented one original turn. Conversation identifiers, the original turn order, the user message, the ChatGPT response, and the associated timestamps were preserved during this restructuring. No conversational content was altered, and each turn remained traceable to its original conversation.

3.3.2 Language Eligibility

The pre-existing Mental Health category contained conversations in multiple languages. The sampling frame was restricted to English-language conversations so that the same coding scheme could be applied consistently across all selected conversations and to reduce interpretive variation introduced by language differences. Language was identified using *langdetect* version 1.0.9, with *DetectorFactory.seed* fixed at 0 to obtain reproducible classifications across runs.

Language was classified separately for each original turn. The user message and corresponding ChatGPT response were stripped of surrounding whitespace and combined into a single text string. If the combined text contained fewer than 20 characters, or if the automated detector could not assign a language, the turn was classified as “unknown”. Language classification therefore referred to the combined user-ChatGPT turn rather than to the user message alone.

Because complete conversations formed the sampling unit, eligibility was determined at the conversation level. A conversation entered the sampling frame only when every turn was automatically classified as English. A conversation containing at least one turn classified as another language or as “unknown” was excluded. This conservative rule prioritised a consistently English-language sampling frame over retaining the maximum number of conversations.

Applying this procedure reduced the source pool from 6,482 conversations containing 39,158 turns to 4,978 conversations containing 28,906 turns. The retained turns were grouped by conversation and ordered according to their original position. All turns belonging to the same conversation were then combined into a single row in the sampling dataset, creating the conversation-level sampling frame used for subsequent stratification and sampling.

3.4 Sampling Strategy

The sampling frame consisted of 4,978 eligible English-language mental-health conversations. Because most conversations were short, the sampling frame was first divided into conversation-length strata. Conversations were then randomly drawn

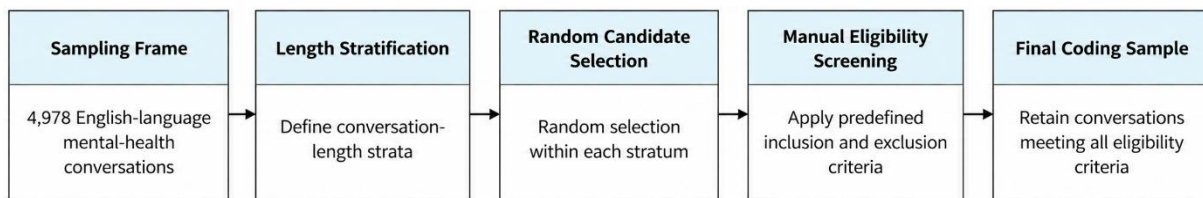
within these strata and manually screened against four predefined eligibility criteria before inclusion in the final sample.

The number of conversations selected from each length stratum was not proportional to the number of conversations available in that stratum. Relatively more conversations were selected from the longer-length strata to increase the representation of longer conversations in the sample.

The strategy was designed to obtain coverage of both shorter and longer conversations while keeping the manual screening and coding workload feasible. It was therefore not intended to reproduce the conversation-length distribution of the complete sampling frame. Figure 4 summarises the progression from the sampling frame through length stratification, candidate selection, and eligibility screening to the final coding sample.

Figure 4

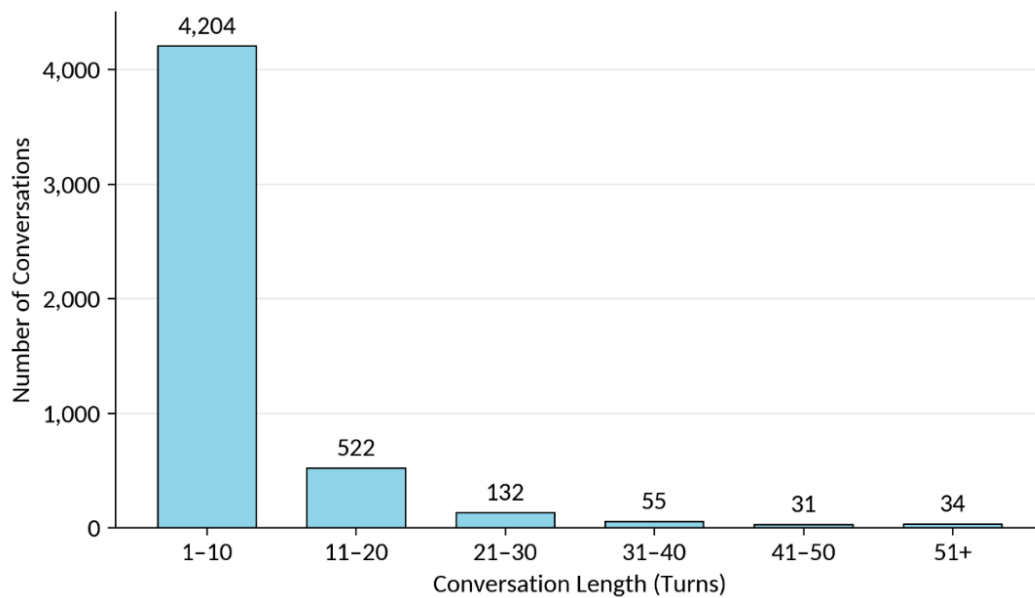
Overview of the Sampling Strategy



3.4.1 Conversation-Length Stratification

Conversation length was defined as the number of original user-ChatGPT turns within a conversation. Conversation length varied substantially within the sampling frame and followed a strongly right-skewed distribution. Of the 4,978 eligible conversations, 4,204 contained between 1 and 10 turns, while progressively fewer conversations occurred at greater lengths (Figure 5). An unrestricted random sample would therefore have been dominated by short conversations and would have provided relatively little coverage of longer interactions. The full conversation-length distribution is reported in Appendix D.1, Figure D1.

Conversation length was particularly relevant because RQ3 examines how self-disclosure varies across relative positions within a conversation. Without stratification, the strong concentration of short conversations would have meant that this analysis was based predominantly on short conversational trajectories. Conversation length was therefore used as the stratification variable to ensure coverage of both shorter and longer interactions.

Figure 5*Distribution of English-Language Conversations Across Conversation-Length Strata*

Note. The bars show the number of conversations in the sampling frame ($N = 4,978$) within each conversation-length stratum.

The sampling frame was divided into six conversation-length strata: 1–10, 11–20, 21–30, 31–40, 41–50, and 51 or more turns. Equal-width 10-turn intervals were used to create a simple and transparent sampling structure that distributed candidate selection across progressively longer conversations. The cut-off points were therefore not derived from theory or assumed to represent distinct types or stages of conversation. Their purpose was to prevent the large concentration of short conversations from determining the composition of the sample. Because only a small number of conversations extended beyond 50 turns, all conversations containing 51 or more turns were combined into one open-ended stratum rather than subdividing this tail into increasingly sparse groups.

3.4.2 Candidate Selection and Screening

Candidate conversations were randomly selected within each conversation-length stratum without replacement. A conversation selected in one draw could therefore not be selected again in a later draw. An approximate target of ten retained conversations per stratum was used as a pragmatic allocation criterion. This ensured that each conversation-length stratum contributed multiple eligible conversations to the analytical sample while avoiding a sample dominated by the much larger short-conversation strata.

Eligibility could only be determined through the manual screening procedure described in Section 3.4.3. The number of candidate conversations required to reach the approximate retention target could therefore not be determined in advance, and candidate selection was conducted iteratively.

The first draw selected 17 conversations from each of the six strata, resulting in 102 candidate conversations. This allocation was chosen to create an initial screening batch of approximately 100 conversations. The number was also constrained by the 41–50-turn stratum, which contained only 31 conversations; selecting 17 allowed a substantial initial draw while leaving unused conversations available for a possible second draw. Manual screening retained 27 conversations from this first draw. A second draw therefore selected another 17 conversations from each stratum where sufficient unused conversations remained. In the 41–50-turn stratum, only 14 unused conversations remained, so all 14 were selected. The second draw contained 99 candidates, of which 30 were retained after screening. This brought the cumulative retained sample to 57 conversations.

After the first two draws, additional coverage was still sought in the four shorter strata. The 41–50 and 51+ strata had been exhausted and were therefore not included in the third draw. Approximately 15 additional retained conversations were sought to bring the sample closer to the intended coverage across strata. Given the approximately 30% retention proportion observed across the first two draws, a final screening batch of 50 candidate conversations was considered sufficient to obtain roughly this number of additional eligible conversations. The candidates were distributed approximately evenly across the four remaining strata: 13 were drawn from both the 1–10 and 11–20 strata, and 12 from both the 21–30 and 31–40 strata. The slight difference in allocation served only to obtain a total batch of 50 candidates and did not reflect a theoretically or statistically derived distinction between the strata. Twelve of these candidates were retained after screening. Across the three draws, 251 candidate conversations were screened and 69 were retained. Table 1 summarises the candidate selection procedure.

Table 1

Candidate Selection and Retention by Conversation-Length Stratum

Conversation-length Stratum	Draw 1	Draw 2	Draw 3	Total screened	Retained
1–10 turns	17	17	13	47	7
11–20 turns	17	17	13	47	16
21–30 turns	17	17	12	46	11
31–40 turns	17	17	12	46	14
41–50 turns	17	14	0	31	7
51+ turns	17	17	0	34	14
Total	102	99	50	251	69

Note. Candidate conversations were sampled without replacement.

Sampling was stopped after the third draw. The approximate target of ten retained conversations per stratum was used as a coverage objective rather than as a fixed quota. At this stage, four strata had reached or exceeded the target, while the 41–50-turn stratum had been exhausted with seven retained conversations. The 1–10-

turn stratum was the only stratum that remained below the target while still containing unused candidates. A fourth draw restricted to the 1–10-turn stratum would have increased the representation of short conversations without improving the longer-conversation coverage required for the conversation-position analysis. Because equal stratum sizes were not an objective of the sampling design, no additional draw was conducted solely to equalise the retained counts.

3.4.3 Candidate Eligibility Screening

Each candidate conversation was manually screened against four predefined eligibility criteria before it could enter the final coding sample. The criteria concerned mental-health relevance, English-language use, interpretability from the available textual data, and minimum conversation length. A conversation was retained only when it satisfied all four criteria presented in Table 2.

For mental-health relevance, English-language use, and interpretability, a 90% threshold was applied to each criterion. The 90% threshold was selected as a pragmatic, conservative operational criterion: it permitted occasional off-topic, non-English, or uninterpretable turns without excluding an otherwise eligible conversation, while requiring the large majority of each conversation to satisfy the criterion. The threshold was not treated as an empirically validated cut-off.

English-language eligibility was manually verified even though the sampling frame had already been restricted using the automated procedure described in Section 3.3.2. This served as a validation step because automatically retained conversations could still contain non-English content. Conversations excluded during the earlier automated filtering stage were not reconsidered.

Table 2
Eligibility Criteria for Candidate Conversations

Criterion	Decision Rule
Mental-health relevance	Retain if at least 90% of the original <user, AI> turns concern the user’s own mental-health or emotional concerns, including personal distress, emotional experiences, relationship difficulties discussed as personal concerns, or coping needs. Exclude conversations primarily concerning academic, fictional, professional, survey-related, or otherwise non-personal discussion of mental health.
English language	Retain if at least 90% of the original turns are judged to be written in English during manual verification; otherwise exclude.
Interpretability without missing files	Retain if at least 90% of the original turns can be interpreted from the available textual conversation data without requiring unavailable images, documents, attachments, or other external content; otherwise exclude.

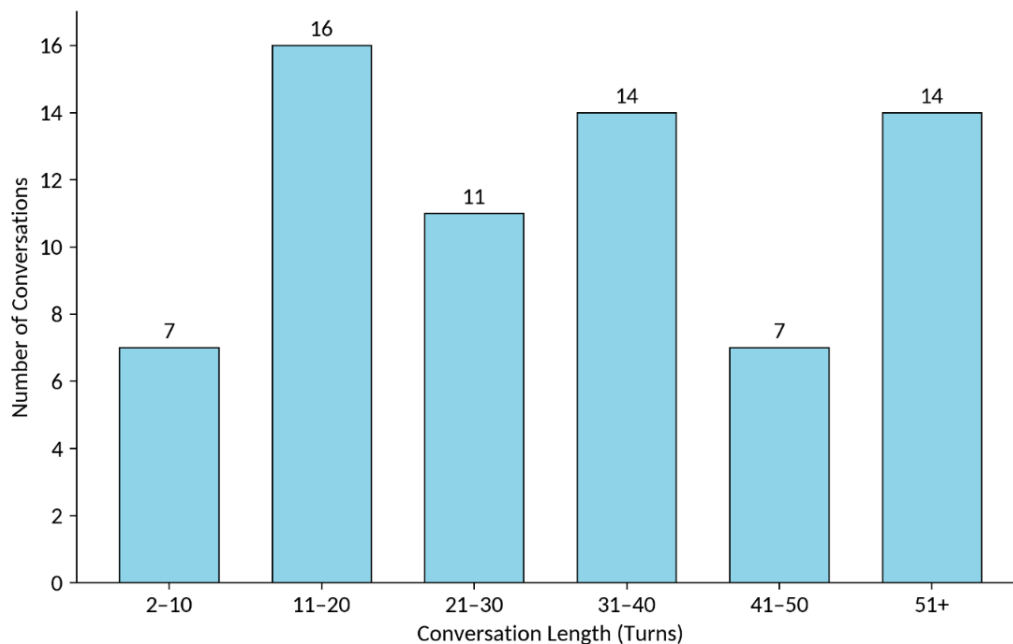
Criterion	Decision Rule
Minimum conversation length	Retain only conversations containing at least two original user–ChatGPT turns. This ensured that each conversation contained at least one ChatGPT message followed by a subsequent user response, which is the message sequence required to construct the analytical <AI, user> turn pair introduced in Section 3.5.

3.4.4 Final Sample

Across the three candidate draws, 69 conversations satisfied all four eligibility criteria and were included in the final coding sample. Duplicate conversation identifiers were checked to ensure that each conversation appeared only once. The final sample comprised 2,708 original <user, AI> turns. Figure 6 shows the distribution of the retained conversations across the six conversation-length strata.

Figure 6

Distribution of Conversations Across Conversation-Length Strata in the Final Sample



Note. The shortest final-sample category is labelled 2–10 turns because one-turn conversations from the original 1–10-turn sampling stratum were excluded during manual screening, as explained in Section 3.5.

3.5 Construction of the <AI, user> Turn Pair

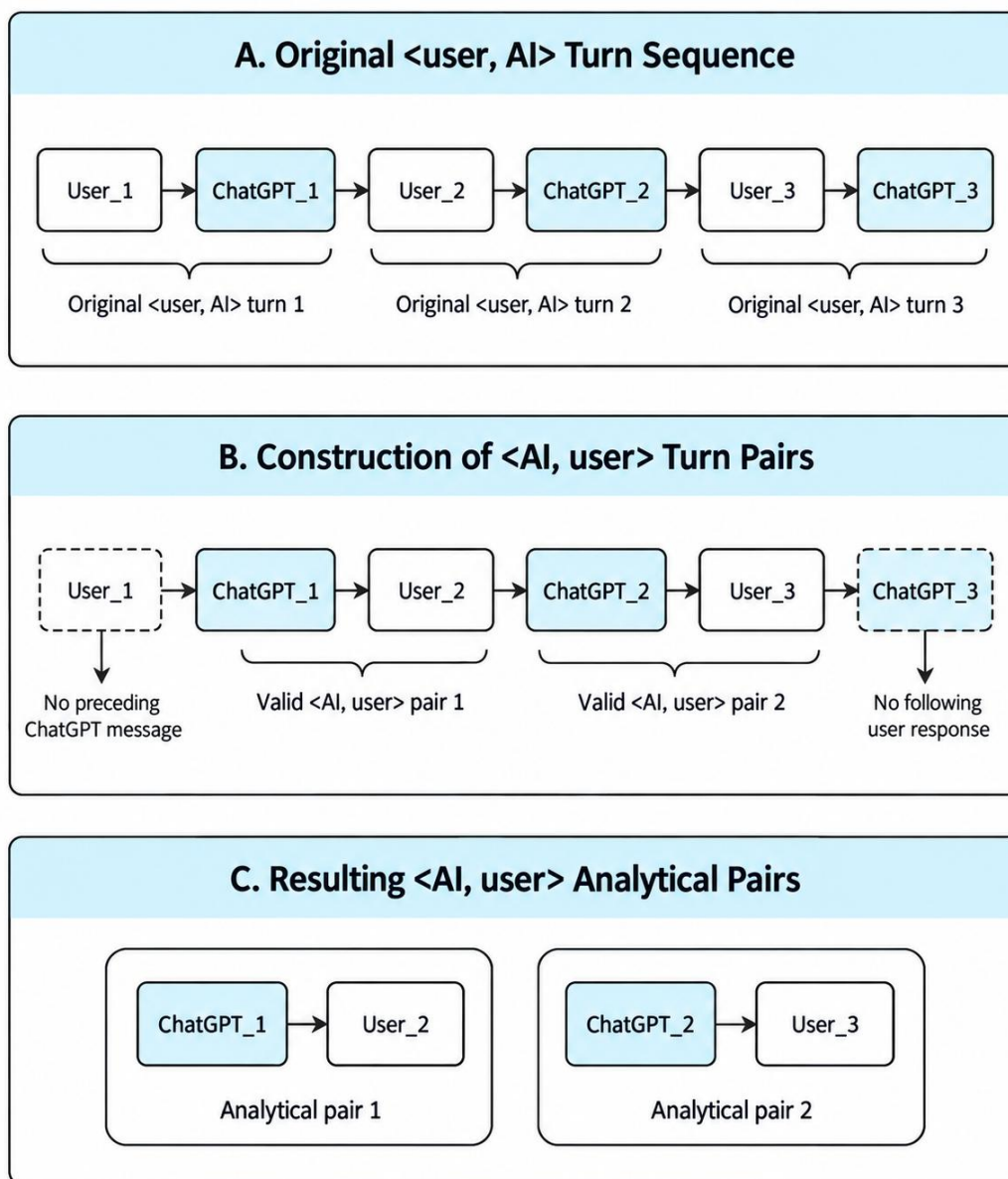
The analytical unit used in this study was an <AI, user> turn pair, consisting of one ChatGPT message and the user response that immediately followed it. This constructed analytical unit differs from the original <user, AI> turn described in Section 3.3.1, which consists of a user message followed by the corresponding ChatGPT response. For example, in the sequence User₁ → AI₁ → User₂ → AI₂,

the first analytical pair is AI_1 → User_2. Figure 7 illustrates this restructuring process.

This restructuring excluded the first user message, because no ChatGPT message preceded it, and the final ChatGPT message, because no subsequent user response was available. Consequently, each conversation produced one fewer analytical <AI, user> pair than its number of original <user, AI> turns. Each pair retained its original conversation identifier and sequential position, preserving its location within the conversation. The sequential position was subsequently used to derive the relative conversation position examined in RQ3.

Figure 7

Construction of the <AI, user> Turn Pair



The <AI, user> turn pair was selected because RQ1 and RQ2 examine whether characteristics of a ChatGPT message are associated with self-disclosure in the

immediately subsequent user response. Using either message alone as the analytical unit would not jointly capture the characteristics of the ChatGPT message and the self-disclosure in the response that immediately followed it.

The coding scheme, introduced in Section 3.6, operationalised depth, breadth, and volume at the level of the complete user response. Splitting a response into separate sentences would therefore divide a single coded response into multiple observations, even though its depth, breadth, and volume were defined and assessed across the response as a whole. Conversely, combining several turns or analysing complete conversations would make it more difficult to relate a particular ChatGPT message to the user response that immediately followed it. The <AI, user> pair therefore provided a consistent unit for relating observable characteristics of a ChatGPT message to self-disclosure in the subsequent user response.

The <AI, user> turn pair captures the local sequential relationship between a ChatGPT message and the immediately subsequent user response. Applying this restructuring to the 2,708 original <user, AI> turns across the 69 retained conversations produced 2,639 valid <AI, user> turn pairs. These pairs constituted the analytical observations used for qualitative coding and subsequent quantitative analysis.

3.6 Qualitative Coding

The 2,639 <AI, user> turn pairs were manually coded to operationalise the constructs examined in the research questions. ChatGPT messages were coded for recommendations and textual anthropomorphic cues, while the immediately subsequent user responses were coded for observable recommendation uptake and self-disclosure across depth, breadth, and volume. The following sections describe the development, application, and quality-assurance procedures of the coding scheme, while the complete codebook and detailed coding procedure are provided in Appendix B.

3.6.1 Code Creation

Building on the predominantly deductive approach described in Section 3.1, the coding scheme used the theoretical constructs developed in Chapter 2 as its initial structure, following directed qualitative content analysis (Hsieh & Shannon, 2005). Limited inductive refinement was subsequently used to specify how these predefined constructs were represented in the sampled conversations.

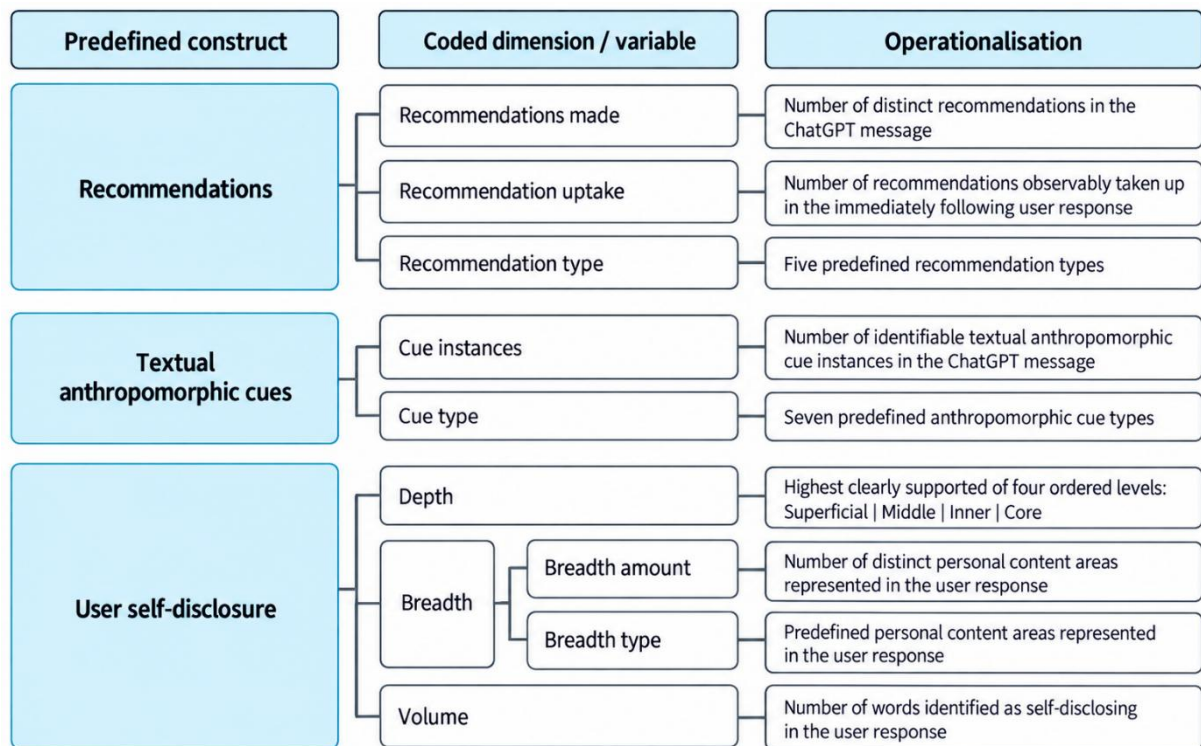
For this refinement, the 2,639 <AI, user> turn pairs produced in Section 3.5 were reviewed once to identify the specific forms in which the predefined constructs occurred in the conversations. During this review, candidate codes were identified for the specific ways in which the predefined constructs appeared in the conversations. These candidate codes formed an initial set of coding categories that was subsequently refined into the final coding scheme. Candidate codes with overlapping meanings or conversational functions were subsequently combined where appropriate to reduce redundancy while retaining distinctions relevant to the

research questions. The resulting coding scheme therefore remained theory-led, while its specific codes were refined based on how the predefined constructs occurred in the analysed conversations.

Figure 8 summarises the final coding scheme by linking the three predefined constructs to the dimensions and variables through which they were operationalised. Sections 3.6.1.1–3.6.1.3 explain how each construct was operationalised, while the complete definitions and decision rules are provided in Appendix B, Sections B.1, B.2 and B.3.

Figure 8

Overview of the Final Coding Scheme and Operationalisation



3.6.1.1 Operationalising Recommendations

Building on the distinction between recommendations and broader conversational steering established in Section 2.3.1, this study used a study-specific operationalisation of recommendations. Recommendations were defined as model-initiated conversational moves that could direct or invite the immediately following user response. This definition extends beyond conventional action-oriented advice. It includes several forms of conversational steering, while preserving their distinct conversational functions in the coding scheme (Feng, 2009; Karnam et al., 2026; Xiao et al., 2020). Within this operationalisation, five forms of conversational steering were included (specified in full in Table B2):

Follow-up Question:	Requested additional information, clarification, or contextual detail about the user's situation or a previous statement.
Ongoing-Support Offer:	Expressed continued availability or invited the user to continue or return to the conversation.
Reflective Question:	Invited the user to reflect on feelings, thoughts, needs, personal meaning, or their interpretation of the situation.
Suggestive Prompt:	Proposed an action, strategy, or next step without presenting it as a question.
Suggestive Question:	proposed a possible action, service, or conversational direction through a question.

These five forms were grouped under the study-specific label of recommendations because each could direct or invite the immediately following user response. When a usable ChatGPT message contained none of these five forms, recommendation type was coded as *No Recommendation*.

The construct was operationalised through three measures: the number of distinct recommendations identified in the ChatGPT message, the number of those recommendations observably taken up in the immediately following user response, and the recommendation type or types present. Observable uptake was identified when the following user response answered a question, performed the requested conversational action, or continued in the proposed direction. This measure therefore captures conversational uptake rather than agreement, acceptance, or real-world compliance.

To align recommendation coding with the <AI, user> analytical unit, only recommendations for which uptake could be evaluated in the immediately following user response were included. Suggestions requiring action outside the conversation or beyond the paired response were therefore excluded. This ensured that recommendation presence and conversational uptake were assessed within the same analytical unit. Complete variable definitions, recommendation-type decision rules, and counting and uptake rules are provided in Appendix B.1.

3.6.1.2 Operationalising Textual Anthropomorphic Cues

As established in Section 2.4.1, textual anthropomorphic cues and perceived anthropomorphism are related but distinct. Textual cues are observable features of chatbot language, whereas perceived anthropomorphism concerns the extent to which users perceive the system as possessing human-like qualities (Greilich et al., 2025; Li et al., 2023).

For this study, textual anthropomorphic cues were operationalised as expressions through which ChatGPT presented itself as possessing human-like agency, thoughts,

emotions, intentions, judgements, or social roles. Because the dataset consisted of written conversations and contained no measure of users' perceptions, the coding was limited to these observable features in the ChatGPT message and does not indicate whether users actually perceived ChatGPT as human-like (Li et al., 2023). Seven study-specific cue types were used to classify these expressions, with the complete definitions and decision rules provided in Table B5:

Agency Framing:	Language presenting ChatGPT as capable of acting, helping, guiding, explaining, noticing, or performing a task.
Belief Framing:	A belief, thought, view, or opinion attributed to ChatGPT.
Collaborative Agency:	Language framing ChatGPT and the user as jointly undertaking an activity, process, or next step.
Emotional Attribution:	Emotional understanding, care, concern, empathy, or another emotional response attributed to ChatGPT.
Human-like Intention:	A want, hope, aim, intention, or personally framed desired outcome attributed to ChatGPT.
Human-like Judgment:	A personal evaluation, judgement, or assessment of the user, the user's actions, or the user's situation.
Social Role Framing:	Language positioning ChatGPT as a socially present interaction partner, such as a listener, supporter, helper, companion, or conversational partner.

To capture both how often textual anthropomorphic cues occurred and the forms in which they occurred, two measures were used: the number of identifiable cue instances in the ChatGPT message and the distinct cue types represented in that message. Each counted cue instance corresponded to one of the seven predefined cue types. Repeated occurrences of the same cue type therefore increased the cue count, whereas each cue type was recorded only once in the type-based measure. This distinction captured both the frequency of anthropomorphic language and the forms in which it occurred. Complete variable definitions, type-specific decision rules, and cue-counting rules are provided in Appendix B.2.

3.6.1.3 Operationalising User Self-Disclosure

Building on the conceptualisation in Section 2.2, this thesis defined user self-disclosure as personal information communicated by the user to ChatGPT about themselves (Omarzu, 2000). In the coding scheme, this included information about the user's own experiences, opinions, attitudes, feelings, relationships, or circumstances. Statements about another person were included only when they also revealed information about the user's own situation, experience, or relationship with that person (Harmsen et al., 2023). Self-disclosure was operationalised across three dimensions: depth, breadth, and volume, which together provided a multidimensional approximation of the construct.

Depth

Depth refers to the intimacy or sensitivity of the disclosed information (Carpenter & Greene, 2015; Omarzu, 2000). In this study, depth was operationalised as the highest level clearly present in the user response. Four ordered levels were used:

- Superficial: No identified self-disclosure, or only general, public, situational, or minimally personal information.
- Middle: Personal but not highly intimate or sensitive information.
- Inner: Clearly personal emotions, concerns, needs, conflicts, psychological difficulties, or sensitive experiences.
- Core: Highly intimate or vulnerable information, including severe mental-health struggles, trauma, suicidal thoughts, deep shame, or fundamental fears about the self.

When multiple levels were present, the highest clearly supported level was recorded. Because responses without identified self-disclosure were also coded as Superficial, this category combines no-disclosure responses with responses containing only low-depth personal information. Detailed decision rules are provided in Appendix B.3, Table B8.

Breadth

Breadth represents the range of personal information disclosed within a user response (Carpenter & Greene, 2015; Omarzu, 2000). In this study, a personal content area was defined as a substantively distinguishable subject of personal information disclosed by the user. Predefined breadth types were used to classify commonly occurring personal content areas, whereas content areas that did not fit a predefined type were represented by the umbrella label *Other*.

Breadth was operationalised through two measures: the breadth types represented in the response and the breadth amount, defined as the number of distinct personal content areas identified. For predefined breadth types, each type contributed a maximum of one to the breadth amount, regardless of repeated references within the same response. The umbrella category *Other* was treated differently: although *Other* was recorded only once as a breadth type, multiple substantively distinct uncategorised personal content areas could each contribute separately to the breadth amount. Complete breadth-type definitions and counting rules are provided in Appendix B.3, Table B9 and Table B10.

Volume

Volume captured the amount of self-disclosing text in the user response. Prior self-disclosure research distinguishes disclosure amount or duration from depth and breadth, and has used word count as one representation of disclosure amount (Omarzu, 2000). However, total response length does not necessarily correspond to self-disclosure; in conversational-agent research, word-count differences can occur

without corresponding differences in manually coded disclosure (Harmsen et al., 2023). Volume was therefore operationalised more narrowly as the number of words within text identified as communicating personal information about the user.

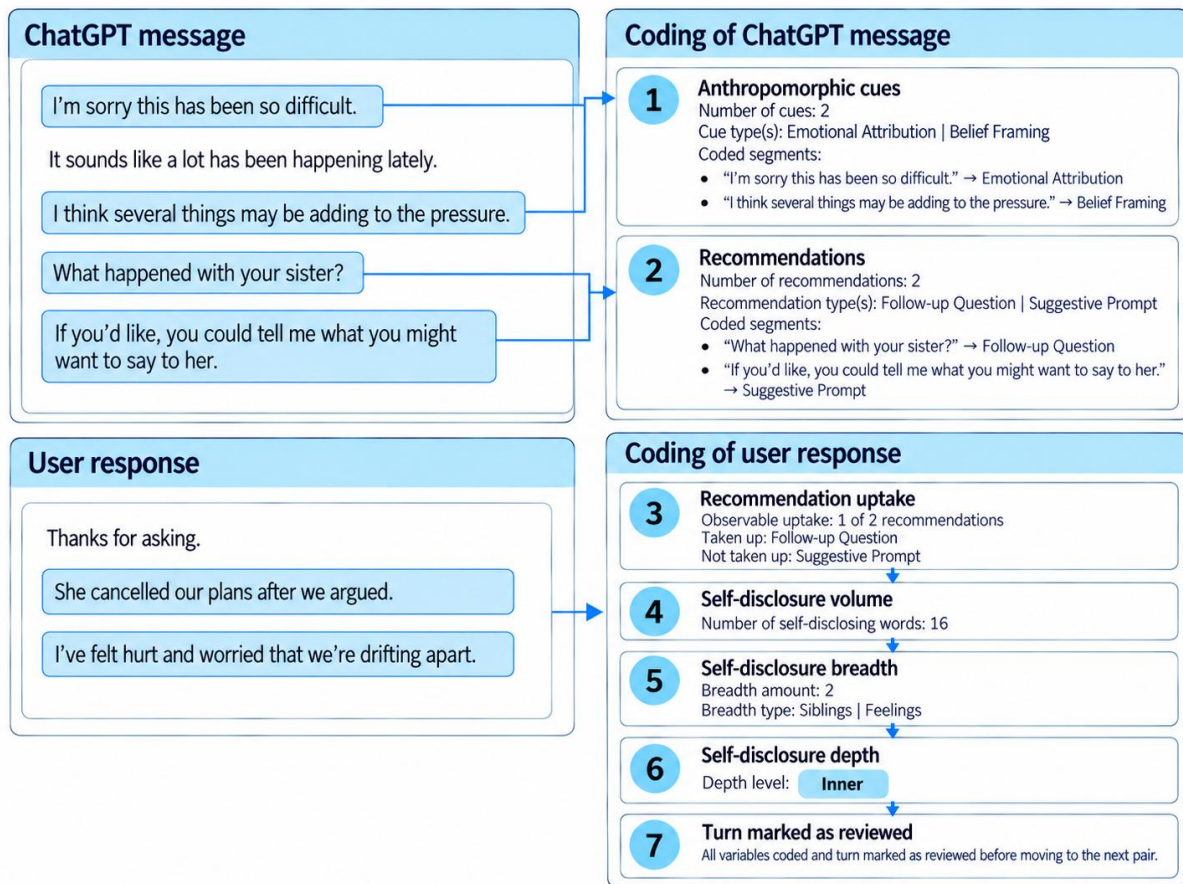
Words that did not communicate self-disclosing information were excluded, as was repeated information within the same response when it added no new personal information. Detailed counting rules are provided in Appendix B.3, Table B11.

3.6.2 Coding Sequence of an <AI, user> Turn

Each <AI, user> turn pair was manually coded using the same fixed sequence. The complete ChatGPT message was first assessed for textual anthropomorphic cues and recommendations. The immediately following user response was then assessed for observable recommendation uptake and self-disclosure. Within the user response, self-disclosure was coded in the operational sequence of volume, breadth, and depth. Self-disclosing text was first identified to determine volume, after which the applicable breadth-type categories and distinct coded personal content areas were identified to determine breadth type and breadth amount, respectively. Finally, the complete user response was assigned a depth level.

When interpretation of content in the target turn pair required earlier conversational context, preceding turns were consulted only to clarify the meaning of content present in that pair. Figure 9 summarises the fixed coding sequence in a synthetic example. The complete coding procedure, including contextual rules, field conventions, ambiguity handling, and detailed application rules, is provided in Appendix B.4.

Figure 9
Coding Sequence for an <AI, user> Turn Pair



Note. The conversational example is synthetic and was not copied, paraphrased, or adapted from donated conversations. For illustrative clarity, the figure indicates which recommendation was observably taken up. In the actual coding, only the number of recommendations taken up was recorded; uptake was not linked to individual recommendation types.

3.6.3 Coding Consistency and Quality Assurance

Coding consistency was supported by applying the same predefined definitions and decision rules to all <AI, user> turn pairs. Ambiguous cases and content eligible for multiple codes were handled using the construct-specific rules documented in Appendix B. Although these rules standardised the coding process, interpretation was still required when the meaning of conversational content was not immediately clear.

Quality checks focused on coding completeness and internal consistency. A turn pair was marked as reviewed only after all applicable coding fields had been completed. Before quantitative analysis, the completed coded dataset was checked for remaining empty fields and codes entered in incorrect fields. These checks supported procedural completeness and internal consistency. Because all coding was performed by one researcher, the coding was not independently replicated and no intercoder-reliability statistic was calculated. Detailed procedures for ambiguous cases and coding-completion checks are provided in Sections B.4.5 and B.4.6.

3.7 Quantitative Analysis

The coded dataset was analysed using descriptive, exploratory, and inferential statistical analyses. These analyses examined the relationships among the self-disclosure dimensions and addressed the three research questions by comparing disclosure across recommendation groups, textual anthropomorphic cues, and relative conversation positions. Supplementary and sensitivity analyses were used where relevant to assess the strength of the observed patterns. Detailed information on the statistical implementation is provided in Appendix C.

3.7.1 Descriptive Statistics

Self-disclosure was summarised separately for depth, breadth, and volume. Depth was characterised by the distribution of responses across the four coded levels. Breadth was described in two ways: breadth amount by its distribution, mean, and median, and breadth type by the frequency and proportion of responses containing each coded content area. Volume was characterised by the proportion of responses with no coded self-disclosure versus at least one coded self-disclosing word, supplemented by the mean, median, and maximum number of words coded as self-disclosing. The purpose was to provide context for interpreting the three self-disclosure outcomes by examining whether disclosures coded at greater depth were also broader in personal content and contained more self-disclosing content. The analysis was exploratory and was not used to validate the operationalisations or to treat the three dimensions as interchangeable.

For the main-text presentation of breadth types, only categories occurring in at least 5% of coded user responses were displayed. This threshold was used solely to maintain figure readability and did not affect the coding or subsequent analyses. All breadth types remained in the analytical dataset, and the complete distribution is reported in Appendix D.2.

Recommendations and observable uptake were summarised using the frequency and proportion of turn pairs containing no recommendation, a recommendation without observable uptake, or a recommendation with observable uptake. Recommendation types were summarised using the frequency and proportion of ChatGPT messages containing each type at least once. The mean numbers of recommendations made and observably taken up per ChatGPT message were also calculated, both across all messages and among messages containing at least one recommendation. Detailed recommendation-type descriptive results are reported in Appendix D.5.

Textual anthropomorphic cues were summarised using the frequency and proportion of ChatGPT messages containing no cue or at least one cue. The number of coded cue instances per message was further summarised using its distribution, mean, median, and range, while textual anthropomorphic-cue types were described using the frequency and proportion of ChatGPT messages containing each type at least once. Detailed descriptive results for textual anthropomorphic cues are reported in Appendix D.6.

3.7.2 Exploratory Analysis of Self-Disclosure Dimensions

A preliminary exploratory analysis examined how breadth amount and self-disclosure volume varied across the four coded depth levels: Superficial, Middle, Inner, and Core. The purpose was to provide context for interpreting the three self-disclosure outcomes by examining whether disclosures coded at greater depth were also broader (breadth amount) in personal content and contained more self-disclosing content (volume). The analysis was exploratory and was not used to validate the operationalisations or to treat the three dimensions as interchangeable. Breadth amount and self-disclosure volume were first summarised descriptively within each depth level using their distributions, means, and medians.

Breadth amount and self-disclosure volume were compared across the four depth levels using separate Kruskal-Wallis tests. This rank-based procedure was selected because both outcomes were numerical and strongly right-skewed and therefore did not require normally distributed outcome values (Kruskal & Wallis, 1952). Epsilon squared (ϵ^2) was reported as the effect-size measure for each overall comparison (Tomczak & Tomczak, 2014). When an overall comparison was statistically supported, follow-up pairwise comparisons were conducted to identify which specific depth levels differed. Because the overall test did not indicate which levels accounted for the difference, all six possible combinations of the four depth levels were examined using Mann-Whitney U tests (Mann & Whitney, 1947). Holm correction was applied across the six comparisons within each outcome.

As a sensitivity analysis, the overall Kruskal-Wallis tests were repeated after excluding observations flagged as potential outliers separately for breadth amount and self-disclosure volume. Potential outliers were identified within each depth level using the conventional Tukey boxplot criterion, whereby observations below $Q_1 - 1.5 \times IQR$ or above $Q_3 + 1.5 \times IQR$ were flagged (Hubert & Vandervieren, 2008). The analysis was used to assess whether the overall depth-level differences depended disproportionately on unusually low or high observations. The complete coded dataset remained the basis of the primary analysis, while the IQR-based analysis served only as a sensitivity check on the stability of the overall pattern.

3.7.3 RQ1: Recommendation Presence and Observable Uptake Analysis

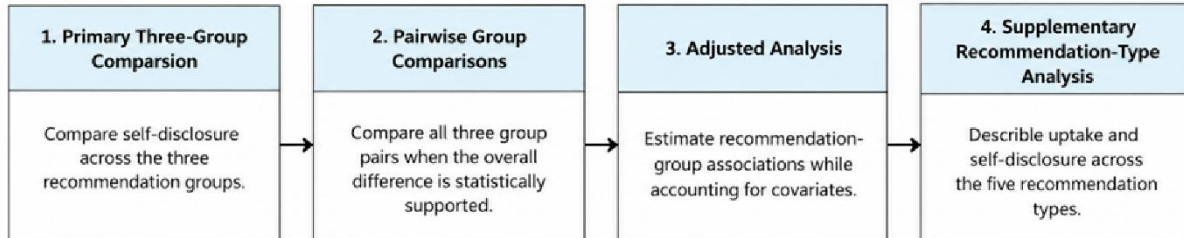
RQ1 examined whether self-disclosure in the subsequent user response differed according to recommendation presence and observable uptake within the <AI, user> turn pair. The analysis consisted of four complementary parts: a primary three-group comparison, follow-up pairwise group comparisons, an adjusted analysis, and a supplementary descriptive analysis across recommendation types. Figure 10 provides an overview of this analytical sequence.

Turn pairs were divided into three mutually exclusive groups: No recommendation, Recommendation without observable uptake, and Recommendation with observable uptake. This distinction separated recommendation presence from observable uptake rather than treating all recommendation-containing messages as a single group. Within each group, breadth amount and volume were summarised using their means

and medians. Depth was summarised as the percentage of responses classified as Superficial, Middle, Inner, or Core.

Figure 10

Analysis Roadmap for RQ1: Recommendation Presence and Observable Uptake



Primary Three-Group Comparison

The primary RQ1 analysis examined overall differences across the three recommendation groups for each self-disclosure dimension. Depth was compared using its complete distribution across Superficial, Middle, Inner, and Core, whereas breadth amount and volume were compared as numerical outcomes.

Pairwise Group Comparisons

For each outcome with a statistically supported overall group difference, follow-up pairwise comparisons were conducted across all three recommendation-group pairs to identify which groups differed. Pairwise comparisons were therefore used to localise statistically supported overall differences rather than as independent primary tests. Holm correction was applied separately across the three pairwise comparisons within each outcome.

Adjusted Analysis

After the primary and pairwise comparisons, adjusted analyses were conducted for breadth amount, ordered depth, and volume. For these analyses, depth was represented as an ordered score from 1 (Superficial) to 4 (Core), allowing it to be examined within the same regression framework as breadth and volume. This numerical representation treats each step from Superficial to Middle, Middle to Inner, and Inner to Core as equally large, although such spacing is not established by the coding scheme. Therefore, the ordered-depth regression was used only as an additional adjusted analysis. The primary depth analysis retained the complete four-category distribution. Recommendation group was the focal predictor, with No recommendation as the reference category. Turn number and textual anthropomorphic-cue presence were included as additional predictors. This allowed the recommendation-group associations to be examined conditional on conversation progression and the presence of the other examined cue. A linear contrast derived from the same fitted models additionally compared recommendations with and without observable uptake because RQ1 explicitly distinguishes between these two responses.

Supplementary Recommendation-type Analysis

Finally, the five coded recommendation types were examined descriptively to assess whether the broader recommendation-group patterns varied across recommendation forms. Observable uptake was summarised by recommendation type, after which depth, breadth amount, and volume were described within each recommendation-type × uptake-group combination. Because multiple recommendation types could occur within the same ChatGPT message and the types differed substantially in frequency, these type-level comparisons were not treated as independent inferential tests. They were therefore used as a supplementary descriptive analysis, with the complete results reported in Appendix D.5.

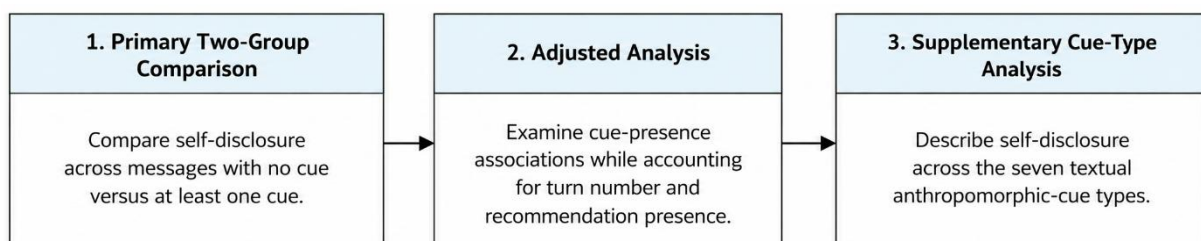
3.7.4 RQ2: Textual Anthropomorphic Cue Analysis

RQ2 examined whether self-disclosure in the subsequent user response differed according to the presence of textual anthropomorphic cues in the preceding ChatGPT message. Figure 11 provides an overview of the analytical sequence used to address this research question.

The analysis consisted of three complementary parts: a primary comparison between responses following messages with and without textual anthropomorphic cues, an adjusted analysis accounting for turn number and recommendation presence, and a supplementary descriptive analysis of the seven individual cue types.

Figure 11

Analysis Roadmap for RQ2: Textual Anthropomorphic Cues



Primary Two-Group Comparison

For the primary analysis, turn pairs were divided into two mutually exclusive groups: messages containing no textual anthropomorphic cue and messages containing at least one textual anthropomorphic cue. Within each group, breadth amount and volume were summarised using their means and medians. Depth was summarised as the percentage of responses classified as Superficial, Middle, Inner, or Core. The two groups were then compared across all three self-disclosure dimensions.

The four depth categories have an inherent order from Superficial to Core. Because RQ2 compared two cue-presence groups, this ordering was used in a Mann-Whitney U test to assess whether responses in one group tended to have higher or lower disclosure depth than responses in the other group (Mann & Whitney, 1947). The

complete four-category distribution was retained descriptively to show how any difference was distributed across the individual depth levels.

Adjusted Analysis

After the primary comparison, adjusted analyses were conducted for ordered depth, breadth amount, and volume. Ordered depth used the same 1 (Superficial) to 4 (Core) representation and supplementary interpretation described for RQ1 in Section 3.7.3. The primary RQ2 depth comparison retained the ordinal ordering without assuming equal distances between categories.

Textual anthropomorphic-cue presence was the focal predictor, with no cue as the reference category. Turn number and recommendation presence were included as additional predictors. Recommendation presence was coded irrespective of observable uptake because presence was determined from the preceding ChatGPT message, whereas observable uptake was classified from the subsequent user response. This allowed cue-presence associations to be examined conditional on conversation progression and recommendation presence in the preceding ChatGPT message.

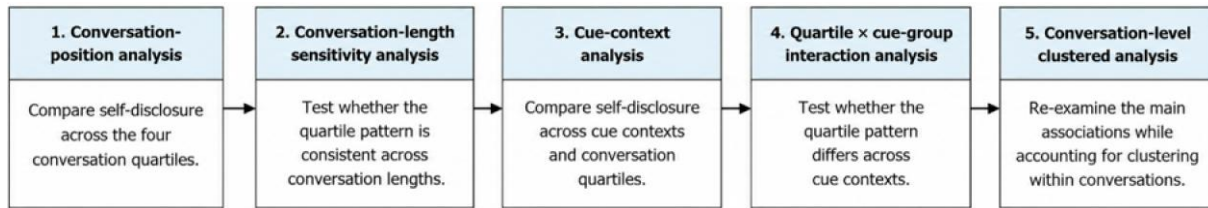
Supplementary Cue-type Analysis

Finally, the seven coded textual anthropomorphic-cue types were examined descriptively to assess whether the broader cue-presence pattern varied across specific forms of anthropomorphic language. For each cue type, depth was summarised using the percentage distribution across Superficial, Middle, Inner, and Core, while mean breadth amount and volume were calculated for the corresponding subsequent user responses. Responses following messages without textual anthropomorphic cues were retained as a descriptive reference category. Because multiple cue types could occur within the same ChatGPT message, a single turn pair could contribute to more than one cue-type category. Cue types also differed substantially in frequency. The type-level analysis was therefore used to describe variation across cue forms rather than to test independent differences between individual cue types. Complete type-level results are reported in Appendix D.6.

3.7.5 RQ3: Conversation Position and Cue-Context Analysis

RQ3 examined how user self-disclosure varied across relative conversation position and whether these patterns differed across conversational cue contexts. Figure 12 provides an overview of the analytical sequence used to address this research question.

The analysis consisted of five complementary parts: a primary comparison across conversation positions, a conversation-length sensitivity analysis, a cue-context analysis, a conversation quartile $\times$ cue-group interaction analysis, and a conversation-level clustered analysis. The final analysis re-examined the central RQ3 associations while accounting for dependence between turn pairs originating from the same conversation.

Figure 12*Analysis Roadmap for RQ3: Conversation Position and Cue Context*

Because the sampled conversations differed substantially in length, conversation position was expressed relative to each conversation rather than by absolute turn number. For each analytical <AI, user> turn pair, relative position was calculated as $(i - 1)/(n - 1)$, where i represents the sequential position of the pair and n the total number of analytical pairs in that conversation. This assigned the first analytical pair a relative position of 0 and the final pair a position of 1. For conversations containing only one analytical pair, relative position was set to 0. The resulting scale allowed positions from conversations of different lengths to be compared on a common relative scale.

For the RQ3 analyses, relative position was divided into four fixed equal-width intervals: Q1 (0–0.25), Q2 (>0.25–0.50), Q3 (>0.50–0.75), and Q4 (>0.75–1.00). Observations exactly at 0.25, 0.50, and 0.75 were therefore assigned to Q1, Q2, and Q3, respectively. These quartiles represented ranges of relative conversation position rather than four groups containing equal numbers of turn pairs.

Conversation-position Analysis

The primary conversation-position analysis compared depth, breadth amount, and volume across the four conversation quartiles. Breadth amount and volume were summarised using means and medians, while depth was represented by its percentage distribution across Superficial, Middle, Inner, and Core. Overall comparisons assessed whether each self-disclosure outcome differed across conversation quartiles. When an overall difference was statistically supported, pairwise quartile comparisons were conducted to identify where the differences occurred.

Conversation-length Sensitivity Analysis

A conversation-length sensitivity analysis then examined whether the pooled quartile pattern was also observed across conversations of different lengths. The quartile analysis was repeated separately within the six conversation-length strata used in the sampling design: 2–10, 11–20, 21–30, 31–40, 41–50, and 51+ turns. Within each stratum, depth, breadth amount, and volume were compared across the four conversation quartiles using the same analytical approach as in the pooled analysis. This analysis was used as a sensitivity check because the same relative quartile can contain different absolute numbers of turns in short and long conversations. It therefore assessed whether the pooled conversation-position pattern was broadly consistent across conversation lengths rather than being concentrated in particular length ranges. Complete stratum-specific results are reported in Appendix D.8.

Cue-context Analysis

The cue-context analysis next examined whether conversation-position patterns differed according to the focal conversational features present in the preceding ChatGPT message. Each turn pair was assigned to one of four mutually exclusive groups: No cue, Recommendation only, Textual anthropomorphic cue only, or Both cues. No cue indicated that neither focal feature was present, whereas Both cues indicated that a recommendation and at least one textual anthropomorphic cue occurred in the same ChatGPT message. Recommendation presence was used irrespective of observable uptake because this analysis concerned the cue context provided by the ChatGPT message rather than the user's subsequent response to the recommendation. Cue-group composition was first summarised across conversation quartiles. Depth, breadth amount, and volume were then examined across the cue-group $\times$ conversation-quartile combinations, both across quartiles within each cue group and across cue groups within each quartile.

Quartile $\times$ Cue-Group Interaction Analysis

A conversation quartile $\times$ cue-group interaction analysis subsequently tested whether the overall conversation-position pattern itself differed across the four cue contexts. The interaction was examined separately for depth, breadth amount, and volume. This analysis served a different purpose from the preceding subgroup comparisons: those comparisons identified where particular groups or quartiles differed, whereas the interaction analysis directly tested whether the pattern across conversation quartiles differed between cue groups.

Conversation-level Clustered Analysis

Finally, an additional conversation-level clustered analysis was conducted because the primary RQ3 analyses treated individual <AI, user> turn pairs as analytical observations, even though multiple turn pairs originated from the same conversation. Observations within the same conversation may be statistically dependent because they occur within a shared conversational sequence, so treating all turn pairs as fully independent can lead to overly precise inference (A. Colin Cameron & Douglas L. Miller, 2015). The clustered analysis therefore re-examined the central conversation-position and conversation quartile $\times$ cue-group associations while retaining all 2,639 turn pairs but adjusting the inference for clustering within the 69 sampled conversations. This provided an additional check of whether the central RQ3 findings remained supported when the nested structure of the data was taken into account.

3.7.6 Statistical Procedures and Reporting Conventions

This section provides the statistical information needed to interpret the results reported in Chapter 4. Table 3 summarises the principal procedures used across the analyses, while detailed implementation and reproducibility information is provided in Appendix C.

Table 3
Overview of Statistical Procedures Across Analyses

Analysis	Primary statistical approach	Additional or sensitivity analysis
Exploratory depth-level analysis	Kruskal-Wallis; Mann-Whitney U	IQR-based sensitivity analysis
RQ1: Recommendations and observable uptake	Permutation χ^2 ; Kruskal-Wallis; pairwise comparisons	Adjusted regression with conversation-level bootstrap
RQ2: Textual anthropomorphic cues	Mann-Whitney U	Adjusted regression with conversation-level bootstrap
RQ3: Conversation position and cue context	Permutation-based group comparisons	Interaction analysis and conversation-clustered analysis

Note. The table summarises the principal statistical procedures. IQR = interquartile range.

Effect sizes were reported alongside statistical tests so that the magnitude of observed differences could be considered separately from their statistical support. Multiple comparisons were controlled using the Holm procedure within predefined families of related tests (Holm, 1979). Statistical support was evaluated at $\alpha = .05$ after the applicable correction. Where both raw and Holm-adjusted p-values are reported, conclusions about statistical support are based on the Holm-adjusted p-values. Unless otherwise indicated, p-values and effect-size estimates were reported to three decimal places, with p-values below .001 reported as $p < .001$.

3.8 Research Quality

Research quality was supported through three complementary elements of the research design. First, construct validity was supported through alignment between the theoretical concepts and their operationalisation. Second, procedural reliability was supported through consistency in the manual coding process. Third, analytical strength was assessed through complementary quantitative checks. Together, these elements provided a systematic basis for translating the research questions into observable variables and applying the coding scheme consistently across the analytical sample. They also supported the evaluation of the stability and uncertainty of the statistical findings. The methodological and contextual limitations that remain despite these procedures, together with their implications for interpretation and generalisability, are discussed in Section 6.3.

Construct Alignment and Operationalisation

Construct alignment was supported by deriving the coding structure from the theoretical concepts developed in Chapter 2 and specifying how those concepts could be observed in the analysed conversations. Consistent with directed qualitative content analysis, the coding scheme was predominantly theory-led, while limited

inductive refinement was used to specify how the predefined constructs were represented in the sampled conversational data (Hsieh & Shannon, 2005).

Self-disclosure was operationalised through the theoretically distinct dimensions of depth, breadth, and volume rather than represented as a single composite measure (Carpenter & Greene, 2015; Omarzu, 2000). Recommendations, observable recommendation uptake, and textual anthropomorphic cues were likewise operationalised as explicitly defined observable variables corresponding to the concepts examined in RQ1 and RQ2. The resulting category definitions, counting rules, and decision criteria were documented in the codebook, providing a traceable link between the conceptual definitions established in Chapter 2 and the variables used in the quantitative analyses.

Coding Consistency and Quality Assurance

Coding consistency was supported through the standardised procedures described in Sections 3.6.2 and 3.6.3. Predefined definitions, construct-specific decision rules, and a fixed coding sequence provided a common basis for classifying all analytical <AI, user> turn pairs. These procedures supported consistent application of the coding framework across the analytical dataset.

Quality assurance further focused on procedural completeness and internal consistency. The completed coded dataset was checked before quantitative analysis for incomplete or incorrectly entered values and inconsistent field use. These checks reduced avoidable procedural inconsistencies before statistical analysis.

Analytical Quality and Sensitivity Checks

Analytical Quality was assessed through complementary procedures addressing alternative analytical specifications, dependence between observations, and sensitivity to selected analytical choices. For RQ1 and RQ2, adjusted regression analyses were conducted alongside the primary group comparisons. Uncertainty in the adjusted estimates was assessed through conversation-level bootstrapping, which preserved the grouping of turn pairs within conversations (Huang, 2018). For RQ3, the central conversation-position and interaction analyses were additionally re-estimated using conversation-clustered inference. This accounted for dependence among observations originating from the same conversation (Cameron & Miller, 2015).

Additional sensitivity analyses examined whether the exploratory depth-level patterns were disproportionately influenced by potential outliers and whether conversation-position patterns were consistent across conversation-length strata. Effect sizes were reported alongside statistical tests, confidence intervals were provided for the adjusted regression analyses, and Holm correction was applied within predefined families of related comparisons (Holm, 1979). Together, these procedures provided complementary information about the stability, magnitude, and uncertainty of the findings.

3.9 Ethical Considerations

This thesis analysed secondary data from previously donated ChatGPT conversation histories, and no new participants were recruited or contacted. Because these conversations may contain personal experiences, health information, emotional distress, relationship difficulties, and other sensitive disclosures, the data were treated as confidential and potentially identifiable throughout the study.

Ethical Approval and Informed Consent

The research was conducted under the required ethical approval from the TU Delft Human Research Ethics Committee. Participants in the original data-donation study provided explicit informed consent for their ChatGPT data to be used for research (Karnam et al., 2026).

Data Protection and Confidentiality

The conversation histories were treated as confidential research data. Health-related disclosures may constitute special-category personal data under Article 9 of the General Data Protection Regulation (GDPR) (Regulation (EU) 2016/679, 2016, Art. 9). Because pseudonymous conversation data may still permit indirect identification through combinations of personal circumstances, the data were handled as potentially identifiable throughout the research process.

Access was restricted to authorised researchers and supervisors. Raw and processed data, coding materials, and analytical outputs containing research data remained on a secured server and were not transferred to personal devices, external cloud services, or other unapproved environments. No raw conversation data or individual messages were provided to generative-AI systems or other external language-processing services. The use of AI and other digital tools is documented separately in Appendix E.

Reporting of Conversation Data

No raw conversation text, direct quotations, or participant-level identifiable information are reported in this thesis. Findings are presented through coded variables and aggregated statistical outputs. Any conversational examples used to illustrate the coding procedure are synthetic and were not copied, paraphrased, or adapted from donated conversations.

Data and Code Availability

The raw and processed ChatGPT conversation histories are not publicly available because they contain potentially sensitive and indirectly identifiable information and remain subject to restricted TU Delft access. No research dataset or analysis code associated with this MSc thesis will be deposited in a public repository. The final thesis will be made publicly available through the TU Delft Repository.

4 Results

This chapter reports the empirical results from the 2,639 analytical <AI, user> turn pairs derived from the 69 retained mental-health conversations. It first provides descriptive context for self-disclosure, recommendations, and anthropomorphic cues, before presenting the analyses addressing RQ1, RQ2, and RQ3. Detailed supplementary results are provided in Appendix D. The chapter focuses on observed patterns and statistical evidence, with brief interpretation of the results where appropriate.

4.1 Descriptive Overview and Exploratory Analysis

This section provides descriptive context for the subsequent research-question analyses by summarising self-disclosure outcomes, recommendation presence and uptake, and textual anthropomorphic cues. It also examines how breadth amount and volume vary across self-disclosure depth levels.

4.1.1 Self-Disclosure Outcomes

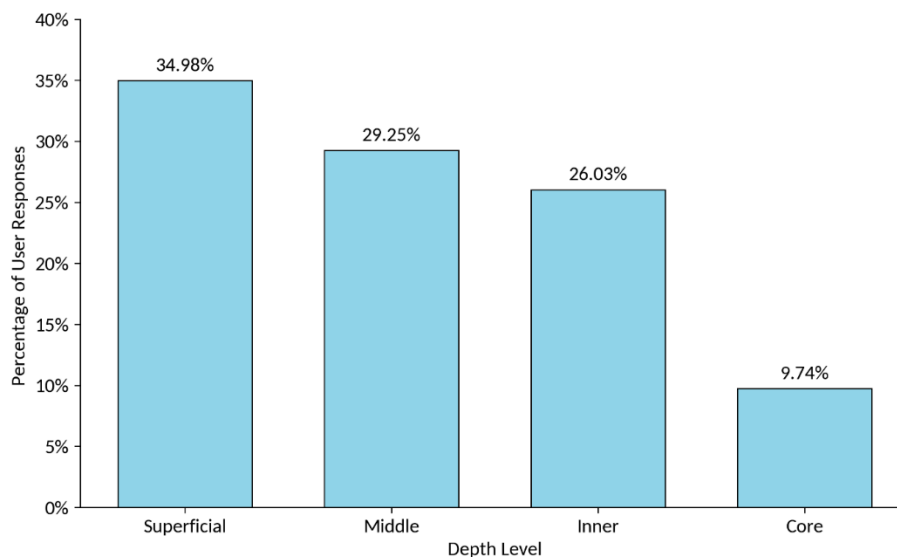
User responses were summarised across the three separately operationalised dimensions of self-disclosure: depth, breadth, and self-disclosure volume. The descriptive results for each dimension are reported below.

Depth Level

Figure 13 shows that user responses were more frequently classified at the lower depth levels. Overall, 64.23% of responses were classified as Superficial or Middle, compared with 35.77% classified as Inner or Core.

Figure 13

Distribution of Self-Disclosure Depth Levels Across User Responses



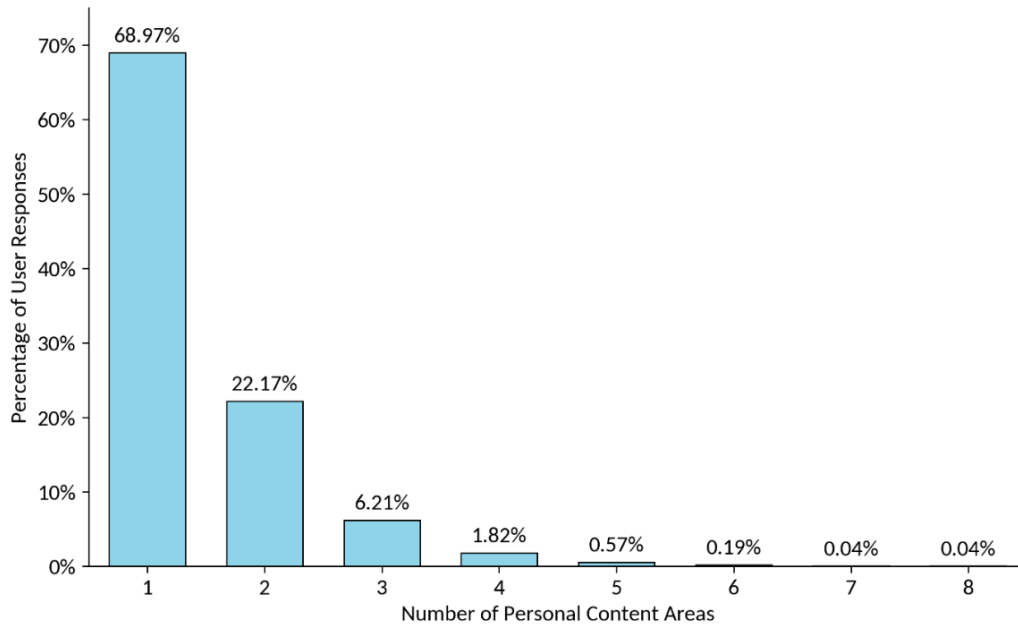
Note. Superficial includes responses with no identified self-disclosure, as defined in Section 3.6.1.

Breadth Amount

No user response had a breadth amount of 0. Mean breadth amount was 1.44 personal content areas per response, with a median of 1. As shown in Figure 14, the distribution was strongly concentrated at lower breadth amounts: 91.14% of responses contained one or two personal content areas, whereas 8.86% contained three or more.

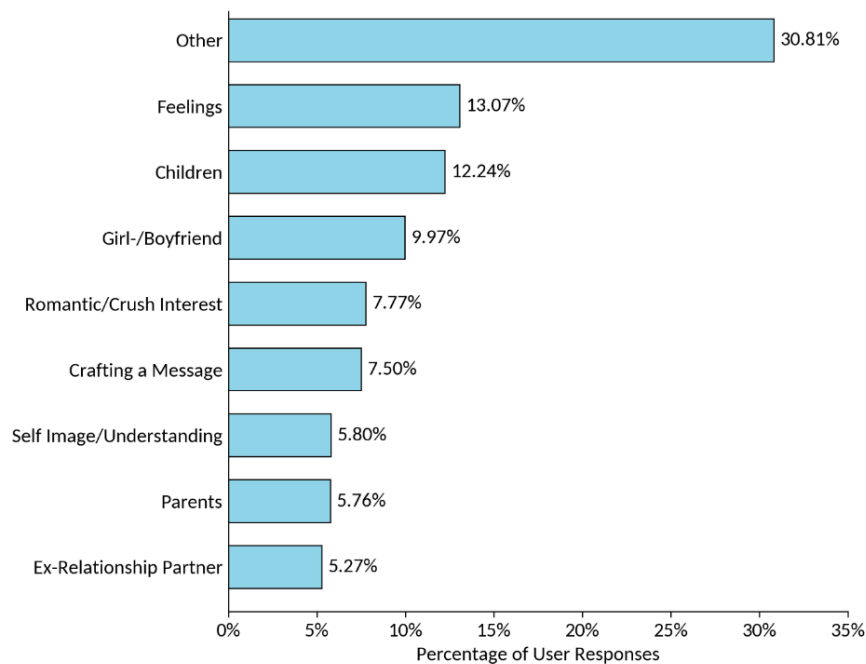
Figure 14

Distribution of Breadth Amount Across User Responses



Breadth Type

Figure 15 shows the breadth types occurring in at least 5% of analysed user responses. *Other* was the most frequently coded breadth type. Among the predefined categories, *Feelings* and *Children* occurred most frequently, while several other frequently occurring breadth types concerned personal relationships. The complete breadth-type distribution is reported in Appendix D.2, Figure D2.

Figure 15*Breadth Types Occurring in at Least 5% of User Responses*

Note. Percentages represent the proportion of user responses in which each breadth type was identified. Multiple breadth types could be coded within the same response; percentages therefore do not sum to 100%. *Other* represents personal content areas not captured by the predefined breadth types.

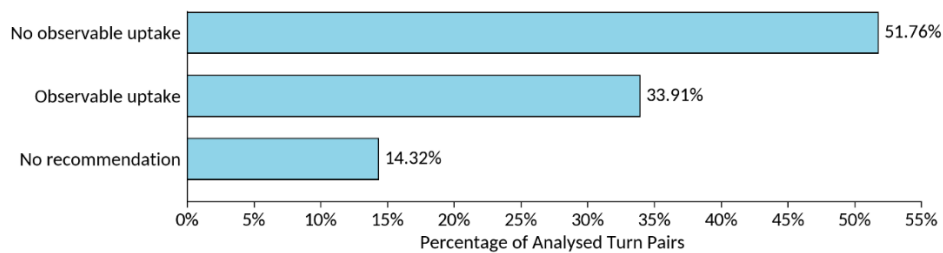
Volume

Volume was zero in 62.79% of user responses, whereas 37.21% contained at least one word coded as self-disclosing information. Across all responses, mean volume was 3.74 words, the median was 0, and the maximum was 96. The zero median relative to the higher mean and maximum indicates a strongly right-skewed distribution, with most responses containing little or no coded self-disclosing content and a smaller number containing substantially larger amounts.

Taken together, the descriptive distributions show that user responses were concentrated at lower depth levels and lower breadth amounts, while volume was zero in most responses. The next two subsections describe the two ChatGPT-side conversational features examined in RQ1 and RQ2: recommendations and textual anthropomorphic cues.

4.1.2 Recommendations and Observable Uptake

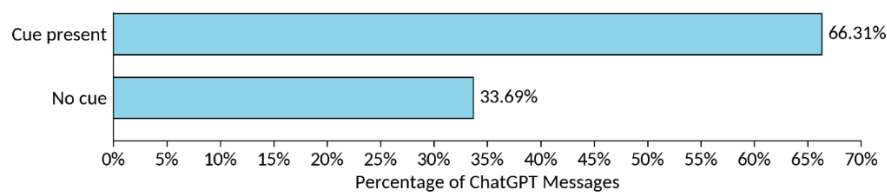
Recommendations were identified in 2,261 of the 2,639 analysed <AI, user> turn pairs (85.68%). As shown in Figure 16, recommendation without observable uptake was more common than recommendation with observable uptake, while no recommendation constituted the smallest group. Among turn pairs containing at least one recommendation, 39.58% showed observable uptake in the immediately following user response and 60.42% did not.

Figure 16*Distribution of Recommendation Groups Across Analysed Turn Pairs*

Note. Recommendation groups are mutually exclusive. Observable uptake refers only to uptake identified in the immediately following user response. Percentages may not sum to exactly 100% due to rounding.

4.1.3 Textual Anthropomorphic Cues

Textual anthropomorphic cues were identified in 1,750 of the 2,639 analysed ChatGPT messages. As shown in Figure 17, messages containing at least one textual anthropomorphic cue constituted approximately two-thirds of the analysed messages.

Figure 17*Presence of Textual Anthropomorphic Cues Across ChatGPT Messages*

Note. Cue present indicates that at least one textual anthropomorphic cue was identified in the ChatGPT message.

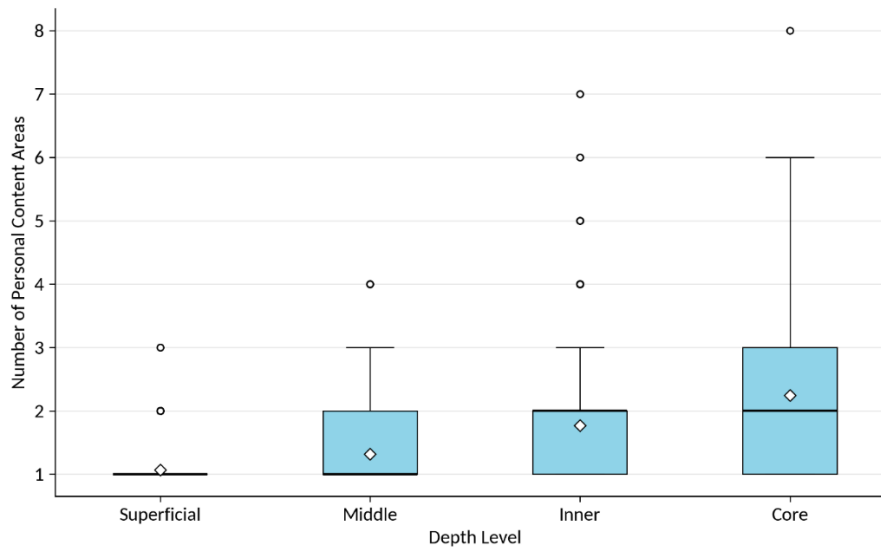
4.1.4 Breadth and Volume Across Depth Levels

Having described the distributions of the disclosure outcomes and conversational features, the exploratory analysis next examined how breadth amount and volume varied across the four depth levels: Superficial, Middle, Inner, and Core.

Breadth Amount Across Depth Levels

Breadth amount increased across the four depth levels. Mean breadth amount rose from 1.07 personal content areas for Superficial responses to 2.25 for Core responses. The median remained at one personal content area for Superficial, Middle, and Inner responses and increased to two for Core responses. Figure 18 also shows greater dispersion at the higher depth levels.

Figure 18
Breadth Amount Across Depth Levels

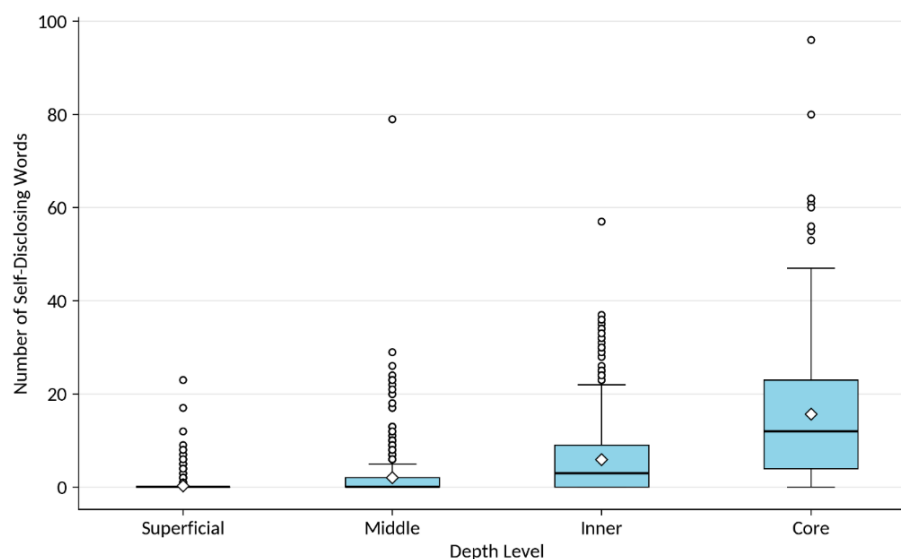


Note. White diamonds indicate group means. Hollow circles indicate observations beyond the boxplot whiskers.

Volume Across Depth Levels

Volume showed the same directional pattern as breadth amount, increasing across the four depth levels. Mean volume increased from 0.25 words for Superficial responses to 2.01 for Middle, 5.94 for Inner, and 15.63 for Core responses. Median volume was 0 words for Superficial and Middle responses, compared with 3 words for Inner and 12 words for Core responses. Figure 19 also shows greater dispersion at the higher depth levels, particularly for Inner and Core responses.

Figure 19
Volume Across Depth Levels



Note. White diamonds indicate group means. Hollow circles indicate observations beyond the boxplot whiskers.

Statistical Comparison Across Depth Levels

As shown in Table 4, separate Kruskal-Wallis tests showed statistically supported overall differences across depth levels for both breadth amount and volume (both $p < .001$). The effect-size estimate was larger for volume ($\epsilon^2 = .383$) than for breadth amount ($\epsilon^2 = .246$).

Table 4
Breadth Amount and Volume Across Depth Levels

Measure	Breadth amount	Volume
Panel A. Primary Analysis		
User responses (N)	2,639	2,639
Kruskal-Wallis H	652.12	1012.52
p	$< .001$	$< .001$
ϵ^2	.246	.383
Panel B. IQR-based Sensitivity Analysis		
N after excluding IQR-flagged observations	2,547	2,445
Kruskal-Wallis $H(3)$ after exclusion	755.07	1177.82
p after exclusion	$< .001$	$< .001$
ϵ^2 after exclusion	.296	.481

Note. The sensitivity analysis repeated the overall Kruskal-Wallis tests after excluding observations flagged within each depth level using the IQR criterion described in Section 3.7.2. ϵ^2 = epsilon squared.

Follow-up pairwise comparisons showed that all six depth-level comparisons remained statistically supported for both breadth amount and volume after Holm correction (all Holm-adjusted $p < .001$). In every comparison, the higher depth level had greater breadth amount and volume than the lower depth level. Complete test statistics and rank-biserial effect sizes are reported in Appendix D.3, Table D1.

The IQR-based sensitivity analysis reproduced the overall depth-level differences for both breadth amount and volume after IQR-flagged observations were excluded (Table 4), indicating that the overall pattern remained after their exclusion.

4.2 RQ1: Recommendation Presence, Observable Uptake, and Subsequent Self-Disclosure

RQ1 examined whether subsequent user self-disclosure differed according to recommendation presence and observable uptake:

How do the depth, breadth, and volume of user self-disclosure differ between responses following no recommendation, a recommendation without observable uptake, and a recommendation with observable uptake?

The results are presented in four steps. First, depth, breadth amount, and volume are compared across the three recommendation groups: no recommendation, recommendation without observable uptake, and recommendation with observable uptake. Where overall differences were statistically supported, pairwise comparisons identify which groups differed. The adjusted analyses then examine the recommendation-group associations conditional on turn number and textual anthropomorphic-cue presence, including the direct contrast between recommendations with and without observable uptake. Finally, the supplementary analysis describes variation across the five recommendation types.

4.2.1 Primary Three-Group Comparison

Table 5 summarises breadth amount and volume across the three recommendation groups. Figure 20 subsequently shows how the depth distributions differed across these groups.

Table 5

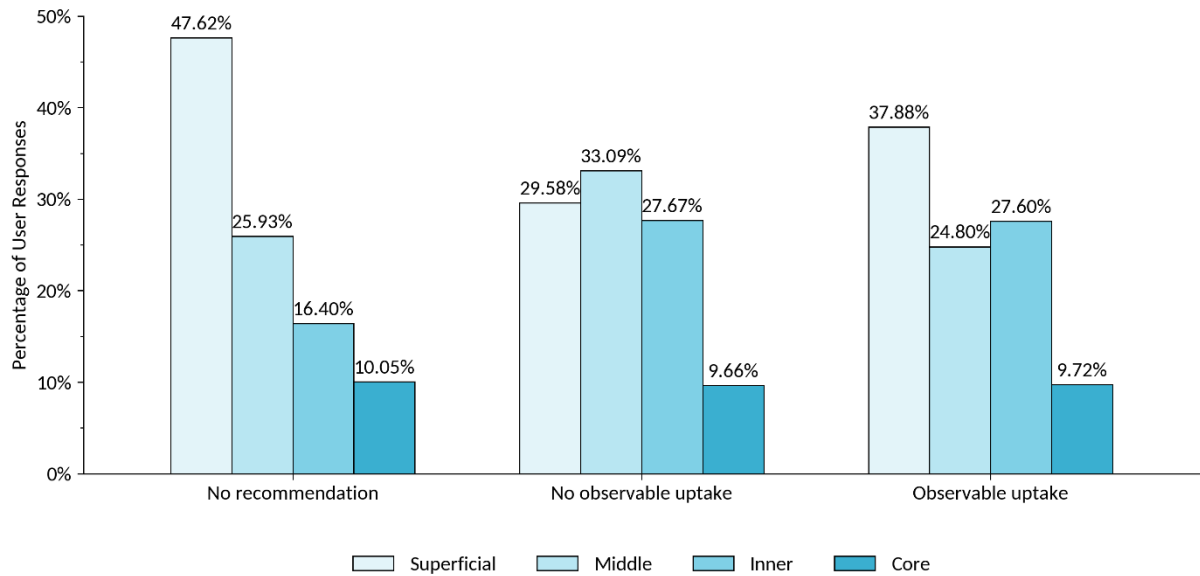
Breadth Amount and Volume by Recommendation Group

Recommendation group	Turn pairs (n)	Breadth amount mean	Breadth amount median	Volume mean	Volume median
No recommendation	378	1.39	1.0	3.43	0.0
Recommendation without observable uptake	1366	1.51	1.0	3.80	0.0
Recommendation with observable uptake	895	1.35	1.0	3.79	0.0

Note. Breadth amount is measured in personal content areas; volume is measured in self-disclosing words.

Breadth amount was highest following recommendations without observable uptake, whereas the other two groups showed lower and relatively similar values. Median breadth amount was one personal content area in all three groups. Volume varied less across the recommendation groups: the two recommendation groups had nearly identical means, while the no-recommendation group had a slightly lower mean. Median volume was zero in all three groups.

The depth distributions showed a different descriptive pattern. As shown in Figure 20, responses following no recommendation were more frequently classified as Superficial and less frequently classified as Inner than responses in either recommendation group. The two recommendation groups showed nearly identical proportions of Inner and Core responses, while their clearest descriptive differences occurred at the Superficial and Middle levels.

Figure 20*Distribution of Self-Disclosure Depth Levels by Recommendation Group*

Note. Percentages were calculated separately within each recommendation group and sum to approximately 100% within each group due to rounding.

Overall differences across recommendation groups were statistically supported for breadth amount, volume, and the complete depth distribution after Holm correction (Table 6). Because these omnibus tests establish only that at least one recommendation group differed from another, Section 4.2.2 examines the specific pairwise differences.

Table 6*Overall Differences Across Recommendation Groups*

Outcome	Test	Test statistic	Raw p	Holm-adjusted p	Effect size
Breadth amount	Kruskal-Wallis	$H = 33.313$	$< .001$	$< .001$	$\epsilon^2 = .012$
Volume	Kruskal-Wallis	$H = 10.865$	$.004$	$.004$	$\epsilon^2 = .003$
Depth distribution	Permutation chi-square	$\chi^2 = 60.944$	$< .001$	$< .001$	Cramér's $V = .107$

Note. Holm correction was applied across the three RQ1 overall outcome tests. ϵ^2 = epsilon squared.

4.2.2 Pairwise Group Comparisons

Because the overall three-group tests were statistically supported for depth, breadth amount, and volume distribution, pairwise comparisons were conducted to identify which recommendation groups differed. Table 7 reports the comparisons that remained statistically supported after Holm correction. The complete set of pairwise results, including comparisons that were not statistically supported, is reported in Appendix D.4.

Table 7*Pairwise Comparisons of Self-Disclosure Across Recommendation Groups*

Outcome and group comparison	Test statistic	Holm-adjusted <i>p</i>	Effect size
Panel A. Depth distribution			
Recommendation without observable uptake – Recommendation with observable uptake	$\chi^2 = 23.793$	< .001	Cramér's <i>V</i> = .103
Recommendation without observable uptake – No recommendation	$\chi^2 = 48.554$	< .001	Cramér's <i>V</i> = .167
Recommendation with observable uptake – No recommendation	$\chi^2 = 20.072$	< .001	Cramér's <i>V</i> = .126
Panel B. Breadth amount			
Recommendation without observable uptake – Recommendation with observable uptake	$U = 679,467.50$	< .001	$r = .112$
Recommendation without observable uptake – No recommendation	$U = 280,879.50$	.004	$r = .088$
Panel C. Volume			
Recommendation without observable uptake – Recommendation with observable uptake	$U = 650,716.50$	.009	$r = .065$

Note. Only comparisons that remained statistically supported after Holm correction are shown. Holm correction was applied separately across the three pairwise comparisons within each outcome. Depth-distribution comparisons used permutation chi-square tests with Cramér's *V* as the effect-size measure. Breadth amount and volume comparisons used Mann-Whitney *U* tests with effect-size *r*.

The complete depth distribution differed between all three recommendation-group pairs. The largest pairwise effect-size estimate occurred between recommendation without observable uptake and no recommendation, followed by recommendation with observable uptake versus no recommendation and recommendation without versus with observable uptake (Table 7). As shown in Figure 20, these differences were mainly reflected in the proportions of responses classified as Superficial, Middle, and Inner, whereas the proportion classified as Core was similar across the three groups.

For breadth amount, recommendation without observable uptake differed from both recommendation with observable uptake and no recommendation. Recommendation with observable uptake did not differ statistically from no recommendation after Holm correction. Together with the descriptive results in Section 4.2.1, this indicates that the highest breadth amount occurred following recommendations without observable uptake rather than recommendations showing observable uptake.

For volume, only the comparison between recommendation without observable uptake and recommendation with observable uptake remained statistically supported after Holm correction. Neither recommendation group differed statistically from no recommendation. Despite this pairwise result, the two recommendation groups had nearly identical mean volumes in Table 5, so the Mann-Whitney result should not be interpreted as evidence of a difference between their means.

Section 4.2.3 next examines whether these recommendation-group associations remain statistically supported when turn number and textual anthropomorphic-cue presence are included in the adjusted analyses.

4.2.3 Adjusted Analysis

The adjusted analyses re-examined the recommendation-group associations conditional on turn number and textual anthropomorphic-cue presence. Table 8 reports the adjusted contrasts for ordered depth, breadth amount, and volume, including the direct comparison between recommendations with and without observable uptake.

Table 8
Adjusted Recommendation-Group Contrasts

Comparison	Coefficient (<i>B</i>)	95% bootstrap CI	Raw bootstrap <i>p</i>	Holm- adjusted <i>p</i>
Panel A. Ordered depth				
Without uptake vs. No recommendation	.270	[−.034, .541]	.081	.244
With uptake vs. No recommendation	.184	[−.139, .451]	.256	.512
With uptake vs. Without uptake	−.086	[−.369, .179]	.538	.538
Panel B. Breadth amount				
Without uptake vs. No recommendation	.103	[−.090, .264]	.287	.573
With uptake vs. No recommendation	−.055	[−.231, .091]	.480	.573
With uptake vs. Without uptake	−.158	[−.293, −.015]	.029	.087
Panel C. Volume				
Without uptake vs. No recommendation	.380	[−1.068, 1.720]	.616	1.000
With uptake vs. No recommendation	.286	[−1.246, 1.787]	.717	1.000

Comparison	Coefficient (<i>B</i>)	95% bootstrap CI	Raw bootstrap <i>p</i>	Holm- adjusted <i>p</i>
With uptake vs. Without uptake	-.094	[-1.404, 1.288]	.912	1.000

Note. Coefficients are from separate linear regression models including turn number and textual anthropomorphic-cue presence as additional predictors. No recommendation is the reference category for the first two contrasts within each outcome; the with- versus without-observable-uptake estimates are linear contrasts derived from the same fitted models. Confidence intervals are unadjusted 95% percentile bootstrap intervals based on 10,000 conversation-level bootstrap resamples. Raw bootstrap *p*-values were Holm-adjusted separately within each outcome across the three recommendation-group contrasts. *B* represents the unstandardized recommendation-group coefficient or contrast.

After Holm correction, none of the adjusted recommendation-group contrasts were statistically supported (Table 8). Neither recommendation without observable uptake nor recommendation with observable uptake differed statistically from the no-recommendation reference group for ordered depth, breadth amount, or volume.

For the direct comparison between the two recommendation groups, breadth amount was the only outcome with a raw bootstrap *p* below .05 ($B = -.158$, raw bootstrap *p* = .029), but this contrast was not statistically supported after Holm correction (Holm-adjusted *p* = .087). The corresponding direct contrasts for ordered depth and volume were also not statistically supported.

Thus, the statistically supported recommendation-group differences observed in the primary analyses were not retained in the adjusted analyses after inclusion of turn number and textual anthropomorphic-cue presence and correction for multiple testing. The final RQ1 subsection examines whether the descriptive patterns varied across recommendation types.

4.2.4 Descriptive Differences Across Recommendation Types

As a supplementary analysis, the five recommendation types were examined descriptively. Because recommendation types could co-occur within the same ChatGPT message and differed substantially in frequency, the results were not treated as independent inferential comparisons. Complete results are reported in Appendix D.5.

Observable uptake differed across recommendation types. Suggestive Questions and Suggestive Prompts occurred more often without observable uptake, whereas Follow-up Questions occurred more often with observable uptake. Reflective Questions showed a smaller difference in favour of observable uptake, while Ongoing Support Offers showed a more balanced uptake distribution.

The corresponding depth, breadth amount, and volume patterns were not uniform across recommendation types. Observable uptake coincided with relatively greater proportions of Inner and/or Core responses for Follow-up Questions, Reflective Questions, and Ongoing Support Offers, but not for Suggestive Questions or

Suggestive Prompts. Breadth amount and volume likewise differed in direction across recommendation types, with observable uptake associated descriptively with higher values for some types and equal or lower values for others.

Overall, the supplementary type-level analysis did not show a consistent descriptive pattern in which observable uptake corresponded to greater depth, breadth amount, or volume across all recommendation types.

4.3 RQ2: Textual Anthropomorphic Cues and Subsequent Self-Disclosure

RQ2 examined whether subsequent user self-disclosure differed according to the presence of textual anthropomorphic cues in the preceding ChatGPT message:

How do the depth, breadth, and volume of user self-disclosure differ between responses following ChatGPT messages with and without textual anthropomorphic cues?

The RQ2 results are presented in three parts. First, depth, breadth amount, and volume are compared between responses following ChatGPT messages with and without textual anthropomorphic cues. The adjusted analysis then examines these associations conditional on turn number and recommendation presence. Finally, the supplementary analysis describes variation across the seven textual anthropomorphic-cue types.

4.3.1 Primary Two-Group Comparison

Table 9 summarises breadth amount and volume for responses following ChatGPT messages with and without textual anthropomorphic cues. Breadth amount was descriptively slightly higher following messages containing at least one cue, while the median was one personal content area in both groups. Volume showed a similarly small descriptive difference, while the median was zero in both groups.

Table 9

Breadth Amount and Volume by Textual Anthropomorphic-Cue Presence

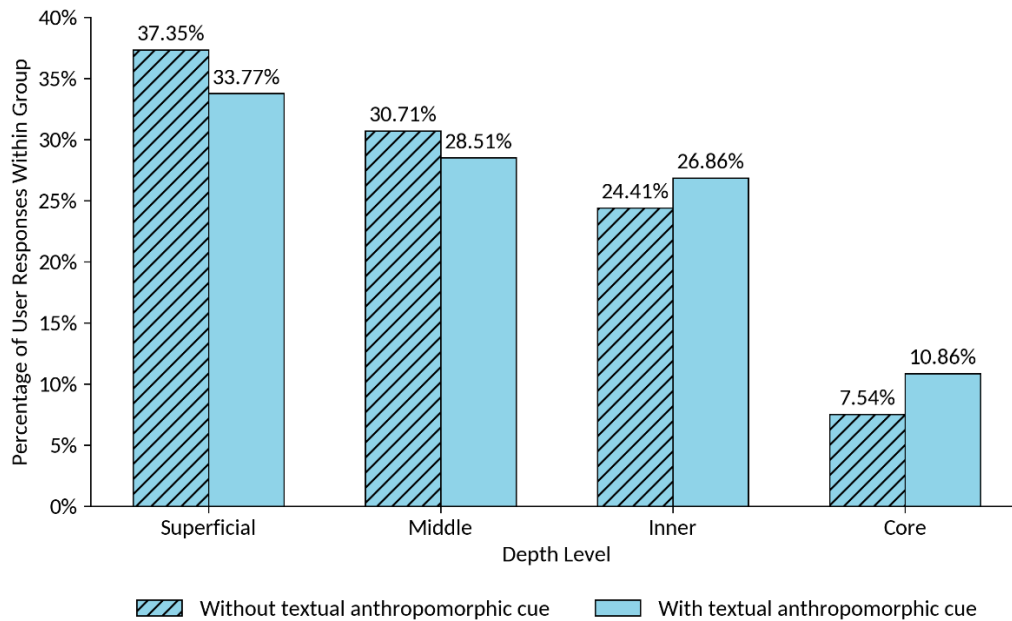
Textual anthropomorphic-cue presence	Turn pairs (n)	Breadth amount mean	Breadth amount median	Volume mean	Volume median
Without textual anthropomorphic cue	889	1.37	1.00	3.50	0.00
With textual anthropomorphic cue	1,750	1.47	1.00	3.87	0.00

Note. Breadth amount is measured in personal content areas; volume is measured in self-disclosing words.

Figure 21 shows the complete four-category depth distribution for the same two groups. Responses following messages containing at least one textual anthropomorphic cue were less frequently classified as Superficial or Middle and more frequently classified as Inner or Core than responses following messages without a cue.

Figure 21

Distribution of Depth Levels by Textual Anthropomorphic-Cue Presence



Note. Percentages were calculated within each cue-presence group and sum to approximately 100% within each group due to rounding.

Table 10 reports the primary inferential comparisons between responses following ChatGPT messages with and without textual anthropomorphic cues. Mann-Whitney U tests were used for ordered depth, breadth amount, and volume. For depth, the ordering from Superficial to Core was used in the inferential comparison, while Figure 21 retains the complete four-category distribution descriptively.

Table 10

Primary Two-Group Comparisons by Textual Anthropomorphic-Cue Presence

Outcome	U	z	Holm-adjusted p	Rank-biserial r
Breadth amount	817,975.00	2.666	.015	.052
Ordered depth	829,777.50	2.935	.010	.067
Volume	782,882.50	0.312	.755	.006

Note. Mann-Whitney U tests compare responses following messages without a textual anthropomorphic cue with responses following messages containing at least one textual anthropomorphic cue. Holm correction was applied across the three RQ2 outcome comparisons.

After Holm correction, differences between the two cue-presence groups were statistically supported for ordered depth and breadth amount, but not for volume

(Table 10). Both statistically supported differences had small effect-size estimates. The primary analysis therefore showed dimension-specific differences according to textual anthropomorphic-cue presence rather than a consistent difference across all three self-disclosure outcomes. Section 4.3.2 next examines whether the cue-presence associations remain statistically supported conditional on turn number and recommendation presence.

4.3.2 Adjusted Analysis

The adjusted analyses re-examined the association between textual anthropomorphic-cue presence and each outcome conditional on turn number and recommendation presence. Separate models were estimated for ordered depth, breadth amount, and volume. Table 11 reports the focal cue-presence coefficient from each model.

Table 11

Adjusted Associations of Textual Anthropomorphic-Cue Presence

Outcome	Coefficient (<i>B</i>)	95% bootstrap CI	Raw bootstrap <i>p</i>	Holm-adjusted <i>p</i>
Ordered depth	.098	[−.082, .251]	.275	.550
Breadth amount	.101	[.012, .180]	.023	.070
Volume	.444	[−.516, 1.371]	.414	.550

Note. No textual anthropomorphic cue is the reference category. Coefficients are from separate linear regression models including turn number and recommendation presence as additional predictors. Confidence intervals are unadjusted 95% percentile bootstrap intervals based on 10,000 conversation-level bootstrap resamples. Raw bootstrap *p*-values were Holm-adjusted across the three focal cue-presence coefficients. *B* represents the unstandardized regression coefficient for textual anthropomorphic-cue presence.

After Holm correction, none of the adjusted cue-presence associations were statistically supported (Table 11). For breadth amount, the raw bootstrap *p* was .023 and the unadjusted 95% bootstrap confidence interval excluded zero, but the association was not statistically supported after Holm correction (*p* adjusted = .070). The adjusted associations with ordered depth and volume were also not statistically supported.

Thus, whereas the primary two-group comparison showed statistically supported differences in ordered depth and breadth amount, neither association remained statistically supported in the adjusted analysis after multiple-testing correction. Section 4.3.3 next examines whether the descriptive patterns varied across textual anthropomorphic-cue types.

4.3.3 Descriptive Variation Across Textual Anthropomorphic Cue Types

As a supplementary analysis, the seven textual anthropomorphic-cue types were examined descriptively. Because multiple cue types could occur within the same ChatGPT message and cue types differed substantially in frequency, the comparisons were not treated as independent inferential tests. Complete results are reported in Appendix D.6.

The descriptive patterns varied across cue types. Human-like Intention showed the largest proportion of Core responses and the highest mean breadth amount and volume, while Emotional Attribution also showed comparatively high breadth amount and volume. Collaborative Agency showed a comparatively large proportion of Core responses and relatively high mean volume. In contrast, Human-like Judgment and Belief Framing contained comparatively larger proportions of Middle responses and lower mean volume. Human-like Intention was relatively uncommon, with only 61 observations, so its descriptive estimates should be interpreted in light of the smaller category size.

Overall, the type-level analysis showed that the descriptive patterns differed across textual anthropomorphic-cue types. Human-like Intention showed the highest values on several descriptive indicators, but the cue types differed substantially in frequency and could co-occur within the same message. These supplementary results therefore describe type-level variation rather than independent cue-type associations with subsequent self-disclosure.

4.4 RQ3: Conversation Position, Cue Context, and Self-Disclosure

RQ3 examined how self-disclosure varied across relative conversation position and whether these patterns differed across recommendation and textual-anthropomorphic-cue contexts:

How does conversation position relate to the depth, breadth, and volume of user self-disclosure across different recommendation and textual-anthropomorphic-cue contexts?

The RQ3 results are presented in five parts. First, the primary conversation-position analysis compares depth, breadth amount, and volume across the four conversation quartiles. Second, the conversation-length sensitivity analysis examines whether the pooled quartile pattern is also observed across conversations of different lengths. Third, the cue-context analysis examines self-disclosure across the four mutually exclusive cue groups and across conversation quartiles within these groups. Fourth, the conversation quartile $\times$ cue-group interaction analysis tests whether the conversation-position pattern differs between cue groups. Finally, the conversation-level cluster analysis re-examines the central conversation-position and interaction associations while accounting for dependence between turn pairs originating from the same conversation.

4.4.1 Self-Disclosure Across Conversation Quartiles

Conversation quartiles represent fixed equal-width intervals of relative conversation position; consequently, the number of analysed turn pairs and contributing conversations differs across quartiles. Table 12 summarises breadth amount and volume across the four conversation quartiles.

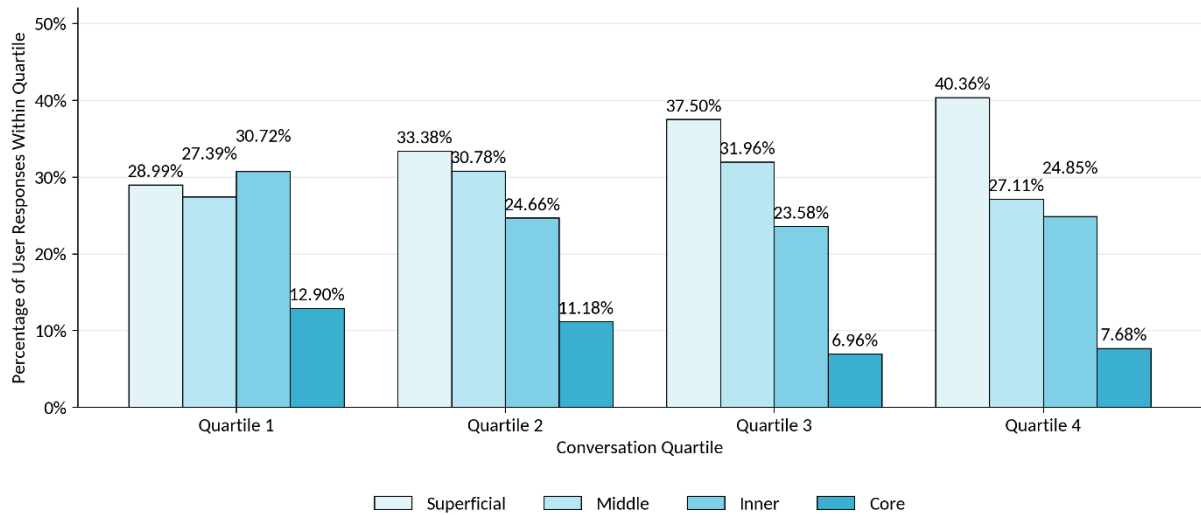
Table 12
Breadth Amount and Volume by Conversation Quartile

Measure	Quartile 1	Quartile 2	Quartile 3	Quartile 4
Analysed turn pairs	690	653	632	664
Contributing conversations (<i>n</i>)	69	66	64	67
Breadth amount mean	1.52	1.41	1.39	1.42
Breadth amount median	1.00	1.00	1.00	1.00
Volume mean	4.36	3.96	3.23	3.37
Volume median	0.00	0.00	0.00	0.00

Note. Conversation counts indicate how many of the 69 sampled conversations contributed at least one analysed turn pair to each quartile. Because conversation quartiles represent fixed intervals of relative position and conversations differed in length, the numbers of turn pairs and contributing conversations need not be equal across quartiles.

Breadth amount and volume were both highest in Quartile 1, decreased through Quartile 3, and increased slightly in Quartile 4. Median breadth amount remained one personal content area and median volume remained zero across all four quartiles. The descriptive patterns were therefore not strictly monotonic across conversation position.

Figure 22 shows a shift toward a larger proportion of responses classified as Superficial across later conversation quartiles. Inner and Core responses were most prevalent in Quartile 1 and occurred less frequently in the later quartiles, although both increased slightly in Quartile 4 relative to Quartile 3. The Middle category showed a less consistent pattern across conversation position.

Figure 22*Depth Distribution Across Conversation Quartiles*

Note. Percentages were calculated within each conversation quartile and sum to 100% across the four depth levels.

Overall differences across conversation quartiles were statistically supported for depth distribution, breadth amount, and volume after Holm correction (Table 13).

Table 13*Overall Differences Across Conversation Quartiles*

Outcome	Test statistic	Holm-adjusted <i>p</i>	Effect size
Depth distribution	$\chi^2(9) = 42.668$	< .001	Cramér's <i>V</i> = .073
Breadth amount	$H(3) = 9.537$	.023	$\epsilon^2 = .003$
Volume	$H(3) = 24.346$	< .001	$\epsilon^2 = .008$

Note. Depth was compared using a permutation chi-square test with Cramér's *V* as the effect-size measure. Breadth amount and volume were compared using permutation Kruskal-Wallis tests with ϵ^2 as the effect-size measure. Holm correction was applied across the three RQ3 overall outcome tests.

Because the overall quartile differences were statistically supported for all three outcomes, follow-up pairwise comparisons were conducted to identify which quartiles differed. Table 14 reports the comparisons that remained statistically supported after Holm correction. The complete pairwise results are provided in Appendix D.7.

Table 14
Statistically Supported Pairwise Comparisons Between Conversation Quartiles

Outcome	Comparison	Test statistic	Holm-adjusted <i>p</i>	Effect size
Depth distribution	Q1–Q3	$\chi^2(3) = 27.293$	< .001	Cramér's <i>V</i> = .144
Depth distribution	Q1–Q4	$\chi^2(3) = 25.784$	< .001	Cramér's <i>V</i> = .138
Breadth amount	Q1–Q2	$U = 241,014.00$	.041	$r = .070$
Volume	Q1–Q2	$U = 242,376.00$	.028	$r = .076$
Volume	Q1–Q3	$U = 246,619.00$	< .001	$r = .131$
Volume	Q1–Q4	$U = 250,578.00$	.004	$r = .094$

Note. Only comparisons that remained statistically supported after Holm correction are shown. Holm correction was applied separately across the six pairwise comparisons within each outcome. Depth distributions were compared using permutation chi-square tests with Cramér's *V*; breadth amount and volume were compared using permutation Mann-Whitney *U* tests with rank-biserial *r*.

Every pairwise comparison that remained statistically supported involved Quartile 1. Breadth amount differed only between Q1 and Q2, volume differed between Q1 and each later quartile, and the complete depth distribution differed between Q1 and Q3 and between Q1 and Q4. No comparison exclusively among Q2–Q4 remained statistically supported after Holm correction.

The statistically supported pairwise differences were therefore concentrated in the earliest conversation quartile. However, the absence of statistically supported differences exclusively among Q2–Q4 indicates that the pattern should not be described as a uniform decline across conversation position.

4.4.2 Conversation-Length Sensitivity Analysis

The conversation-length sensitivity analysis examined whether the pooled conversation-position pattern reported in Section 4.4.1 was also observed across conversations of different lengths. The primary quartile analysis was repeated within each of the six predefined conversation-length strata. Table 15 summarises the outcomes for which overall quartile differences remained statistically supported after Holm correction. Complete stratum-specific descriptive statistics, overall tests, and pairwise comparisons are reported in Appendix D.8.

Table 15
Conversation-Length Sensitivity Analysis of Quartile Differences

Conversation length	Conversations	Analysed turn pairs (<i>n</i>)	Statistically supported quartile differences
2–10 turns	7	19	None
11–20 turns	16	191	None
21–30 turns	11	254	None
31–40 turns	14	458	None
41–50 turns	7	296	Depth
51+ turns	14	1,421	Depth, Breadth, Volume

Note. Outcomes are listed when the overall comparison across the four conversation quartiles remained statistically supported after Holm correction. The number of available turn pairs differed substantially across the conversation-length strata.

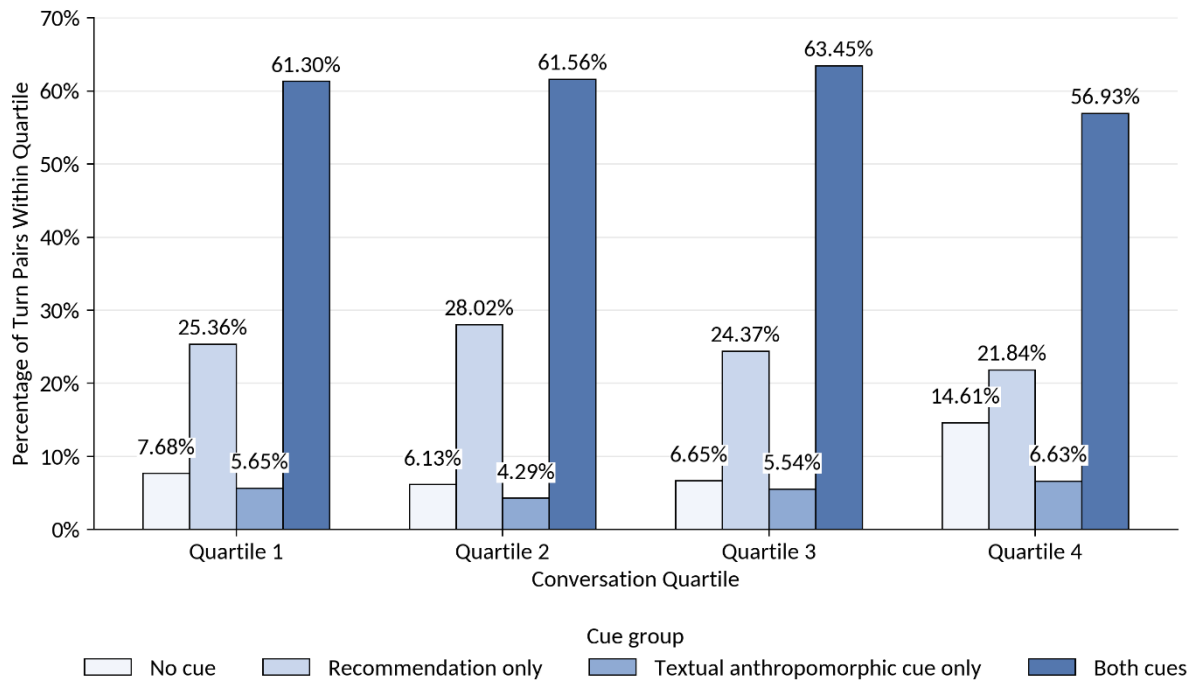
Statistical support for conversation-position differences varied across conversation lengths. No outcome showed a statistically supported overall quartile difference in the four shortest strata. In the 41–50-turn stratum, only the depth distribution differed across quartiles, whereas in the 51+ stratum, differences were statistically supported for depth distribution, breadth amount, and volume.

The sensitivity analysis therefore showed that the pooled Quartile-1-centred pattern was most evident in the longer-conversation strata, particularly the 51+ stratum, but was not observed consistently across all conversation lengths.

4.4.3 Self-Disclosure by Cue Group and Conversation Quartile

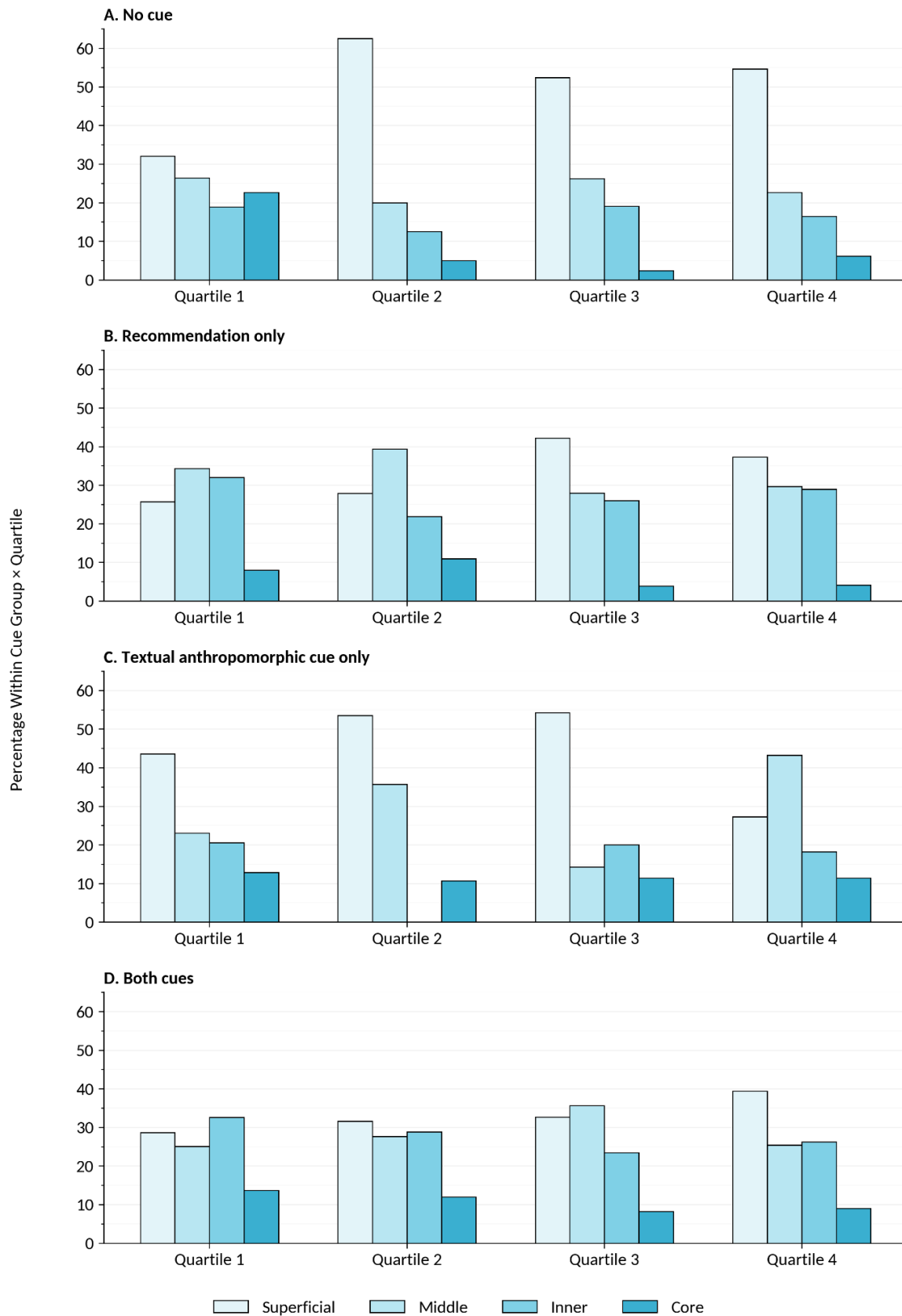
The cue-context analysis examined self-disclosure across four mutually exclusive groups based on the focal conversational features present in the preceding ChatGPT message: No cue, Recommendation only, Textual anthropomorphic cue only, and Both cues.

Figure 23 shows that Both cues was the largest cue group in every conversation quartile, followed by Recommendation only. Cue-group composition was relatively similar across Quartiles 1–3, whereas Quartile 4 contained a larger proportion of No cue turn pairs and a smaller proportion of Both cues turn pairs.

Figure 23*Cue-Group Distribution Across Conversation Quartiles*

Note. Cue groups are mutually exclusive. No cue indicates that neither a recommendation nor a textual anthropomorphic cue was identified in the preceding ChatGPT message; Recommendation only indicates recommendation presence without a textual anthropomorphic cue; Textual anthropomorphic cue only indicates cue presence without a recommendation; Both cues indicates the presence of both features. Percentages were calculated within each conversation quartile.

Figure 24 shows that depth patterns across conversation quartiles differed descriptively between cue groups. In the No cue group, the proportion of Superficial responses increased from 32.08% in Quartile 1 to 52.38–62.50% across Quartiles 2–4, while the proportion of Core responses decreased from 22.64% to 2.38–6.19%. The Both cues group showed a smaller shift, with Superficial responses becoming more prevalent and Core responses less prevalent in Quartile 4 than in Quartile 1. The Recommendation only and Textual anthropomorphic cue only groups showed less consistent patterns across conversation position. Complete numerical depth distributions are reported in Appendix D.9.

Figure 24*Depth Distribution Across Conversation Quartiles by Cue Group*

Note. Percentages were calculated within each cue-group × conversation-quartile combination and sum to approximately 100% across the four depth levels due to rounding.

Table 16 presents mean breadth amount and volume across the same cue-group × conversation-quartile combinations.

Table 16

Mean Breadth Amount and Volume Across Conversation Quartiles by Cue Group

Quartile	No cue	Recommendation only	Anthropomorphism only	Both cues
Panel A. Breadth amount				
Q1	1.38	1.42	1.36	1.59
Q2	1.48	1.31	1.39	1.46
Q3	1.36	1.36	1.40	1.40
Q4	1.35	1.41	1.45	1.44
Panel B. Volume				
Q1	4.19	4.25	4.00	4.47
Q2	3.92	3.61	2.46	4.24
Q3	2.88	2.96	1.80	3.50
Q4	3.21	3.02	4.52	3.41

Note. Values are group means. Breadth amount is measured in personal content areas; volume is measured in self-disclosing words.

Breadth amount did not show a consistent directional pattern across conversation quartiles within the four cue groups. Mean volume decreased from Quartile 1 to Quartile 3 in all four cue groups. Between Quartiles 3 and 4, volume increased in the No cue, Recommendation only, and Textual anthropomorphic cue only groups, while decreasing slightly in the Both cues group.

To assess which of these descriptive differences were statistically supported, two sets of secondary comparisons were conducted: differences across conversation quartiles within each cue group and differences across cue groups within each conversation quartile. Table 17 reports the comparisons that remained statistically supported after Holm correction. Complete results are reported in Appendix D.10.

Table 17*Statistically Supported Cue-Context Comparisons Across Quartiles and Cue Groups*

Cue group/quartile	Outcome	Test statistic	Holm-adjusted p	Effect size
Panel A. Quartile Differences Within Cue Groups				
No cue	Depth distribution	$\chi^2(9) = 20.807$	.031	Cramér's $V = .173$
Recommendation only	Depth distribution	$\chi^2(9) = 24.834$	.009	Cramér's $V = .112$
Both cues	Depth distribution	$\chi^2(9) = 32.031$	< .001	Cramér's $V = .082$
Both cues	Breadth amount	$H(3) = 13.456$	.016	$\epsilon^2 = .007$
Both cues	Volume	$H(3) = 15.817$	.005	$\epsilon^2 = .008$
Panel B. Cue-Group Differences Within Quartiles				
Q1	Depth distribution	$\chi^2(9) = 19.583$	.035	Cramér's $V = .097$
Q1	Breadth amount	$H(3) = 14.245$	.008	$\epsilon^2 = .016$
Q2	Depth distribution	$\chi^2(9) = 37.381$	< .001	Cramér's $V = .138$
Q3	Depth distribution	$\chi^2(9) = 21.086$	.035	Cramér's $V = .106$
Q4	Depth distribution	$\chi^2(9) = 21.903$	.035	Cramér's $V = .105$

Note. Only comparisons that remained statistically supported after Holm correction are shown. Depth distributions were compared using permutation chi-square tests with Cramér's V as the effect-size measure; breadth amount and volume were compared using permutation Kruskal-Wallis tests with ϵ^2 as the effect-size measure. For quartile-difference tests, Holm correction was applied across the four cue groups separately within each outcome. For cue-group-difference tests, Holm correction was applied across the four conversation quartiles separately within each outcome.

Within cue groups, depth distributions differed across conversation quartiles for No cue, Recommendation only, and Both cues, but not for Textual anthropomorphic cue only. Both cues was the only cue context in which statistically supported quartile differences were also observed for breadth amount and volume.

When cue groups were compared within individual conversation quartiles, depth distributions differed across cue groups in all four quartiles. Breadth amount differed across cue groups only in Quartile 1, while volume showed no statistically supported cue-group difference within any quartile.

These subgroup comparisons identify where cue-context differences occurred but do not establish whether the overall pattern across conversation quartiles differed between cue groups. The conversation quartile $\times$ cue-group interaction is therefore examined directly in Section 4.4.4.

4.4.4 Conversation Quartile $\times$ Cue-Group Interaction

The subgroup comparisons in Section 4.4.3 identified differences across conversation quartiles and cue groups but did not directly test whether the conversation-position pattern differed between cue contexts. Conversation quartile $\times$ cue-group interactions were therefore tested for depth distribution, breadth amount, and volume. Table 18 presents the results.

Table 18

Conversation Quartile $\times$ Cue-Group Interaction Tests

Outcome	Test statistic	Raw p	Holm-adjusted p	ΔR^2
Depth distribution	$F(9, 2623) = 2.171$	$< .001$	.002	.007
Breadth amount	$F(9, 2623) = 0.657$	.745	1.000	.002
Volume	$F(9, 2623) = 0.595$	.802	1.000	.002

Note. Two-sided p -values are based on Freedman-Lane permutation tests. Holm correction was applied across the three interaction tests. ΔR^2 represents the additional variance explained by adding the conversation quartile $\times$ cue-group interaction to the corresponding model containing the main effects of conversation quartile and cue group. Depth was represented using categorical depth indicators, breadth amount in its original numerical form, and volume as $\log_{10}(\text{volume})$, as specified in Appendix C.2.

The conversation quartile $\times$ cue-group interaction was statistically supported for depth distribution after Holm correction (Table 18). The interaction explained an additional $\Delta R^2 = .007$ beyond the corresponding model containing the main effects of conversation quartile and cue group. This indicates that the pattern of depth across conversation quartiles differed across the four cue groups.

The omnibus interaction test does not identify which specific cue-group trajectories differed from one another. The descriptive patterns reported in Section 4.4.3 therefore provide context for this result but should not be interpreted as separately tested interaction contrasts.

No statistically supported conversation quartile $\times$ cue-group interaction was observed for breadth amount or volume after Holm correction (Table 18). The analysis therefore provided statistical support for different conversation-position patterns across cue groups for depth distribution, but not for breadth amount or volume.

4.4.5 Conversation-Level Cluster Analysis

The preceding RQ3 analyses treated individual <AI, user> turn pairs as analytical observations, although multiple turn pairs originated from the same conversation. The conversation-level cluster analysis therefore re-examined the central RQ3 associations using all 2,639 turn pairs nested within the 69 sampled conversations

while accounting for within-conversation dependence. Separate models tested the overall association between conversation quartile and each outcome and the conversation quartile $\times$ cue-group interaction. Table 19 presents the omnibus results.

Table 19

Conversation-Level Clustered Analysis of Conversation Position and Cue-Group Interactions

Outcome	Test statistic	Raw p	Holm-adjusted p	ΔR^2
Panel A. Conversation-Position Tests				
Ordered depth	$F(3, 68) = 3.816$	.014	.041	.014
Breadth amount	$F(3, 68) = 2.223$	.093	.093	.004
Volume (log1p-transformed)	$F(3, 68) = 3.487$	.020	.041	.008
Panel B. Conversation Quartile $\times$ Cue-Group Interaction Tests				
Ordered depth	$F(9, 68) = 2.341$	.023	.069	.006
Breadth amount	$F(9, 68) = 0.801$	.617	.702	.002
Volume (log1p-transformed)	$F(9, 68) = 1.135$	.351	.702	.002

Note. $N = 2,639$ turn pairs nested within 69 conversations. Inference used conversation-clustered covariance estimates with small-sample and degrees-of-freedom corrections. Depth was represented as an ordered score from 1 (Superficial) to 4 (Core), breadth amount in its original numerical form, and volume as log1p(volume). Holm correction was applied separately across the three conversation-position tests and the three interaction tests. For Panel A, ΔR^2 represents the additional variance explained by conversation quartile relative to an intercept-only model. For Panel B, ΔR^2 represents the additional variance explained by adding the conversation quartile $\times$ cue-group interaction to the corresponding main-effects model.

Under conversation-clustered inference, the association with conversation quartile remained statistically supported for ordered depth and log1p-transformed volume after Holm correction, but not for breadth amount (Table 19). The statistically supported breadth-amount difference observed in the primary conversation-position analysis was therefore not reproduced under the clustered specification.

None of the conversation quartile $\times$ cue-group interactions remained statistically supported after Holm correction (Table 19). Thus, under the conversation-clustered specification, the analysis did not provide statistical support for different conversation-position patterns across cue groups for ordered depth, breadth amount, or log1p-transformed volume.

For breadth amount and volume, the absence of statistically supported interactions was consistent with the primary interaction analysis in Section 4.4.4. For depth, however, the statistically supported interaction in the primary categorical-depth analysis was not reproduced in the clustered analysis using ordered depth. Because the two analyses represent depth differently, this difference cannot be attributed to conversation-level clustering alone. The depth interaction was therefore not reproduced under the clustered ordered-depth specification.

5 Interpretation of the Findings

This chapter interprets the findings for the three research questions in relation to the theoretical expectations and prior empirical evidence discussed in Chapter 2. Broader academic contributions, practical implications, limitations and generalisability, and future research are addressed in Chapter 6.

5.1 RQ1: Recommendation Presence, Observable Uptake, and Subsequent User Self-Disclosure

After adjustment for turn number and textual anthropomorphic-cue presence, none of the recommendation-group contrasts remained statistically supported for depth, breadth, or volume after Holm correction. This does not establish equivalence between the groups; rather, the adjusted analyses did not provide sufficient evidence to retain the differences observed in the primary comparisons.

This pattern is compatible with previous research showing that responses to conversational steering depend on how steering is embedded within an interaction. Advice has been evaluated more favourably when preceded by emotional support and problem inquiry, whereas unsolicited recommendations that conflict with an individual's existing preferences can evoke resistance (Feng, 2009; Fitzsimons & Lehmann, 2004). Follow-up questions can also elicit more informative and specific user responses (Xiao et al., 2020). Together, these studies suggest that responses to conversational steering can vary across interactional conditions (Feng, 2009; Fitzsimons & Lehmann, 2004; Xiao et al., 2020). These interactional conditions provide a possible explanation for why responses to recommendation-related steering may vary across exchanges. However, the adjusted findings show no consistent association between recommendation presence and greater subsequent self-disclosure.

Consistent with the distinction established in Section 2.3.3, the adjusted results also provide no evidence that observable uptake corresponds to greater subsequent self-disclosure. Recommendations with and without observable uptake did not differ statistically in depth, breadth, or volume after adjustment. Observable uptake and subsequent self-disclosure should therefore remain conceptually separate outcomes rather than being treated as interchangeable indicators of user responsiveness (Karnam et al., 2026).

The supplementary recommendation-type analysis adds nuance to this interpretation. The observed variation cautions against assuming that different forms of recommendation-related steering are empirically interchangeable, which is consistent with the distinctions developed in Chapter 2 (Feng, 2009; Xiao et al., 2020). Taken together, RQ1 provides no adjusted evidence that recommendation presence or observable uptake is consistently associated with greater depth, breadth, or volume of subsequent user self-disclosure.

5.2 RQ2: Textual Anthropomorphic Cues and Subsequent User Self-Disclosure

After adjustment for turn number and recommendation presence, the initially supported differences in depth and breadth were no longer statistically supported after Holm correction. These differences concerned responses following messages with versus without textual anthropomorphic cues. Volume had not differed in the primary analysis. This does not establish that textual anthropomorphic cues are unrelated to disclosure under all conditions. Rather, the adjusted analyses did not provide evidence of a consistent association with subsequent self-disclosure.

This finding qualifies, rather than directly contradicts, previous anthropomorphism research. Earlier studies have shown that human-like cues can increase perceived anthropomorphism and social presence, while social presence and trust have subsequently been associated with disclosure intentions or self-reported disclosure (J. Lee et al., 2024; Li et al., 2023). Kim and Wang (2025) similarly linked socially oriented chatbot design to greater self-reported disclosure through lower psychological reactance. These studies primarily concern users' perceptions, intentions, or post-interaction self-reports, whereas the present study examines observable textual cues in naturally occurring ChatGPT messages and coded disclosure in the immediately subsequent response. The adjusted findings therefore do not support extending earlier perception- and intention-based evidence to observable textual anthropomorphic cues. They also do not support the broader claim that such cues are consistently associated with greater subsequent depth, breadth, or volume.

The absence of an adjusted cue-presence association does not necessarily conflict with earlier anthropomorphism research, because textual cue presence is conceptually distinct from users' perceptions and psychological responses to those cues (Greulich et al., 2025; Li et al., 2023). These perceptual and psychological mechanisms were not measured in the present study and therefore remain possible explanations rather than mechanisms demonstrated by the analysis.

The supplementary cue-type analysis adds nuance to this conclusion. Variation across cue types and across depth, breadth, and volume cautions against treating all textual anthropomorphic cues as empirically interchangeable. Taken together, RQ2 provides no adjusted evidence that textual anthropomorphic-cue presence is consistently associated with greater depth, breadth, or volume of subsequent user self-disclosure.

5.3 RQ3: Conversation Position, Cue Context, and Subsequent User Self-Disclosure

RQ3 showed that subsequent self-disclosure varied across relative conversation position, with the supported pooled differences concentrated around the earliest quartile and conversation-clustered inference retaining associations for depth and

volume, but not breadth. The pattern was therefore neither a uniform early-to-late progression nor consistent across all three dimensions of self-disclosure.

This within-conversation pattern should not be interpreted as evidence of relationship development. Social Penetration Theory describes disclosure in relation to developing interpersonal relationships, while longitudinal human-chatbot studies examine change across repeated interactions over longer periods (Carpenter & Greene, 2015; Y. C. Lee et al., 2020; Skjuve et al., 2023). Relative position within a single conversation therefore represents a different type of change over time. The present findings also do not show a stage-like progression, because the clearest pooled differences were concentrated around the earliest quartile rather than forming a consistent sequence across later positions. Direct evidence on systematic earlier-versus-later disclosure within individual natural conversational-AI interactions remains limited (Papneja & Yadav, 2025).

The conversation-length sensitivity analysis limits how broadly the pooled Q1-centred pattern can be interpreted. The pattern was not reproduced consistently across conversation lengths, and the amount of available evidence differed substantially between strata. Moreover, because the analysis repeated quartile comparisons within length strata rather than testing a conversation-length $\times$ position interaction, it does not establish conversation length as a moderator.

Several processes could plausibly underlie the concentration of pooled differences around Q1. Longitudinal chatbot research suggests that disclosure can change as perceived rewards, costs, novelty, or interaction experience develop over time (Skjuve et al., 2023). Because these mechanisms have primarily been examined across repeated interactions, they provide hypotheses for interpreting within-conversation variation rather than direct explanations of the present Q1-centred pattern.

Cue context provides little additional evidence that the position pattern systematically depends on the conversational features preceding the user response. The primary analysis supported a depth interaction, but this was not reproduced under the conversation-clustered ordered-depth specification, while breadth and volume showed no supported interaction. Because depth was represented differently across these analyses, the discrepancy cannot be attributed to clustering alone. The evidence is therefore insufficient to conclude that recommendation context or textual anthropomorphic-cue context systematically changes the relationship between conversation position and subsequent self-disclosure.

Taken together, RQ3 shows that subsequent self-disclosure differs depending on where an exchange occurs within a conversation. However, these differences do not follow a simple pattern in which disclosure consistently increases or decreases as the conversation progresses. The evidence is strongest for depth and volume, but the associations are modest, are not reproduced uniformly across conversation lengths, and are not shown to depend systematically on the examined cue contexts.

6 Discussion

Building on the RQ-specific interpretations in Chapter 5, this chapter brings the findings together to consider their broader academic and practical meaning. It examines what the combined findings contribute to understanding self-disclosure in conversational AI, what they imply for the design and evaluation of such systems, and which methodological and contextual boundaries constrain these conclusions. These boundaries also provide the basis for identifying the main directions in which future research can extend the present study.

6.1 Academic Contributions to Self-Disclosure in Conversational AI

The combined findings identify three distinctions that are important for interpreting self-disclosure in conversational-AI research. First, self-disclosure should be treated as a multidimensional outcome. Second, observable conversational features and user responses should be distinguished from the psychological mechanisms proposed to explain them. Third, change within a single conversation should be distinguished from change across repeated interactions or longer-term human-chatbot relationship development (Papneja & Yadav, 2025; Skjuve et al., 2023). Together, these distinctions provide an empirically grounded basis for specifying and interpreting associations between conversational features and user self-disclosure more precisely.

First, the findings provide turn-level empirical evidence for analysing self-disclosure as a multidimensional outcome. Social Penetration Theory distinguishes depth from breadth, while Omarzu (2000) additionally treats the amount or duration of disclosure as a distinct dimension. These theoretical distinctions support examining different dimensions of self-disclosure separately rather than treating disclosure as a single outcome (Carpenter & Greene, 2015; Omarzu, 2000). These distinctions also matter empirically in the present study: the examined conversational features did not show consistent adjusted associations across depth, breadth, and volume, while conversation position was related more consistently to depth and volume than to breadth. The present findings therefore show that conclusions about self-disclosure can differ depending on the dimension examined, supporting the separate analysis of depth, breadth, and volume in conversational-AI research.

Second, the findings highlight the importance of separating observable conversational features, observable user responses, and the psychological mechanisms proposed to explain them. Prior research identifies social presence, trust, privacy-related perceptions, supportive sequencing, and psychological reactance as possible mechanisms connecting conversational design with disclosure-related responses (Feng, 2009; Fitzsimons & Lehmann, 2004; Li et al., 2023). However, a textual anthropomorphic cue is not the same construct as perceived anthropomorphism, just as observable recommendation uptake is not equivalent to psychological acceptance or real-world compliance (Karnam et al., 2026; Li et al.,

2023). The present study contributes by showing why these distinctions matter when interpreting naturally occurring turn-level conversational data: observable system features and user responses can be examined empirically without treating them as direct measures of the psychological mechanisms proposed to explain them.

Third, the findings reinforce the importance of distinguishing change within a single conversation from change across repeated interactions or longer-term human-chatbot relationship development. Longitudinal studies examine disclosure across repeated interactions over days or weeks, whereas the present study concerns relative position within a single conversation (Y. C. Lee et al., 2020; Skjuve et al., 2023). The RQ3 findings indicate that within-conversation position can be analytically relevant, but it represents a different type of change over time. Findings from repeated or longer-term interaction should therefore not automatically be assumed to explain earlier-versus-later variation within a single conversation (Papneja & Yadav, 2025; Skjuve et al., 2023). The present study therefore contributes by operationalising and empirically examining relative within-conversation position as a distinct temporal level of analysis, separate from longer-term change across repeated interactions or human–chatbot relationship development.

Taken together, these contributions provide a more bounded basis for interpreting self-disclosure in conversational AI. Disclosure dimensions, observable conversational features, observable user responses, psychological mechanisms, and different forms of change over time should not be collapsed into single constructs (Carpenter & Greene, 2015; Karnam et al., 2026; Li et al., 2023; Omarzu, 2000; Papneja & Yadav, 2025; Skjuve et al., 2023). Keeping these distinctions explicit supports more precise interpretation and comparison of findings across conversational-AI research.

6.2 Practical Implications

The findings are most directly relevant to organisations and product teams responsible for designing, evaluating, and governing conversational-AI services in disclosure-sensitive contexts. The central managerial implication is that self-disclosure should not automatically be treated as evidence of successful engagement. Personal information can enable more tailored interactions, but disclosure also introduces privacy and confidentiality concerns, particularly in sensitive contexts such as mental health (Meadi et al., 2025; Papneja & Yadav, 2025). Product teams therefore face a trade-off: conversational features may support interaction and personalisation while simultaneously eliciting information that is sensitive or unnecessary for the intended purpose. The relevant managerial question is consequently not whether a feature produces more interaction or more disclosure, but whether it contributes to the intended user purpose while requiring an appropriate level of personal information.

A first implication concerns how conversational-AI performance is evaluated. Recommendation uptake, conversational continuation, and self-disclosure represent different outcomes and should not be collapsed into a single measure of successful

interaction. Observable recommendation uptake indicates that a user continued in the direction proposed by the system, but does not establish psychological acceptance, user benefit, or real-world compliance (Karnam et al., 2026). Product teams should therefore define the intended outcome of a conversational feature before evaluating it and assess that outcome separately from disclosure-related consequences. Where self-disclosure is relevant to evaluation, depth, breadth, and volume should also be considered separately rather than interpreted as a single indicator of engagement.

This distinction is particularly important for recommendation-related steering. Previous research shows that advice can be evaluated differently depending on the surrounding interaction: emotional support and problem inquiry can improve advice evaluations, whereas unsolicited recommendations that conflict with existing preferences can produce psychological reactance (Feng, 2009; Fitzsimons & Lehmann, 2004). Recommendation features should therefore be evaluated in relation to how and when they are delivered and against the specific outcome they are intended to achieve. Observable responses to recommendations should not be maximised as an objective in themselves unless they are shown to contribute to that intended outcome.

A second implication concerns anthropomorphic conversational design. The adjusted RQ2 findings provide no basis for assuming that making conversational language more human-like will consistently increase the depth, breadth, or volume of subsequent user self-disclosure. Prior research indicates that anthropomorphic design can increase perceptions such as social presence and trust, but can also raise context-dependent concerns related to privacy, anxiety, heightened expectations, and responses to overly human-like systems (Greilich et al., 2025; Li et al., 2023). Anthropomorphic language should therefore be treated as a design variable whose value depends on the intended function and interaction context, rather than as a general mechanism for increasing engagement. When product teams introduce or modify human-like conversational cues, they should evaluate whether those cues achieve their intended purpose while also considering possible disclosure and privacy consequences.

A third implication concerns when disclosure-sensitive evaluation should occur. The RQ3 findings show that differences in subsequent self-disclosure were already visible in the earliest part of conversations. Privacy-sensitive evaluation should therefore not focus only on prolonged interactions or situations in which a relationship with the system has already developed. Privacy and confidentiality are recurring ethical concerns in conversational AI for mental-health contexts because interactions can involve highly sensitive personal information (Meady et al., 2025). Product teams should consequently include early conversational exchanges when evaluating whether prompts, recommendations, or other conversational features elicit personal information that is proportionate to the intended purpose. This does not imply that disclosure should be minimised indiscriminately: personal information may be relevant to useful personalisation or support. The managerial issue is instead whether

the type and extent of disclosure elicited are justified by what the interaction is intended to achieve (Papneja & Yadav, 2025).

Taken together, the findings suggest a practical evaluation logic for disclosure-sensitive conversational AI. Product teams should define the intended purpose of a conversational feature, the outcome against which it will be evaluated, and the personal information required to support that purpose. Observable uptake, continued conversation, human-like language, and self-disclosure should then be evaluated as distinct aspects of the interaction rather than as interchangeable indicators of success. Privacy-sensitive evaluation should also include early conversational exchanges, where disclosure-related variation may already be present. From a Management of Technology perspective, this makes conversational-AI design both a governance issue and an interface-design issue. Organisations need to balance functionality, personalisation, and user engagement against the responsible handling of personal information, and define these evaluation criteria before treating conversational behaviour as evidence of technological success.

6.3 Limitations and Generalisability

The findings of this study should be interpreted within several methodological and contextual boundaries. These limitations concern the observational design and use of secondary conversation data, the construction of the dataset and analytical sample, and the operationalisation and coding of the examined constructs. They also concern the next-turn unit of analysis and representation of conversation position, remaining statistical uncertainty, and the coverage and reproducibility of the literature review. Rather than treating these issues as separate methodological caveats, this section considers how they affect the interpretation, evidential strength, and generalisability of the findings. Each limitation is therefore discussed in relation to the conclusions it constrains and, where relevant, the analytical procedures used to reduce its impact.

Table 20

Overview of Study Limitations and Interpretation Boundaries

Area	Main limitation	Interpretation boundary
Observational design and secondary data	Residual confounding, unavailable psychological measures, model changes, and behavioural scope	Associations do not establish causal effects, psychological mechanisms, clinical effectiveness, safety, or user benefit/harm
Dataset and sampling	Non-probability source data, study-specific eligibility thresholds, disproportionate length sampling, and no power-derived sample size	Frequencies are not population estimates and transfer is limited to comparable conversations and users

Area	Main limitation	Interpretation boundary
Measurement and coding	Study-specific operational measures, same-response uptake measurement, restricted recommendation coding, aggregated feature types, and single-researcher coding	Measures should not be interpreted as complete psychological constructs or independently validated classifications
Unit of analysis and conversation position	Next-turn pairs capture immediate sequence; initial user message is excluded; quartiles approximate relative position	Findings do not establish delayed conversational effects, initial disclosure willingness, or relationship-development stages
Statistical uncertainty	Repeated observations, incomplete clustering adjustment, sparse subgroups, non-equivalent clustered depth specification, and multiple testing families	Some estimates remain imprecise and supplementary findings have weaker evidential strength than the principal analyses
Literature review	Restricted database/search coverage, incomplete historical search records, single-researcher screening, English-only scope, and review cut-off	The research gap reflects the reviewed evidence base rather than exhaustive identification of all relevant literature

6.3.1 Observational Design and Secondary Data

The use of naturally occurring conversation histories allows this study to examine associations between recommendations, textual anthropomorphic cues, conversation position, and subsequent self-disclosure in actual ChatGPT interactions rather than researcher-created scenarios (Karnam et al., 2026). However, because these conversational characteristics were neither experimentally manipulated nor randomly assigned, the observed relationships may also reflect unmeasured characteristics of the user or preceding interaction. Recommendations and textual anthropomorphic cues may, for example, partly respond to information users had already disclosed, while user goals, emotional state, preceding context, and offline circumstances were unavailable. The adjusted analyses accounted for selected observable characteristics, and conversation-level procedures addressed dependence between repeated exchanges, but these procedures cannot eliminate residual confounding. The RQ1 and RQ2 findings therefore concern associations rather than the effects that introducing or removing these features would produce. Likewise, the RQ3 findings do not establish that progression through the conversation itself caused the observed differences in self-disclosure.

A related limitation is that the conversation histories do not directly capture how users perceived individual messages or why they disclosed particular information.

Psychological processes such as perceived anthropomorphism, social presence, trust, and psychological reactance were therefore not measured directly. The findings can consequently establish relationships between observable conversational characteristics and subsequent disclosure behaviour, but not the psychological mechanisms that may explain those relationships.

The donated conversation histories also span a period in which ChatGPT's underlying models changed, while model version and calendar time were not incorporated into the analyses (Karnam et al., 2026). Variation between system configurations may therefore have contributed to heterogeneity in both the examined conversational characteristics and users' responses. The findings should consequently be interpreted as patterns across the included ChatGPT histories rather than estimates attributable to a single fixed model configuration.

Finally, the study is limited to observable conversational behaviour. The analyses do not establish whether ChatGPT is clinically effective or safe, whether its recommendations are clinically appropriate, or whether greater self-disclosure benefits or harms users. These outcomes were not measured and therefore fall outside the conclusions of this study.

6.3.2 Dataset and Sampling

The empirical findings are bounded by both the composition of the source data and the construction of the analytical sample. The source data do not constitute a probability sample of the broader population of ChatGPT users. Participants in the original data-donation study were recruited through Prolific and were required to have sustained prior experience with ChatGPT, including at least 100 conversations and 90 days of activity (Karnam et al., 2026). The source population therefore primarily represents relatively experienced ChatGPT users who also chose to donate their conversation histories. In addition, this thesis had access only to the pre-existing Mental Health category identified in the original study, meaning that relevant conversations assigned to another category could not enter the sampling frame.

Additional selection occurred through the English-language and study-specific eligibility criteria. Conversations were retained only when all turns satisfied the English-language requirement, which may have excluded otherwise relevant conversations near the classification boundary. The automated language-classification procedure also treated combined user-ChatGPT turns containing fewer than 20 characters as unknown. This threshold was study-specific and alternative thresholds were not evaluated. Eligibility screening also excluded single-turn conversations and conversations that could not be interpreted sufficiently from the available text. The subsequent manual screening applied a 90% threshold to mental-health relevance, English-language use, and interpretability. This threshold was a study-specific operational decision rather than an empirically validated cut-off, and different choices could therefore have produced a somewhat different analytical sample. These procedures supported a more consistent basis for coding and analysis, but also narrowed the range of conversations represented.

The sampling strategy introduced a further trade-off. Longer conversations were deliberately represented more strongly than their prevalence in the sampling frame because RQ3 required sufficient coverage across relative positions within conversations. The final sample of 69 conversations was therefore not intended to reproduce the conversation-length distribution of the source pool. Its size reflected the stratified sampling strategy, the availability of eligible conversations, and the practical requirements of manual coding rather than a formal statistical-power calculation. Descriptive frequencies in the analytical sample should consequently not be interpreted as prevalence estimates for ChatGPT mental-health conversations or ChatGPT use more broadly.

The empirical findings therefore apply most directly to donated, English-language, multi-turn mental-health conversations from sustained ChatGPT users represented by this study-specific sampling design. Transfer to other conversational-AI systems, languages, user populations, conversational domains, or differently structured interaction samples requires independent examination. This boundary is narrower for the empirical patterns than for the conceptual distinctions developed in the thesis. Although these distinctions may provide a useful basis for research in other conversational-AI contexts, whether the specific empirical relationships identified here recur across systems, populations, languages, or conversational domains requires independent replication.

6.3.3 Measurement and Coding

Self-disclosure was operationalised through observable content in user responses across depth, breadth, and volume using coding categories developed specifically for this study rather than validated measurement instruments. These measures should therefore be interpreted as operational approximations of selected dimensions of self-disclosure rather than direct or exhaustive measurements of the underlying construct. Conversation text cannot establish whether disclosed information was factually accurate or how vulnerable, meaningful, or deliberate the disclosure was experienced by the user. Prior research distinguishes dimensions such as honesty and conscious intent from depth and breadth, further indicating that observable personal statements do not capture every property of self-disclosure (Papneja & Yadav, 2025).

Each disclosure dimension also has a specific operational boundary. Depth represents the highest coded sensitivity identified anywhere in the complete user response rather than its average depth, while the Superficial category includes both responses without identified self-disclosure and responses containing only minimally personal information. Breadth represents the number of distinct personal content areas captured by the study's coding taxonomy rather than every possible dimension of personal disclosure (Carpenter & Greene, 2015; Omarzu, 2000). Volume represents the number of words coded as self-disclosing content. It therefore captures disclosure amount in words, but not the informational or psychological significance of what is disclosed (Harmsen et al., 2023; Omarzu, 2000).

The conversational-feature variables introduce additional measurement boundaries. Observable recommendation uptake and self-disclosure were coded from the same subsequent user response. Differences between uptake groups therefore describe how observable uptake and self-disclosure co-occur within that response rather than an effect of uptake on disclosure. Uptake was also limited to behaviour observable in the immediately following response, meaning that recommendations requiring later or offline action were outside the RQ1 operationalisation. Textual anthropomorphic cues were similarly measured as observable linguistic features rather than users' perceptions of those cues (Li et al., 2023). Finally, the principal RQ1 and RQ2 analyses combined multiple recommendation and cue forms into broader categories. These comparisons may therefore conceal differences between specific recommendation or anthropomorphic-cue types.

Finally, all final coding was performed by one researcher. The predefined codebook, fixed coding sequence, ambiguity rules, and post-coding consistency checks supported systematic application of the coding scheme, but they do not demonstrate that another researcher would reproduce the same classifications. No intercoder-reliability statistic was therefore available as an independent assessment of reproducibility. This is particularly relevant because conversational self-disclosure can require interpretive decisions when content is ambiguous or varies in granularity (Skjuve et al., 2023). The findings should consequently be understood as relationships among systematically operationalised observable variables rather than independently validated measurements of the underlying constructs.

6.3.4 Unit of Analysis and Conversation Position

The next-turn analytical design links characteristics of a ChatGPT message to self-disclosure in the immediately subsequent user response, but captures only a local part of the broader conversational process. User responses may reflect information or interactional developments accumulated across earlier turns, while recommendations or textual anthropomorphic cues may influence disclosure only after several subsequent exchanges. The <AI, user> turn pair therefore captures immediate sequential associations and may not detect delayed or cumulative relationships.

The analytical pair also excludes the user's initial message because no ChatGPT message precedes it. The earliest disclosure analysed is therefore the first user response following a ChatGPT message rather than the disclosure with which the user initiated the conversation. Consequently, findings for the earliest conversation position should not be interpreted as evidence about users' initial willingness to disclose when opening a conversation.

Conversation quartiles represent a user response's relative sequential position within a conversation rather than elapsed time, relationship duration, or a stage of human-chatbot relationship development. Conversations of different lengths also contribute differently to these relative-position categories, and the conversation-length sensitivity analysis showed that the pooled position pattern was not reproduced

consistently across all conversation lengths. The RQ3 findings should therefore be interpreted as associations with relative within-conversation position rather than as a uniform disclosure trajectory or a measure of relationship development.

6.3.5 Statistical Uncertainty

The analytical strategy addressed several sources of statistical uncertainty, but some dependence between observations remains relevant to the interpretation of the findings. Multiple <AI, user> turn pairs originated from the same conversation and therefore cannot be assumed to represent fully independent observations. The adjusted RQ1 and RQ2 analyses addressed this structure by resampling complete conversations, while the central RQ3 conversation-position and interaction analyses were additionally re-estimated using conversation-clustered inference. These procedures provided checks on the sensitivity of the principal findings to within-conversation dependence. However, the primary unadjusted group comparisons, subgroup analyses, pairwise comparisons, and type-level analyses remained turn-level analyses without an equivalent conversation-level adjustment. Possible dependence between multiple conversations contributed by the same donor was also not modelled. The conversation-level procedures therefore reduced, but did not eliminate, dependence-related uncertainty across the complete analytical framework.

The conversation-clustered sensitivity analysis for the RQ3 quartile $\times$ cue-group interaction also represented depth differently from the primary interaction analysis. The primary analysis used categorical depth, whereas the clustered specification used ordered depth. Consequently, the failure to reproduce the statistically supported primary depth interaction under the clustered specification cannot be attributed to conversation-level dependence alone, because the two analyses differ in both their dependence structure and their representation of depth. This limits the extent to which that sensitivity comparison can isolate the effect of conversation-level clustering.

Precision also varied across the analyses because some recommendation types, textual anthropomorphic-cue types, conversation-length strata, and cue-group $\times$ quartile combinations contained relatively few observations. Estimates based on these smaller groups are consequently less precise and more sensitive to individual observations than estimates based on larger groups. A lack of statistical support in these analyses should therefore be interpreted together with the available sample sizes, effect sizes, confidence intervals where applicable, and descriptive patterns rather than as evidence that an association is absent. This is particularly important for the supplementary analyses, which were intended to clarify or test the boundaries of the principal findings rather than to establish equally well-supported standalone relationships.

A further consideration is the number of statistical comparisons conducted across the thesis. Holm correction was applied to control for multiple comparisons within predefined families of related tests (Holm, 1979). The correction was not applied

globally across every exploratory, primary, supplementary, sensitivity, and interaction analysis in the thesis. Statistical support should therefore be interpreted in relation to the analytical family in which a comparison was tested rather than as if one experiment-wide correction had been applied across the entire study.

Taken together, these considerations mean that the strength of the statistical evidence is best evaluated across the complete pattern of results rather than through individual p-values in isolation. The principal interpretations therefore draw on the consistency of descriptive patterns, effect sizes, confidence intervals where available, adjusted analyses, conversation-level sensitivity procedures, and sensitivity analyses. Supplementary type-level, conversation-length, and cue-context findings are most appropriately used to qualify and contextualise the main results. This is particularly important where subgroup sizes are limited. These supplementary findings should therefore not be treated as independently established relationships with equal evidential strength.

6.3.6 Literature Review Coverage and Reproducibility

A final limitation concerns the coverage and reproducibility of the literature review underlying the theoretical background and research gap. Web of Science was the only academic database searched directly, and systematic backward and forward citation searching was not conducted. Relevant publications indexed elsewhere or identifiable through citation-based searching may therefore have been missed. In addition, the complete historical search log and formal counts of publications screened and excluded were not retained, which limits exact reconstruction of the literature-identification process. Source selection and evaluation were conducted by a single researcher, and the review was restricted to English-language publications for which the required source material could be accessed. Because no independent agreement procedure was used for screening or source evaluation, another review process could identify additional relevant publications or evaluate the relevance and strength of individual sources differently. Supplementary searching continued until May 2026, meaning that later publications fall outside the reviewed evidence base.

These limitations principally constrain claims about the completeness of the literature review and the research gap derived from it rather than the internal results of the empirical analysis. The research gap should therefore be understood as the conclusion supported by the literature identified and critically synthesised within the documented review process, rather than as evidence that no relevant research exists outside that source base.

6.4 Future Research

Future research should test whether recommendation-related steering and textual anthropomorphic cues causally affect subsequent self-disclosure and examine the mechanisms through which such relationships may occur. Controlled experiments could manipulate specific recommendation characteristics or textual anthropomorphic cues while holding other interactional features constant. Such

designs could also measure processes such as perceived anthropomorphism, social presence, trust, or psychological reactance, which previous research has linked to anthropomorphic design and responses to advice (Feng, 2009; Kim & Wang, 2025; Li et al., 2023). This would distinguish observable feature-disclosure relationships from the psychological processes proposed to explain them.

Second, future research should examine self-disclosure beyond the immediately subsequent user response and across multiple turns within a conversation. Independent samples could first be used to assess whether the within-conversation position patterns identified in this study are reproduced. Multi-turn designs could then examine whether relationships involving recommendations, textual anthropomorphic cues, and self-disclosure emerge immediately, accumulate across several exchanges, or appear only after a delay. Such designs could also capture recommendation uptake occurring after the immediately following response and incorporate users' initial disclosure when opening the conversation. This would allow subsequent change to be distinguished from pre-existing disclosure patterns while retaining a clear distinction between within-conversation change and longer-term development across repeated interactions (Y. C. Lee et al., 2020; Skjuve et al., 2023).

Third, future studies should strengthen the measurement and validation of the constructs examined here. Independent coding by multiple researchers would allow the reproducibility of classifications for recommendations, observable uptake, textual anthropomorphic cues, and self-disclosure to be assessed. Alternative operationalisations of depth, breadth, and volume could also be compared. The breadth taxonomy could be refined using a larger inductively coded sample, while the word-based measure of self-disclosure volume could be compared with other measures of disclosure amount, such as coded disclosure statements (Harmsen et al., 2023; Omarzu, 2000). More granular analyses of recommendation and anthropomorphic-cue types could additionally examine whether the broader feature categories used in the principal analyses conceal type-specific relationships. These extensions would help determine whether the substantive conclusions persist across alternative measurement and coding choices.

Fourth, future research should examine the consequences of additional self-disclosure rather than assuming that greater disclosure is inherently beneficial or harmful. Self-disclosure can support more personalised interactions while also increasing privacy exposure in sensitive conversational contexts (Papneja & Yadav, 2025). Future studies could therefore connect observed disclosure behaviour to outcomes such as perceived usefulness, privacy concerns, satisfaction, or user well-being. This would provide a stronger empirical basis for determining when additional disclosure contributes to the intended purpose of an AI service and when it occurs without a corresponding user benefit.

Finally, the empirical patterns identified in this thesis should be replicated across broader technological and conversational contexts before being treated as transferable. Comparative research across LLM-based systems, model versions, languages, user populations, and conversational domains could establish which

patterns depend on the specific setting examined here. This is particularly relevant because the source data represent sustained ChatGPT users and span changes in ChatGPT's underlying models (Karnam et al., 2026). Larger and more diverse samples, with sample-size planning aligned to the intended analyses, could improve precision for less frequent recommendation types, anthropomorphic-cue types, and contextual combinations. Where the required identifiers are available, multilevel analyses could more explicitly account for dependence between turn pairs, conversations, and users. More focused confirmatory analyses could also reduce uncertainty associated with extensive multiple testing. Across these extensions, future studies should preserve clear distinctions between disclosure dimensions, observable conversational features, user responses, and psychological mechanisms. Doing so would support more meaningful comparisons across systems and contexts.

7 Conclusion

This chapter concludes the thesis by answering the three research questions and integrating their findings into an overall answer to the main research question.

7.1 Answers to the Research Questions

The three research questions are answered below based on the analyses presented in Chapter 4. Where subsequent analyses qualified the primary findings, these qualifications are incorporated into the final answers.

7.1.1 RQ1: Recommendation Presence and Observable Uptake

How do the depth, breadth, and volume of user self-disclosure differ between responses following no recommendation, a recommendation without observable uptake, and a recommendation with observable uptake?

Recommendation presence and observable uptake did not retain a statistically supported association with subsequent depth, breadth, or volume in the adjusted analyses after multiple-testing correction. Although the primary analyses identified differences between the recommendation groups, these differences were not retained after adjustment. Neither recommendations without observable uptake nor recommendations with observable uptake differed statistically from no recommendation. The direct comparison between the two recommendation groups was also not statistically supported for depth, breadth, or volume. The supplementary recommendation-type analysis further showed no consistent descriptive pattern in which observable uptake corresponded to greater depth, breadth, or volume across recommendation types. Taken together, RQ1 does not support a consistent adjusted association between recommendation presence or observable uptake and subsequent user self-disclosure.

7.1.2 RQ2: Textual Anthropomorphic Cues

How do the depth, breadth, and volume of user self-disclosure differ between responses following ChatGPT messages with and without textual anthropomorphic cues?

Textual anthropomorphic-cue presence did not retain a statistically supported association with subsequent depth, breadth, or volume in the adjusted analyses after multiple-testing correction. In the primary comparison, responses following messages containing textual anthropomorphic cues showed slightly greater depth and breadth, while no statistically supported difference was found for volume. These primary differences therefore did not persist after adjustment. The supplementary cue-type analysis showed descriptive variation across individual cue types, but these patterns did not establish cue-specific associations with subsequent self-disclosure. Taken

together, RQ2 provides no adjusted evidence of a consistent association between textual anthropomorphic-cue presence and subsequent depth, breadth, or volume of user self-disclosure.

7.1.3 RQ3: Conversation Position and Cue Context

How does conversation position relate to the depth, breadth, and volume of user self-disclosure across different recommendation and textual-anthropomorphic-cue contexts?

Conversation position was associated with depth and volume, but not breadth, under conversation-clustered inference. In the pooled quartile analyses, every statistically supported pairwise comparison involved the first conversation quartile. However, this Q1-centred pattern was not reproduced consistently across conversation lengths, with statistical support concentrated in the longer-conversation strata.

Evidence that these position-related patterns systematically differed across cue contexts was insufficient. The primary interaction analysis supported different conversation-position patterns across cue groups for depth distribution, but not for breadth or volume; however, the depth interaction was not reproduced in the clustered analysis using ordered depth, and none of the clustered interactions remained statistically supported after multiple-testing correction. Taken together, RQ3 indicates that relative conversation position is associated with depth and volume, while the evidence is insufficient to conclude that these position-related patterns systematically differ across the four cue contexts.

7.2 Answer to the Main Research Question

How do ChatGPT recommendations, textual anthropomorphic cues, and conversation position relate to the depth, breadth, and volume of subsequent user self-disclosure in naturally occurring mental-health conversations?

This study examined naturally occurring, multi-turn mental-health conversations with ChatGPT to understand how recommendations, human-like language, and the position of an exchange within a conversation relate to the personal information users disclose. The clearest finding was that self-disclosure differed depending on where an exchange occurred within the conversation. Conversation position was related most clearly to how sensitive the disclosed information was and how much personal information users shared, while the number of different areas of users' personal lives that they discussed showed a less clear pattern. These position-related patterns were modest and were not reproduced uniformly across conversations of different lengths. Whether ChatGPT gave a recommendation, and whether the user responded in a way that followed or continued that recommendation, did not show a consistent relationship with the depth, breadth, or volume of the personal

information the user disclosed. The same was true for whether ChatGPT used more human-like language in the preceding message.

The findings also show why self-disclosure should not be treated as a single outcome. The sensitivity of what users share, the range of personal topics they discuss, and the amount of personal information they provide can vary differently. These aspects should therefore be examined separately when studying conversational AI. The study also shows why features that can be observed directly in a conversation should be kept separate from users' underlying perceptions or psychological responses.

For organisations and product teams designing conversational AI for sensitive contexts, the main takeaway is that more self-disclosure should not automatically be treated as a sign of a better interaction. Conversational features should instead be evaluated according to the purpose they are intended to serve, while the personal information users disclose is considered separately. Privacy-sensitive evaluation should also include the earlier parts of conversations rather than assuming that disclosure-related concerns arise only after prolonged interaction. The broader management challenge is therefore to design and evaluate conversational AI in a way that achieves its intended purpose while ensuring that the personal information users disclose is appropriate to that purpose.

8 Bibliography

- A. Colin Cameron, & Douglas L. Miller. (2015). A Practitioner's Guide to Cluster-Robust Inference. *Journal of Human Resources*, 50(2), 317–372.
<https://doi.org/10.3368/jhr.50.2.317>
- Carpenter, A., & Greene, K. (2015). Social Penetration Theory. In *The International Encyclopedia of Interpersonal Communication* (pp. 1–4). Wiley.
<https://doi.org/10.1002/9781118540190.wbeic160>
- Feng, B. (2009). Testing an integrated model of advice giving in supportive interactions. *Human Communication Research*, 35(1), 115–129.
<https://doi.org/10.1111/j.1468-2958.2008.01340.x>
- Fitzsimons, G. J., & Lehmann, D. R. (2004). Reactance to Recommendations: When Unsolicited Advice Yields Contrary Responses. *Marketing Science*, 23(1), 82–94.
<https://doi.org/10.1287/mksc.1030.0033>
- Frazier, P. A., Tix, A. P., & Barron, K. E. (2004). Testing moderator and mediator effects in counseling psychology research. *Journal of Counseling Psychology*, 51(1), 115–134. <https://doi.org/10.1037/0022-0167.51.1.115>
- Greilich, A., Bremser, K., & Wüst, K. (2025). Consumer Response to Anthropomorphism of Text-Based AI Chatbots: A Systematic Literature Review and Future Research Directions. *International Journal of Consumer Studies*, 49(5), e70108. <https://doi.org/10.1111/ijcs.70108>
- Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. <http://arxiv.org/abs/2203.05794>
- Harmesen, W. N., Van Waterschoot, J., Hendrickx, I., & Theune, M. (2023). Eliciting User Self-disclosure using Reciprocity in Human-Voicebot Conversations. *Proceedings of the 5th International Conference on Conversational User Interfaces, CUI 2023*, 1–6. <https://doi.org/10.1145/3571884.3604301>
- Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2), 65–70.
- Hsieh, H. F., & Shannon, S. E. (2005). Three approaches to qualitative content analysis. *Qualitative Health Research*, 15(9), 1277–1288.
<https://doi.org/10.1177/1049732305276687>
- Huang, F. L. (2018). Using Cluster Bootstrapping to Analyze Nested Data With a Few Clusters. *Educational and Psychological Measurement*, 78(2), 297–318.
<https://doi.org/10.1177/0013164416678980>
- Hubert, M., & Vandervieren, E. (2008). An adjusted boxplot for skewed distributions. *Computational Statistics and Data Analysis*, 52(12), 5186–5201.
<https://doi.org/10.1016/j.csda.2007.11.008>
- Karnam, S. K., Dash, A., Gummadi, Krishna., Mukherjee, A., Weber, I., & Zannettou, S. (2026). *Bowling with ChatGPT: On the Evolving User Interactions with Conversational AI Systems*. 9711–9721.
<https://doi.org/10.1145/3774904.3792978>
- Kelly, S., White, K. M., Kaye, S. A., & Oviedo-Trespalacios, O. (2025). Tapping into Key Drivers: Self-Disclosure in Sensitive Health Conversations with ChatGPT. *International Journal of Human-Computer Interaction*, 41(24), 15668–15678.
<https://doi.org/10.1080/10447318.2025.2499656>

- Kim, H., & Wang, Y. (2025). Unveiling the human touch: how AI chatbots' emotional support and human-like profiles reduce psychological reactance to promote user self-disclosure in mental health services. *International Journal of Advertising*, 54(6), 1499–1523.
<https://doi.org/10.1080/02650487.2025.2558479>
- Kruskal, W. H., & Wallis, W. A. (1952). Use of Ranks in One-Criterion Variance Analysis. *Journal of the American Statistical Association*, 47(260), 583–621.
<https://doi.org/10.1080/01621459.1952.10483441>
- Landis, R. S. (2014). *Wiley StatsRef: Statistics Reference Online* (N. Balakrishnan, T. Colton, B. Everitt, W. Piegorisch, F. Ruggeri, & J. L. Teugels, Eds.).
<https://doi.org/10.1002/9781118445112.stat06588>
- Lee, J., Lee, D., & Lee, J. gil. (2024). Influence of Rapport and Social Presence with an AI Psychotherapy Chatbot on Users' Self-Disclosure. *International Journal of Human-Computer Interaction*, 40(7), 1620–1631.
<https://doi.org/10.1080/10447318.2022.2146227>
- Lee, Y. C., Yamashita, N., Huang, Y., & Fu, W. (2020, April 21). 'I Hear You, i Feel You': Encouraging Deep Self-disclosure through a Chatbot. *Conference on Human Factors in Computing Systems - Proceedings*.
<https://doi.org/10.1145/3313831.3376175>
- Li, Q., Luximon, Y., & Zhang, J. (2023). The Influence of Anthropomorphic Cues on Patients' Perceived Anthropomorphism, Social Presence, Trust Building, and Acceptance of Health Care Conversational Agents: Within-Subject Web-Based Experiment. *Journal of Medical Internet Research*, 25, e44479.
<https://doi.org/10.2196/44479>
- Mann, H. B., & Whitney, D. R. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1), 50–60. <https://doi.org/10.1214/aoms/1177730491>
- Mead, M. R., Sillekens, T., Metselaar, S., van Balkom, A., Bernstein, J., & Batelaan, N. (2025). Exploring the Ethical Challenges of Conversational AI in Mental Health Care: Scoping Review. *JMIR Mental Health*, 12, e60432.
<https://doi.org/10.2196/60432>
- Omarzu, J. (2000). A Disclosure Decision Model: Determining How and When Individuals Will Self-Disclose. *Personality and Social Psychology Review*, 4(2), 174–185. https://doi.org/10.1207/S15327957PSPR0402_05
- Papneja, H., & Yadav, N. (2025). Self-disclosure to conversational AI: a literature review, emergent framework, and directions for future research. *Personal and Ubiquitous Computing*, 29(2), 119–151. <https://doi.org/10.1007/s00779-024-01823-7>
- Phipson, B., & Smyth, G. K. (2010). Permutation P-values should never be zero: Calculating exact P-values when permutations are randomly drawn. *Statistical Applications in Genetics and Molecular Biology*, 9(1), Article 39.
<https://doi.org/10.2202/1544-6115.1585>
- Pustejovsky, J. E., & Tipton, E. (2018). Small-sample methods for cluster-robust variance estimation and hypothesis testing in fixed effects models. *Journal of Business & Economic Statistics*, 36, 672–683.
<https://doi.org/10.1080/07350015.2016.1247004>

- Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation), Official Journal of the European Union L 119, 1–88 1 (2016). <https://gdpr-info.eu/art-9-gdpr/>
- Siddals, S., Torous, J., & Coxon, A. (2024). “It happened to be the perfect thing”: experiences of generative AI chatbots for mental health. *Npj Mental Health Research*, 3(1), 48. <https://doi.org/10.1038/s44184-024-00097-4>
- Skjuve, M., Følstad, A., & Brandtzæg, P. B. (2023). A Longitudinal Study of Self-Disclosure in Human-Chatbot Relationships. *Interacting with Computers*, 35(1), 24–39. <https://doi.org/10.1093/iwc/iwad022>
- Tomczak, M., & Tomczak, E. (2014). The need to report effect size estimates revisited. An overview of some recommended measures of effect size. *Trends in Sport Sciences*, 21(1), 19–25.
- Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M., & Nichols, T. E. (2014). Permutation inference for the general linear model. *NeuroImage*, 92, 381–397. <https://doi.org/10.1016/j.neuroimage.2014.01.060>
- Xiao, Z., Zhou, M. X., Liao, Q. V., Mark, G., Chi, C., Chen, W., & Yang, H. (2020). Tell Me About Yourself: Using an AI-Powered Chatbot to Conduct Conversational Surveys with Open-ended Questions. *ACM Transactions on Computer-Human Interaction*, 27(3). <https://doi.org/10.1145/3381804>

Appendix A:

Literature-Search and Review Process

This appendix documents the literature-search process used to develop Chapter 2. The search aimed to identify theoretical and empirical literature on self-disclosure in mental-health conversations with ChatGPT, focusing on recommendations, textual anthropomorphic cues, and conversation position in relation to the breadth, depth, and volume of disclosure. The review followed a targeted and iterative approach in which search terms were refined and supplementary searches were conducted as relevant concepts and evidence gaps emerged. The procedure is reported to make the identification, selection, and synthesis of sources transparent and traceable; it is not presented as an exhaustive systematic review.

A.1 Review Approach and Scope

The literature review followed a targeted and iterative approach, while the resulting literature was critically evaluated and conceptually synthesised. The review was intended to establish the theoretical foundations of the study, compare relevant empirical evidence, and identify the unresolved relationships addressed by the research. Rather than aiming for exhaustive coverage of a single narrowly defined literature, sources were selected according to their relevance to the concepts and relationships examined in Chapter 2.

The review prioritised evidence closest to the empirical context of this thesis: text-based interactions with ChatGPT and other generative conversational-AI systems involving mental health, health, or similarly sensitive personal disclosure. Where direct evidence was limited, studies of other text-based conversational agents were used as adjacent evidence. Research involving voice-based agents, visual or profile-based anthropomorphic cues, and other non-textual interaction features was included when it contributed to construct definition or theoretical explanation. However, this research was not treated as direct evidence for textual characteristics of ChatGPT messages. Foundational human-communication research was similarly used to define relevant concepts and inform theoretical explanations where necessary.

The review was restricted to English-language academic literature but applied no fixed publication-year boundary. This allowed recent research on generative conversational AI to be considered alongside established theoretical work required to conceptualise self-disclosure, recommendations, anthropomorphic communication, and the change over time aspects of interactions.

A.2 Search Strategy

Web of Science was the only academic database searched directly. The main searches were conducted between February and April 2026. Searches used the Topic, Title, Abstract, and All Fields options, depending on the purpose of the search. The TU Delft Library was used to obtain full-text access to publications that had already been identified but were not directly accessible.

The thesis topic was divided into searchable concept groups corresponding to the principal literature streams developed in Chapter 2. Table A1 presents the core search concepts and representative terms used during the search process. Because the search vocabulary developed iteratively and the complete historical search log was no longer available, these terms represent the principal search vocabulary rather than an exhaustive record of every term or combination used.

Table A1

Main Search Concepts and Representative Search Terms

Search concept	Representative terms
Conversational AI	<i>ChatGPT; chatbot; AI; large language model; LLM; human-chatbot; human-AI interaction; generative AI</i>
Interaction context	<i>mental health; health; privacy; sensitive disclosure; personal information; mental-health support; mental-health conversation</i>
Self-disclosure	<i>disclosure; self-disclosure; depth; breadth; volume; hypothetical willingness; self-reported disclosure; personal disclosure; personal information; disclosure behaviour</i>
Recommendations	<i>recommendation; advice; steering; suggestion; guidance; conversational steering; suggested action; course of action; recommendation message</i>
Anthropomorphism	<i>anthropomorphism; anthropomorphic cue; textual anthropomorphic cues; perceived anthropomorphism; human-like cues; humanlike cues; social cues; anthropomorphic features; human-like communication; humanness</i>
Conversation position	<i>conversation position; longitudinal; interaction sequence; conversational sequence; within-conversation progression; turn position</i>

Search terms were used individually and combined manually using the Boolean operators AND and OR. Combinations depended on the purpose of each search. For example, conversational-AI terms were combined with terms relating to self-disclosure, recommendations, anthropomorphism, or changes in interaction over time. When review literature was specifically sought, *systematic review* was combined with relevant topic terms. Additional terminology encountered while

reading relevant publications informed subsequent searches, as described in Section A.3.

Web of Science refinement and ordering functions were used selectively rather than through a single fixed search protocol. Depending on the purpose of a search, results were examined using the Review Articles, Highly Cited Papers, and Hot Papers options. Result lists were also reordered by relevance or publication date. English was not consistently imposed as a database filter; instead, language eligibility was assessed when potentially relevant publications were evaluated.

Because the complete historical Web of Science search log was no longer available when the search process was documented, an exhaustive reconstruction of every historical query was not possible. Table A1 and the description above therefore represent the core search vocabulary and procedures. The resulting reproducibility limitation is discussed in Section 6.3.6.

A.3 Search Development and Supplementary Searching

As relevant publications were reviewed, newly encountered terminology and theoretical concepts were used to conduct targeted follow-up searches in Web of Science. This iterative process allowed the search strategy to develop in response to the literature identified during the review rather than remaining restricted to the initial search vocabulary reported in Section A.2.

ChatGPT was occasionally used as a supplementary literature-discovery aid to identify potentially relevant search concepts and publication leads. When ChatGPT was used to identify literature, it was asked to provide the DOI of the suggested publication. The DOI was then independently checked and used to access the original publication. Every publication identified with AI assistance and subsequently used in the thesis was read and evaluated by the researcher. Text generated by ChatGPT was not used as academic evidence. Further details on the use and verification of generative AI are provided in Appendix E.

The study by Karnam et al. (2026) was identified through a recommendation from the thesis supervisor. Because the dataset analysed in the present study is a subset of the dataset used in Karnam et al. (2026), the paper became an important source for understanding the dataset and its broader research context. The main database-search phase was completed by the end of April 2026, while supplementary literature identification and targeted follow-up searching continued until May 2026.

A.4 Source Evaluation and Selection

Potentially relevant publications were evaluated according to their relevance to the research problem and their academic suitability for the role they would serve in Chapter 2. Relevance concerned whether a publication contributed substantively to defining concepts, informing theoretical explanations, evaluating empirical

relationships, or providing counter-evidence. It also concerned whether the publication helped identify methodological limitations or assess unresolved relationships relevant to the conceptual framework. Academic credibility was assessed according to whether a source was sufficiently reliable for the specific purpose for which it was used in Chapter 2. Peer-reviewed journal articles were preferred for empirical claims, while peer-reviewed conference publications were retained where they provided relevant evidence. Review articles supported broader synthesis, while established academic reference chapters were used for foundational theoretical concepts. Non-peer-reviewed academic publications were assigned lower evidential weight and were not used for central claims when stronger peer-reviewed evidence was available. Table A2 summarises the criteria used to evaluate source relevance, academic suitability, and evidential weight during the literature-selection process.

Table A2

Criteria for Source Evaluation, Selection, and Evidential Weighting

Criterion	Retained when...	Excluded or given lower evidential weight when...
Relevance to Chapter 2	The source contributed directly to defining a focal concept, explaining a proposed relationship, providing empirical evidence or counter-evidence, identifying a methodological limitation, or assessing an unresolved relationship addressed in Chapter 2.	The source did not materially contribute to any concept, relationship, empirical argument, or limitation required for Chapter 2.
Contextual fit	The study examined ChatGPT, conversational AI, self-disclosure, recommendations, anthropomorphic communication, or conversation position in a context sufficiently close to the thesis, or provided clearly relevant adjacent evidence.	Differences in technology, modality, setting, population, or outcome substantially limited the relevance of the findings to the research problem. Such studies could still be retained for theory but received lower weight as empirical evidence.
Academic credibility	The source was peer reviewed or was an established academic reference source appropriate for the claim it supported. Peer-reviewed empirical studies were preferred for empirical claims, while reviews and academic reference chapters were used primarily for synthesis or theoretical foundations.	The source was non-peer-reviewed, its academic status was unclear, or stronger peer-reviewed evidence was available for the same claim.

Criterion	Retained when...	Excluded or given lower evidential weight when...
Evidential assessability	The complete publication was available and provided enough information to evaluate its research design, sample or context, measures, findings, and relevant limitations.	The full text was unavailable or the publication did not provide sufficient information to determine how the evidence had been produced or what it could support.

Initial screening considered the title, abstract, and conclusion of potentially relevant publications. Publications that remained relevant were subsequently obtained and read in full before being used substantively in the thesis. A publication was not used substantively when its full text could not be obtained. Formal screening counts were not recorded. The resulting limitation is discussed in Section 6.3.6.

During full-text evaluation, each study was assessed to determine how strongly it could support claims in Chapter 2. This assessment considered the research design, sample and setting, measurement approach, and whether the outcome reflected observed behaviour, self-reported disclosure, or hypothetical willingness. It also considered whether the focal characteristic could be distinguished from other manipulated or contextual factors. Studies that were further removed from the thesis context were not automatically excluded, but their findings were interpreted more cautiously. Relevant mixed, null, and contradictory findings were retained alongside supportive findings.

A.5 Literature Documentation and Synthesis

A literature matrix was created during the main literature-search period between February and April 2026 to document the publications used in developing Chapter 2. The matrix was initially compiled manually by the researcher and updated as the literature review developed. It recorded bibliographic information and the relevant Chapter 2 theme or section, together with each study's context and sample, research design and method, constructs and measures, main findings, methodological limitations and boundary conditions, and relevance to the thesis. Relevant page numbers were also recorded to support traceability between the matrix entries and the underlying source material.

During later revision of the matrix, ChatGPT was used to flag incomplete or inconsistent entries and to improve the organisation and readability of information already recorded. Any suggested factual changes were checked by the researcher against the original publications before being retained. The researcher remained responsible for interpreting the sources, determining the thematic organisation, and developing the final synthesis used in Chapter 2. Further details on generative-AI use, verification, and researcher responsibility are provided in Appendix E.

The literature was synthesised thematically rather than discussed publication by publication in Chapter 2. The principal literature streams concerned self-disclosure

and its dimensions, recommendations and conversational steering, textual anthropomorphic cues, and conversation position. The literature matrix recorded each publication separately and linked it to the relevant themes and corresponding sections of Chapter 2. Within these streams, studies were compared in terms of their empirical findings, research context, methodological approach, relevance to the thesis, and limitations.

During synthesis, the evidence was compared to identify areas of convergence, divergence, uncertainty, and remaining limitations. The methodological and contextual considerations described in Section A.4 informed how cautiously individual findings were interpreted during synthesis. No formal weighting or quality-scoring system was applied. The thematic streams were subsequently integrated in Section 2.6 to establish what the reviewed literature collectively supported, where findings remained conditional or uncertain, and which relationships remained insufficiently addressed. This synthesis informed the formulation of the research gap and the conceptual framework presented in Chapter 2.

Appendix B: Codebook

This appendix presents the codebook used to code the <AI, user> turn pairs in the final sample. It provides the variables, code values, definitions, and decision rules used for recommendations, textual anthropomorphic cues, and self-disclosure together with synthetic examples created solely to illustrate the coding rules. All examples are synthetic and do not reproduce content from the donated conversations.

B.1 Recommendation Coding

Following the operationalisation described in Section 3.6.1.1, recommendation is used in this thesis as a study-specific umbrella for model-initiated conversational moves that can direct the immediately subsequent user response. Conventional advice and recommendation research generally treats recommendations as messages that propose or favour a course of action or choice (Feng, 2009; Fitzsimons & Lehmann, 2004). This action-oriented definition does not capture all ways in which ChatGPT can direct the next part of an interaction. Follow-up questions, reflective questions, ongoing-support offers, suggestive prompts, and suggestive questions can also influence what is discussed next, although these forms are not theoretically equivalent to conventional advice (Feng, 2009; Karnam et al., 2026; Xiao et al., 2020). The coding scheme therefore grouped these forms under a study-specific recommendation construct while preserving their distinct conversational functions through the *recommendation_type* variable.

Three recommendation variables were coded within each <AI, user> turn pair. *recommendation_num_made* recorded the number of distinct recommendations in the ChatGPT message, while *recommendation_num_followed* recorded how many of these were observably taken up in the immediately following user response. *recommendation_type* recorded the type or types of recommendation present in the ChatGPT message. In this coding scheme, recommendation uptake therefore refers only to whether the immediately following user response observably followed, answered, or acted on a recommendation made in the preceding ChatGPT message. It does not establish whether the user agreed with, accepted, or acted on the recommendation outside the conversation. An overview of the three recommendation variables and their coding rules is provided in Table B1.

Only recommendations that could be observably taken up within the immediately subsequent user response were coded. Suggestions requiring action outside the conversation or beyond the immediately following response were excluded. This restriction ensured that recommendation presence and observable uptake were defined within the same <AI, user> turn-pair boundary.

Table B1

Recommendation Variables and Coding Rules

Variable	Operational Definition	Codes	Decision Rule
<i>recommendation_num_made</i>	Number of distinct recommendations in the ChatGPT message.	NA; 0–10	Count each distinct recommendation according to Table B2 and Table B3. Code 0 when a usable ChatGPT message contains no recommendation and NA when recommendation coding cannot be performed.
<i>recommendation_num_followed</i>	Number of recommendations from the ChatGPT message that are observably taken up in the immediately following user response.	NA; 0–10	Count each recommendation that is observably taken up according to Table B3. Code 0 when at least one recommendation was present but none was taken up, and NA when no recommendation was available to follow or uptake could not be coded.
<i>recommendation_type</i>	Recommendation-type value or values assigned to the ChatGPT message.	NA; No Recommendation; Follow-up Question; Ongoing Support Offer; Reflective Question; Suggestive Prompt; Suggestive Question	Record each applicable type once using Table B2. Record multiple applicable types separated by a vertical bar ().

The 0–10 range used for *recommendation_num_made* and *recommendation_num_followed* was an operational interface constraint rather than a theoretical maximum. Preliminary manual screening did not identify ChatGPT messages containing more than ten codable recommendations. A fixed 0–10 dropdown was therefore used to standardise manual data entry and reduce errors associated with free-text numerical input.

Table B2 provides the decision rules used to distinguish the values of *recommendation_type*. Classification was based on the conversational function of the expression in the context. Multiple values could be assigned when different recommendation types occurred within the same ChatGPT message.

Table B2

Recommendation-Type Decision Rules

Code	Decision Rule	Synthetic Example
NA	Select when: the <i>recommendation_type</i> variable cannot be coded because no usable ChatGPT message is available. Do not select when: a usable ChatGPT message is available but contains none of the defined recommendation types; use <i>No Recommendation</i>.	No usable ChatGPT text is available for the turn pair.
No Recommendation	Select when: a usable ChatGPT message contains none of the recommendation types defined below. Do not select when: at least one defined recommendation type is present.	“That sounds like a difficult situation.”
Follow-up Question	Select when: ChatGPT requests additional information, clarification, or contextual detail about the user's situation or a previous statement. Do not select when: the primary function is to encourage reflection on feelings, thoughts, needs, or meaning; use <i>Reflective Question</i>. If the question proposes an action or direction, use <i>Suggestive Question</i>.	“What happened immediately before you started feeling this way?”
Ongoing Support Offer	Select when: ChatGPT expresses continued availability or generally invites the user to continue or return to the conversation. Do not select when: ChatGPT proposes a specific action, task, service, or conversational direction; use the applicable suggestive code.	“I'm here if you want to talk more about this.”

Code	Decision Rule	Synthetic Example
Reflective Question	Select when: ChatGPT invites the user, through a question, to reflect on their feelings, thoughts, needs, personal meaning, interpretation of the situation, or on information provided by ChatGPT in the same message. Do not select when: the question primarily requests factual or contextual information; use <i>Follow-up Question</i>. If it proposes an action or direction, use <i>Suggestive Question</i>.	“What do you think makes this situation feel so painful for you?”
Suggestive Prompt	Select when: ChatGPT proposes an action, strategy, or next step without presenting it as a question. Do not select when: the message only provides information or acknowledgement without proposing what the user could do. If the suggestion is formulated as a question, use <i>Suggestive Question</i>.	“You could write down what you want to say before speaking with them.”
Suggestive Question	Select when: ChatGPT proposes a possible action, service, or conversational direction through a question. Do not select when: the question only requests information or clarification; use <i>Follow-up Question</i>. If it invites reflection without proposing an action or direction, use <i>Reflective Question</i>.	“Would you like me to create a summary that you could show your therapist?”

Table B3 specifies how distinct recommendations were counted in the ChatGPT message and how observable uptake was counted in the immediately following user response. These rules distinguish recommendation counts from recommendation-type categories and restrict uptake coding to the paired user response.

Table B3*Recommendation Counting and Uptake Rules*

Coding Situation	Counting or Uptake Rule	Synthetic Example
One distinct recommendation is present.	Count one recommendation made.	“What happened immediately before you started feeling this way?” → recommendation_num_made = 1; recommendation_type = Follow-up Question.
Multiple distinct recommendations are present.	Count each distinct recommendation separately.	“What happened immediately before you started feeling this way? Would you like me to draft a reply?” → recommendation_num_made = 2; recommendation_type = Follow-up Question Suggestive Question.
Multiple recommendations belong to the same type.	Count each distinct recommendation separately in <i>recommendation_num_made</i> , but record the shared recommendation type only once in <i>recommendation_type</i> .	Two distinct follow-up questions → recommendation_num_made = 2; recommendation_type = Follow-up Question.
The same recommendation is restated without introducing a new action or direction.	Count the recommendation once when subsequent wording merely restates or elaborates the same recommendation without introducing a distinct additional action or direction.	“You could write your thoughts down. Putting them on paper may help.” → recommendation_num_made = 1.
Several alternatives form a single choice or invitation.	Count the combined invitation as one recommendation when the alternatives form one decision point.	“We can talk about what happened with your friend or about your sister. Which would you prefer?” → recommendation_num_made = 1.

Coding Situation	Counting or Uptake Rule	Synthetic Example
The user observably takes up a recommendation.	Count the recommendation as followed when the immediately following user response performs the requested action, answers the question, or continues in the proposed conversational direction.	ChatGPT asks what happened; the user explains the event → <code>recommendation_num_followed = 1</code> .
The user takes up only some of multiple recommendations.	Count only the recommendations that are observably taken up.	ChatGPT asks for an explanation and offers to draft a message; the user only explains → <code>recommendation_num_followed = 1</code> .
The user takes up none of the recommendations.	Code <code>recommendation_num_followed = 0</code> .	ChatGPT suggests discussing the user's feelings; the user changes to an unrelated topic → <code>recommendation_num_followed = 0</code> .
ChatGPT proposes an action that cannot be completed or observably followed within the immediately following user response.	Do not count the suggestion as a recommendation. Only recommendations that could be taken up within the paired user response were included.	“You should discuss this with your doctor tomorrow.” → not counted as a recommendation.

B.2 Textual Anthropomorphic Cue Coding

Following the operationalisation described in Section 3.6.1.2, anthropomorphism was coded through observable textual features in the ChatGPT message of each <AI, user> turn pair. Textual anthropomorphic cues were defined as expressions through which ChatGPT presented itself as possessing human-like agency, thoughts, emotions, intentions, judgements, or social roles (Greilich et al., 2025; Li et al., 2023). The coding therefore captures how ChatGPT was linguistically presented rather than whether the user actually perceived ChatGPT as human-like (Li et al., 2023).

Two variables were coded: *anthropomorphic_cues*, which recorded the number of identifiable textual anthropomorphic cue instances in the ChatGPT message, and *anthropomorphic_type*, which recorded the distinct anthropomorphic types represented by

those cue instances. Each cue instance contributed separately to *anthropomorphic_cues*, including repeated occurrences of the same type, whereas each anthropomorphic type was recorded only once per ChatGPT message. An overview of both variables and their coding rules is provided in Table B4.

Table B4

Textual Anthropomorphism Variables and Coding Rules

Variable	Operational Definition	Allowed Values	Decision Rule
<i>anthropomorphic_cues</i>	Number of identifiable textual anthropomorphic cue instances in the ChatGPT message.	NA; 0–10	Count each identifiable cue instance through which ChatGPT attributes human-like agency, thoughts, emotions, intentions, judgements, or social roles to itself. Repeated occurrences are counted separately, including repeated occurrences of the same cue type (Table B5). When one expression independently communicates multiple anthropomorphic meanings that satisfy different type definitions, count each meaning as a separate cue instance. Code 0 when a usable ChatGPT message contains no anthropomorphic cue and NA when no usable ChatGPT message is available.
<i>anthropomorphic_type</i>	Distinct anthropomorphic type or types represented in the ChatGPT message.	NA; No Anthropomorphism; Agency Framing; Belief Framing; Collaborative Agency; Emotional Attribution; Human-like Intention; Human-like Judgment; Social Role Framing	Record each applicable anthropomorphic type once per ChatGPT message using Table B5, regardless of how often that type occurs. When different types are present, separate them with a vertical bar ().

The 0–10 range used for *anthropomorphic_cues* was an operational interface constraint rather than a theoretical maximum. Preliminary manual screening did not identify ChatGPT messages containing more than ten identifiable anthropomorphic cue instances. A fixed 0–10 dropdown was therefore used to standardise manual data entry and reduce errors associated with free-text numerical input.

Table B5 provides the decision rules used to distinguish the values of *anthropomorphic_type*. Classification was based on the anthropomorphic meaning expressed in context rather than on the presence of first-person words such as “I” or “we” alone. When distinct anthropomorphic meanings occurred within the same ChatGPT message, multiple type codes could be assigned.

Table B5

Textual Anthropomorphic Cue-Type Decision Rules

Code/value	Decision Rule	Synthetic Example
NA	Select when: no usable ChatGPT message is available and <i>anthropomorphic_type</i> therefore cannot be coded. Do not select when: a usable message is available but contains no anthropomorphic cue; use <i>No Anthropomorphism</i> .	No usable ChatGPT message is available.
No Anthropomorphism	Select when: a usable ChatGPT message does not attribute human-like agency, thoughts, emotions, intentions, judgements, or social roles to itself. Do not select when: at least one defined anthropomorphic cue is present.	“Here are three possible options.”
Agency Framing	Select when: ChatGPT presents itself as an entity capable of acting, helping, guiding, noticing, explaining, or performing a task. Do not select when: the message only provides information without presenting ChatGPT as an acting entity, or primarily positions ChatGPT relationally as a social interaction partner.	“I can help you organise these options.”
Belief Framing	Select when: ChatGPT attributes a belief, thought, view, or opinion to itself. Do not select when: information is presented without framing it as ChatGPT’s own belief or thought.	“I think this situation sounds emotionally exhausting.”

Code/value	Decision Rule	Synthetic Example
Collaborative Agency	Select when: ChatGPT frames itself and the user as jointly undertaking an activity, process, or next step. Do not select when: only ChatGPT or only the user is presented as acting, without a shared frame. General relational presence alone is coded as <i>Social Role Framing</i>.	"We can work through these options one step at a time."
Emotional Attribution	Select when: ChatGPT attributes emotional understanding, care, concern, empathy, gladness, sadness, or another emotional response to itself. Do not select when: the message only identifies or describes the user's emotions without attributing emotional involvement to ChatGPT.	"I'm sorry you're going through this."
Human-like Intention	Select when: ChatGPT attributes a want, hope, aim, intention, or personally framed desired outcome to itself. Do not select when: the message merely describes an action ChatGPT can perform without expressing a want, hope, or intention.	"I want to help you make sense of what happened."
Human-like Judgment	Select when: ChatGPT expresses a personal evaluation, judgement, or assessment of the user, the user's actions, or the user's situation. Do not select when: the message provides factual, descriptive, or procedural information without a personal evaluative stance.	"You were very brave to talk about this."
Social Role Framing	Select when: ChatGPT positions itself as a socially present interaction partner, such as a listener, supporter, helper, companion, or conversational partner. Do not select when: the expression only communicates a belief, judgement, emotion, intention, or ability without positioning ChatGPT relationally within the interaction.	"I'm here to listen."

Table B6 specifies how individual textual anthropomorphic cue instances were counted within a ChatGPT message.

Table B6*Counting Rules for Textual Anthropomorphic Cues*

Coding Situation	Coding Rule	Synthetic Example
No usable ChatGPT message is available	Code <i>anthropomorphic_cues</i> = NA and <i>anthropomorphic_type</i> = NA.	The ChatGPT message is unavailable.
A usable ChatGPT message contains no anthropomorphic cue instance.	Code <i>anthropomorphic_cues</i> = 0 and <i>anthropomorphic_type</i> = <i>No Anthropomorphism</i> .	“A useful next step is to write down your thoughts.”
One anthropomorphic cue instance is present.	Count one cue and record the corresponding anthropomorphic type.	“I’m here to listen.” → <i>anthropomorphic_cues</i> = 1; <i>anthropomorphic_type</i> = Social Role Framing.
Two anthropomorphic cue instances of different types are present.	Count each cue instance separately and record each applicable type.	“I’m sorry this feels so difficult, and we can work through it together.” → <i>anthropomorphic_cues</i> = 2; <i>anthropomorphic_type</i> = Emotional Attribution Collaborative Agency.
The same anthropomorphic type occurs in separate cue instances.	Count each occurrence separately, but record the shared type only once in <i>anthropomorphic_type</i> .	“I can explain the first option. I can also help you compare the alternatives.” → <i>anthropomorphic_cues</i> = 2; <i>anthropomorphic_type</i> = Agency Framing.
One expression communicates several distinct anthropomorphic meanings.	Count each independently identifiable anthropomorphic meaning as a separate cue instance when each satisfies a distinct type definition in Table B5.	“I’m glad you told me, and I want to help you think this through.” → <i>anthropomorphic_cues</i> = 2; <i>anthropomorphic_type</i> = Emotional Attribution Human-like Intention.

B.3 User Self-Disclosure Coding

Self-disclosure was coded in the user response of each <AI, user> turn pair. It was operationalised as information through which users revealed information about their own experiences, opinions, attitudes, feelings, relationships, or circumstances (Omarzu,

2000). Statements concerning another person were treated as self-disclosure only when they also communicated information about the user's own situation, experience, or relationship with that person (Harmsen et al., 2023).

Self-disclosure was examined across three dimensions: depth, breadth, and volume. Depth captures the intimacy or sensitivity of the disclosed information, breadth captures the range of personal content areas disclosed (Carpenter & Greene, 2015; Omarzu, 2000), and volume captures the amount of written content identified as self-disclosing (Harmsen et al., 2023; Omarzu, 2000). These dimensions were operationalised through four substantive coding variables: *depth_code*, *breadth_amount*, *breadth_type*, and *volume*. An overview of these variables is provided in Table B7.

Table B7

User Self-Disclosure Variables and Coding Rules

Variable	Operational Definition	Allowed Codes	Decision Rule
<i>depth_code</i>	Highest level of personal intimacy or sensitivity clearly present in the user response.	Superficial; Middle; Inner; Core	Assign the highest depth level clearly supported by the disclosed meaning in context. Depth concerns the intimacy or sensitivity of the information rather than the number of subjects or words disclosed (Carpenter & Greene, 2015). If no self-disclosing information is identified, code <i>depth_code</i> = <i>Superficial</i> . Detailed decision rules are provided in Table B8.
<i>breadth_amount</i>	Number of distinct coded personal content areas represented in the user response.	0–10	Count each applicable predefined breadth type once. When one or more uncategorised personal content areas are present, record <i>Other</i> once in <i>breadth_type</i> , but count each distinct uncategorised area separately in <i>breadth_amount</i> . Repetition or further elaboration of the same coded personal content area does not increase the count. Code 0 when no personal content area is identified.

Variable	Operational Definition	Allowed Codes	Decision Rule
<i>breadth_type</i>	Distinct breadth type or types represented in the user response.	NA; predefined breadth-type categories; Other	Record each applicable predefined breadth type once per user response using Table B9. When multiple predefined types are present, separate them with a vertical bar (). When one or more uncategorised personal content areas are present, record <i>Other</i> once, regardless of how many distinct uncategorised areas are counted in <i>breadth_amount</i> . When no breadth type is identified and <i>breadth_amount</i> = 0, code <i>breadth_type</i> = NA.
<i>volume</i>	Number of words identified as communicating self-disclosing information in the user response.	Numerical count ≥ 0	Identify the words or phrases that communicate personal information about the user and count the words within those spans. Code 0 when no words are identified as self-disclosing. Detailed counting rules are provided in Table B11.

The 0–10 range used for *breadth_amount* was an operational interface constraint rather than a theoretical maximum. Preliminary manual screening did not identify user responses containing more than ten distinct coded personal content areas. Although a value of 0 was permitted by the coding scheme when no personal content area was identified, no user responses in the final analytical dataset received a *breadth_amount* of 0.

Depth

Depth represents the intimacy or sensitivity of the information disclosed by the user. It was coded using four ordered levels: *Superficial*, *Middle*, *Inner*, and *Core*, based on the layered understanding of disclosure in Social Penetration Theory (Carpenter & Greene, 2015). When several depth levels occurred within one response, the highest level clearly supported by the disclosed information was assigned. Responses in which no self-disclosing information was identified were coded as *Superficial*. The Superficial category therefore includes both responses without identified self-disclosure and responses containing only general, public, situational, or minimally personal information. Table B8 specifies the decision rules used to distinguish these four levels.

Table B8*Self-Disclosure Depth Decision Rules*

Depth code	Select this code when...	Do not select this code when...	Synthetic example
<i>Superficial</i>	The response contains no identified self-disclosing information, or contains only general, public, situational, or minimally personal information without revealing a clearly private or sensitive aspect of the user.	The response communicates clearly personal daily-life information, opinions, relationships, emotions, concerns, or sensitive experiences.	"I'm going to take a shower and watch a movie."
<i>Middle</i>	The response communicates personal but relatively non-sensitive information about daily life, preferences, opinions, ordinary experiences, family, friendships, work, or relationships.	The disclosure provides substantial insight into the user's emotional or psychological inner state or contains highly vulnerable information.	"I had a difficult week at work."
<i>Inner</i>	The response communicates clearly personal emotions, psychological difficulties, worries, needs, relationship conflicts, sensitive experiences, or other information providing meaningful insight into the user's inner state.	The information is primarily ordinary or mildly personal; use <i>Core</i> when the disclosure is clearly highly vulnerable or among the user's most sensitive personal information.	"My wife is excluding me from the family trip, and I feel really stressed and hurt about it."
<i>Core</i>	The response communicates highly intimate, vulnerable, or sensitive information, such as severe psychological distress, suicidal thoughts, deeply personal trauma, intense shame, or fundamental fears or negative beliefs about the self.	The response contains personal emotions or difficulties but does not reach this degree of vulnerability or sensitivity.	"I feel completely worthless and sometimes wonder whether anyone would care if I disappeared."

Depth was determined from the disclosed meaning in context rather than from the breadth category alone. The same general subject could therefore receive different depth codes depending on the intimacy or sensitivity of what the user disclosed (Carpenter

& Greene, 2015; Omarzu, 2000). For example, “My wife and I have argued several times this week” illustrates *Middle* disclosure, whereas “My wife and I have argued several times this week, and I feel rejected and afraid our relationship is falling apart” illustrates *Inner* disclosure because the latter provides substantially greater insight into the user’s emotional state. Similarly, “I have been feeling anxious and overwhelmed lately” illustrates *Inner* disclosure, whereas “I feel completely worthless and sometimes wonder whether anyone would care if I disappeared” illustrates *Core* disclosure because it communicates substantially greater vulnerability and sensitivity. The depth level was therefore assigned from the full meaning of the response rather than from the presence of a particular breadth type alone.

Breadth

Breadth represents the range of personal content areas disclosed within the user response (Carpenter & Greene, 2015; Omarzu, 2000). It was operationalised through *breadth_amount* and *breadth_type*. *breadth_type* recorded which predefined personal content-area categories were represented in the response, whereas *breadth_amount* recorded the number of distinct coded personal content areas. For predefined categories, each category contributed a maximum of one to *breadth_amount*, regardless of how often content belonging to that category occurred within the same response. The *Other* category functioned as an umbrella label for uncategorised content: *Other* was recorded only once in *breadth_type*, but each distinct uncategorised personal content area could contribute separately to *breadth_amount*. *breadth_amount* should therefore be interpreted as the range of coded personal content areas rather than as the number of *breadth_type* labels or a literal count of conversational topics.

Breadth types were assigned independently according to the decision rules in Table B9. When one disclosure satisfied more than one predefined breadth-type definition, all clearly applicable types were recorded. For example, “I feel lonely” was coded as Feelings | Loneliness and therefore received *breadth_amount* = 2. Repeated mentions or elaborations belonging to the same predefined breadth type did not increase *breadth_amount*. For content coded as *Other*, however, distinct uncategorised personal content areas were counted separately in *breadth_amount* even though *Other* was recorded only once in *breadth_type*. When the applicable breadth type was ambiguous, only the type or types clearly supported by the disclosed meaning and conversational context were assigned.

Table B9*Self-Disclosure Breadth-Type Decision Rules*

Breadth Type Codes	Decision Rule	Synthetic Example
<i>Parents</i>	Select when the disclosure concerns the user's mother, father, parent, or parental relationship. Use <i>Family</i> instead when the disclosure concerns the family generally without identifying a specific parental relationship.	"My mother never listens when I tell her how I feel."
<i>Grandparents</i>	Select when the disclosure concerns the user's grandmother, grandfather, or relationship with a grandparent.	"I still miss my grandmother a lot."
<i>Siblings</i>	Select when the disclosure concerns the user's brother, sister, or sibling relationship.	"My sister and I barely talk anymore."
<i>Children</i>	Select when the disclosure concerns the user's child, children, or relationship with them.	"I worry that my daughter won't pass to the next school year."
<i>Family</i>	Select when the disclosure concerns the user's family generally or a family dynamic not represented by a more specific family code.	"My family expects me to solve everyone's problems."
<i>Husband</i>	Select when the disclosure specifically concerns the user's husband or marital relationship with him.	"My husband and I keep arguing about this."
<i>Wife</i>	Select when the disclosure specifically concerns the user's wife or marital relationship with her.	"My wife has been distant from me lately."
<i>Friends</i>	Select when the disclosure concerns the user's friends, friendships, or lack of friendship.	"I don't really have friends I can talk to."
<i>Ex-Relationship Partner</i>	Select when the disclosure concerns a former romantic partner or the user's relationship with that person.	"I still think about my ex every day."

Breadth Type Codes	Decision Rule	Synthetic Example
<i>Romantic Relationships</i>	Select when the disclosure concerns the user's romantic relationship generally and is not represented more specifically by another relationship code.	"My relationship has been unstable lately."
<i>Romantic/Crush Interest</i>	Select when the disclosure concerns someone in whom the user has a romantic interest or crush.	"I like someone, but I'm scared to tell them."
<i>Girl-/Boyfriend</i>	Select when the disclosure concerns the user's girlfriend, boyfriend, or relationship with them.	"My boyfriend doesn't understand my anxiety."
<i>Dating</i>	Select when the disclosure concerns the user's dating experiences or dating situation.	"Dating makes me feel insecure."
<i>Kissing</i>	Select when the disclosure concerns kissing or an experience involving kissing.	"I kissed someone yesterday and now I feel confused."
<i>Sex</i>	Select when the disclosure concerns sex or sexual experiences.	"I feel uncomfortable about what happened sexually."
<i>Divorce</i>	Select when the disclosure concerns divorce, separation, or the effects of a divorce on the user.	"My divorce still affects me."
<i>Loneliness</i>	Select when the user discloses feeling alone, isolated, disconnected, or lacking meaningful social contact.	"I feel alone even when other people are around."
<i>Feelings</i>	Select when the user discloses their own emotional state, such as sadness, anxiety, fear, shame, stress, happiness, or love. Do not select merely because an emotional term appears without describing the user's own emotional state.	"I feel anxious whenever I have to go back there."

Breadth Type Codes	Decision Rule	Synthetic Example
<i>Self Image/Understanding</i>	Select when the disclosure concerns the user's identity, self-worth, self-image, or understanding of themselves.	"I keep thinking that I'm a failure."
<i>Mental Illness</i>	Select when the disclosure concerns a psychological condition, diagnosis, symptoms, or ongoing psychological difficulty.	"My depression has been much worse recently."
<i>Trauma</i>	Select when the disclosure concerns a traumatic experience or trauma-related effects.	"I still have nightmares about what happened."
<i>Suicide</i>	Select when the disclosure concerns suicidal thoughts, self-harm intentions, or not wanting to live.	"Sometimes I don't want to be alive anymore."
<i>Therapy</i>	Select when the disclosure concerns therapy, counselling, or professional mental-health support.	"I talked about this with my therapist."
<i>Health/Medication</i>	Select when the disclosure concerns the user's physical health, medication, treatment, or other medical issues.	"My medication makes me feel numb."
<i>Irritations and Frustrations</i>	Select when the disclosure concerns the user's irritation, frustration, anger, or annoyance.	"I get frustrated when nobody takes me seriously."
<i>Future Desires</i>	Select when the disclosure concerns the user's hopes, plans, goals, or desired future circumstances.	"I want to move away and start over."
<i>Dreams</i>	Select when the disclosure concerns dreams experienced while sleeping.	"I keep having the same dream about my childhood."
<i>Death/Funeral</i>	Select when the disclosure concerns death, grief, funerals, or personal loss.	"My father's funeral still feels unreal."

Breadth Type Codes	Decision Rule	Synthetic Example
<i>House maintenance/moving</i>	Select when the disclosure concerns the user's housing situation, moving, or home-related matters.	"Moving between houses has made me very anxious."
<i>Work/Colleagues</i>	Select when the disclosure concerns the user's work, job, colleagues, workplace, or work-related stress.	"My colleagues make me feel excluded."
<i>Addiction</i>	Select when the disclosure concerns addiction or dependency.	"I think I'm becoming dependent on alcohol."
<i>Drugs</i>	Select when the disclosure concerns the user's use of drugs.	"I used drugs again last weekend."
<i>Crafting a Message</i>	Select when the disclosed content concerns a message the user is writing or preparing with ChatGPT's help.	"Help me write to my mother that I'm hurt."
<i>Other</i>	Select when one or more self-disclosing personal content areas do not fit any of the established breadth-type categories. Use only after the established categories have been considered. Record <i>Other</i> only once in <i>breadth_type</i> , regardless of how many distinct uncategorised personal content areas are present. Each distinct uncategorised personal content area may nevertheless contribute separately to <i>breadth_amount</i> according to Table B10.	A personal disclosure concerning an uncategorised life domain.

Table B10 specifies how *breadth_amount* was counted within each user response. The count reflects the number of distinct coded personal content areas. Each applicable predefined breadth type contributed a maximum of one to *breadth_amount*, whereas distinct uncategorised personal content areas represented under *Other* could contribute separately even though *Other* was recorded only once in *breadth_type*.

Table B10*Self-Disclosure Breadth Counting Rules*

Coding Situation	Coding Rule	Synthetic Example
No personal content area is identified	Code <i>breadth_amount</i> = 0 and <i>breadth_type</i> = NA.	“Can you explain what depression means?” → <i>breadth_amount</i> = 0.
One predefined breadth type is identified	Record the applicable predefined type once and add one to <i>breadth_amount</i> .	“Work has been stressful lately.” → Work/Colleagues → <i>breadth_amount</i> = 1.
Several predefined breadth types are identified	Record each applicable predefined type once. Each distinct predefined type contributes one to <i>breadth_amount</i> .	“I feel anxious about work and I am considering therapy.” → Feelings Work/Colleagues Therapy → <i>breadth_amount</i> = 3.
The same predefined breadth type occurs several times	Count the predefined breadth type once. Repetition, clarification, or further detail belonging to the same coded personal content area does not increase <i>breadth_amount</i> .	“Work has been stressful. My workload keeps increasing and my colleagues expect too much.” → Work/Colleagues → <i>breadth_amount</i> = 1.
One disclosure satisfies more than one predefined breadth-type definition	Record each clearly applicable predefined type once; each contributes one to <i>breadth_amount</i> .	“I feel lonely.” → Feelings Loneliness → <i>breadth_amount</i> = 2.
One or more uncategorised personal content areas are identified	Record Other once in <i>breadth_type</i> . Count each distinct uncategorised personal content area separately in <i>breadth_amount</i> . Thus, two distinct uncategorised areas produce <i>breadth_type</i> = Other but contribute two to <i>breadth_amount</i> .	“I’m studying architecture at university, and I play violin in a local orchestra.” → <i>breadth_type</i> = Other; <i>breadth_amount</i> = 2.

Coding Situation	Coding Rule	Synthetic Example
Predefined and uncategorised personal content areas occur together	Count each applicable predefined breadth type once and each distinct uncategorised personal content area separately. Record <i>Other</i> only once in <i>breadth_type</i> .	“My sister and I barely talk anymore, I’m studying architecture at university, and I play violin in a local orchestra.” → breadth_type = Siblings Other; breadth_amount = 3.
A predefined breadth type is implied rather than directly named	Count the type when the meaning of the response and conversational context clearly support the corresponding <i>breadth_type</i> code.	A response clearly describing an ongoing therapeutic session may therefore be coded as Therapy even when the word “therapy” is not used.

Volume

Volume represents the amount of self-disclosing text in the user response. Self-disclosure research distinguishes the amount or duration of disclosure from its depth and breadth (Omarzu, 2000), while conversational-agent research has used word count to capture response length (Harmsen et al., 2023). In this study, volume was operationalised more narrowly as the number of words within text identified as self-disclosing. When the same personal information was repeated within the same response without adding new information, the repeated content was excluded from the volume count. If the repeated wording added new personal information, this additional information was included in the volume count.

Table B11 specifies the decision rules used to identify and count self-disclosing words, including cases involving mixed self-disclosing and non-self-disclosing content, repeated information, and additional personal detail.

Table B11*Self-Disclosure Volume Decision Rules*

Coding Situation	Coding Rule	Synthetic Example
Word-count unit	For volume coding, count words as space-separated text units. A contraction or hyphenated expression without spaces counts as one word.	“don’t” = 1 word; “do not” = 2 words; “well-being-wise” = 1 word.
No self-disclosing information is present.	Code <i>volume</i> = 0.	“Can you explain what anxiety means?” → <i>volume</i> = 0.
A short self-disclosing span is present.	Highlight the words that communicate personal information about the user and count only those highlighted words.	“I feel anxious.” → <i>volume</i> = 3.
A long disclosure concerns one personal content area.	Highlight all words belonging to substantive self-disclosing text that communicate new personal information within the response. Repetition of information already disclosed in that response does not increase <i>volume</i> .	A detailed account of loneliness may have a high volume score while remaining focused on one personal content area.
Several personal content areas are disclosed.	Highlight the self-disclosing text associated with all applicable personal content areas. <i>Breadth</i> and <i>volume</i> are coded independently.	“I feel lonely, my mother ignores me, and work is too much.” contains several separate self-disclosing spans.
Another person is mentioned.	Count the words only when the statement also communicates information about the user’s own relationship, situation, or experience.	“My brother is ill and I am scared” communicates information about the user’s family situation and emotional state; “Some people have ill brothers” does not.
A statement is vague or general.	Count it only when it clearly communicates information about the user’s own experience.	“I feel bad all the time” is self-disclosing; “That sounds bad” is not.

Coding Situation	Coding Rule	Synthetic Example
The response contains non-disclosing surrounding text.	Exclude surrounding instructions, general discussion, repetitions of ChatGPT's wording, or question framing when those words do not themselves communicate information about the user.	In "Can you explain why this happens? I feel anxious every morning," only "I feel anxious every morning" contributes to volume.
The same self-disclosing information is repeated within the same user response.	Count the disclosed information only once. Repetition of information already disclosed within that response does not increase <i>volume</i> . If the same information is disclosed again in a later user response, it is coded again for that later turn pair.	"I feel lonely. I really feel lonely." → the repeated statement does not contribute a second time to volume.

The coding interface additionally stored *volume_word_indices*, which recorded the positions of the words identified as self-disclosing. This was a technical traceability field rather than a separate measure of self-disclosure and allowed the coded disclosure spans underlying the volume score to be reconstructed and checked.

B.4 Coding Procedure

This section documents the detailed procedure used to apply the coding scheme in Sections B.1–B.3 to the final sample of <AI, user> turn pairs. It specifies the coding sequence, the contextual information available during coding, the conventions used for entering codes, and the procedures used to track coding completion and consistency. Detailed construct definitions and decision rules are provided in Tables B1–B11.

B.4.1 Unit of Analysis and Coding Context

The unit of coding was one <AI, user> turn pair, consisting of a ChatGPT message and the user response that immediately followed it. The number and type of recommendations and textual anthropomorphic cues were coded from the ChatGPT message. Observable recommendation uptake and the depth, breadth, and volume of self-disclosure were coded from the paired user response. The construction of the <AI, user> turn pair is described in Section 3.5.

Coding decisions were based on content present in the current <AI, user> turn pair. Earlier turns from the same conversation remained available when needed to interpret references, implied meanings, or statements that could not be interpreted reliably from the pair alone. Earlier context was used only to clarify content present in the current turn pair. Information appearing exclusively in earlier turns was not transferred into the coding of the current pair unless it was communicated again in the current user response.

B.4.2 Coding Sequence and Field Conventions

Each <AI, user> turn pair was coded using the same fixed sequence. The ChatGPT message was assessed first for textual anthropomorphic cues and recommendations. The immediately following user response was then assessed for observable recommendation uptake and self-disclosure. Within the user response, volume was coded first, followed by breadth amount and breadth type, and finally depth.

A distinction was maintained between empty fields, 0 values, and NA values. An empty field indicated that the variable had not yet been coded. A value of 0 indicated an observed absence for a variable for which zero was a valid value. NA indicated that the variable was not applicable or could not validly be coded under its predefined decision rules. For example, when a ChatGPT message contained no usable text, recommendation and textual anthropomorphic-cue variables could not be coded and were therefore recorded as NA. The variable-specific conditions for assigning 0 and NA are specified in Tables B1–B11.

B.4.3 Coding the ChatGPT Message

The ChatGPT message was coded for textual anthropomorphic cues and recommendations before the paired user response was assessed. For textual anthropomorphic cues, each identifiable cue instance satisfying a type definition in Table B5 was identified and counted according to Table B6. Repeated occurrences were counted separately, including repeated occurrences of the same cue type. When one expression independently communicated multiple anthropomorphic meanings that satisfied different type definitions, each meaning was counted as a separate cue instance. Each distinct anthropomorphic type present in the message was recorded once in *anthropomorphic_type*, regardless of how often that type occurred.

Recommendations were then identified using the definitions and decision rules in Tables B1–B3. Only recommendations that could be observably taken up within the immediately following user response were included. Suggestions requiring action outside the conversation or beyond the paired response were excluded. Each distinct eligible recommendation was counted once in *recommendation_num_made*, including multiple distinct recommendations of the same type, whereas each recommendation type was recorded once in *recommendation_type*. Restatements or elaborations of the same recommendation did not increase *recommendation_num_made* unless they introduced a distinct additional action or conversational direction. Recommendation uptake was assessed only after the paired user response was read, as described in Section B.4.4.

B.4.4 Coding the User Response

After the ChatGPT message had been coded, the immediately following user response was assessed for recommendation uptake and self-disclosure. Each recommendation identified in the preceding ChatGPT message was evaluated separately for observable uptake using the decision rules in Table B3, and the number showing uptake was recorded in *recommendation_num_followed*.

Self-disclosure was then coded in the operational sequence of volume, breadth, and depth. For volume, the coder selected the words that communicated personal information about the user. The number of selected words constituted the volume score, while their positions were stored in *volume_word_indices* to preserve the underlying coded text for later review. The detailed coding rules used for volume coding are specified in Table B11.

Breadth was subsequently coded by identifying the applicable predefined breadth types and any distinct uncategorised personal content areas. Each applicable predefined breadth type was recorded once in *breadth_type*. When one or more uncategorised personal content areas were present, *Other* was recorded once in *breadth_type*. *breadth_amount* was then determined from the number of distinct coded personal content areas: each applicable predefined breadth type contributed one, while each distinct uncategorised personal content area represented under *Other* contributed separately, following the counting rules in Table B10.

Finally, the complete user response was assigned a depth level using the decision rules in Table B8. The highest depth level clearly supported by the self-disclosing information in the response was recorded as *depth_code*.

B.4.5 Contextual and Ambiguous Cases

Ambiguous cases were resolved using the predefined definitions and decision rules in Tables B1–B11. When the meaning of content in the target turn pair depended on earlier conversational context, earlier turns were consulted only to clarify the meaning of content present in the target pair, as specified in Section B.4.1. Information from earlier turns was not introduced into the coding of the target pair unless it was communicated again in that pair. A code was assigned only when the content of the target turn pair, interpreted with earlier context where necessary, provided sufficient evidence for that code.

For breadth coding, a breadth type could therefore be assigned even when it was not explicitly named in the current user response if the response and preceding context made the relevant personal content area clear. All clearly applicable breadth types were retained, while *Other* was used only when self-disclosing information did not sufficiently fit any predefined breadth type in Table B9. Similarly, for the multi-code variables *recommendation_type*, *anthropomorphic_type*, and *breadth_type*, multiple values were retained when the target content independently satisfied more than one type definition rather than forcing the content into a single type.

B.4.6 Coding Completion and Consistency Checks

After all applicable coding fields for an <AI, user> turn pair had been completed, the pair was marked as reviewed. The reviewed status was used as a completion indicator during the coding process.

Before quantitative analysis, the completed coded dataset was checked for remaining empty fields and codes entered in incorrect fields. These checks supported coding completeness and internal consistency.

Appendix C:

Statistical Implementation Details

This appendix provides the technical implementation details for the quantitative analyses described in Sections 3.7.2–3.7.5. It documents the statistical procedures, outcome representations, effect-size calculations, multiple-testing corrections, regression and resampling procedures, conversation-clustered inference, and reproducibility settings used in the analyses.

C.1 Statistical Procedures by Analysis

Table C1 provides the mapping between the analyses described in Chapter 3 and the corresponding statistical procedures. Rank-based procedures were used for the principal numerical group comparisons because breadth amount and self-disclosure volume were strongly right-skewed and these procedures do not require normally distributed outcome values.

Table C1
Statistical Procedures Used Across the Analyses

Analysis	Depth test	Breadth amount test	Volume test
Exploratory depth-level overall comparison	—	Kruskal-Wallis	Kruskal-Wallis
Exploratory depth-level pairwise comparisons	—	Mann-Whitney U	Mann-Whitney U
IQR sensitivity analysis	—	Kruskal-Wallis	Kruskal-Wallis
RQ1 overall recommendation-group comparison	Permutation χ^2	Kruskal-Wallis	Kruskal-Wallis
RQ1 pairwise recommendation-group comparisons	Permutation χ^2	Mann-Whitney U	Mann-Whitney U
RQ1 adjusted analysis	Adjusted linear regression on ordered depth with conversation-level bootstrap	Adjusted linear regression with conversation-level bootstrap	Adjusted linear regression with conversation-level bootstrap

Analysis	Depth test	Breadth amount test	Volume test
RQ2 primary two-group comparison	Mann-Whitney U test on ordered depth	Mann-Whitney U	Mann-Whitney U
RQ2 adjusted analysis	Adjusted linear regression on ordered depth with conversation-level bootstrap	Adjusted linear regression with conversation-level bootstrap	Adjusted linear regression with conversation-level bootstrap
RQ3 overall and subgroup comparisons	Permutation χ^2	Permutation-based Kruskal-Wallis test	Permutation-based Kruskal-Wallis
RQ3 pairwise quartile comparisons	Permutation χ^2	Permutation-based Mann-Whitney U	Permutation-based Mann-Whitney U
RQ3 conversation quartile $\times$ cue-group interaction	Freedman-Lane permutation test using categorical depth indicators	Freedman-Lane permutation	Freedman-Lane permutation test on $\log_{10}(\text{volume})$
RQ3 conversation-level clustered analysis	Linear regression on ordered depth with conversation-clustered covariance	Linear regression with conversation-clustered covariance	Linear regression on $\log_{10}(\text{volume})$ with conversation-clustered covariance

Note. The analytical purpose and sequence of these procedures are described in Sections 3.7.2–3.7.5.

The exploratory depth-level sensitivity analysis repeated the overall Kruskal-Wallis comparisons after excluding potential outliers separately for breadth amount and volume. Potential outliers were identified within each depth level using the Tukey boxplot criterion of values below $Q1 - 1.5 \times IQR$ or above $Q3 + 1.5 \times IQR$ (Hubert & Vandervieren, 2008). The complete coded dataset remained the basis of the primary analysis.

C.2 Outcome Representations, Rank Handling, and Effect Sizes

Self-disclosure depth was represented differently across analyses because the coding scheme establishes an ordering from Superficial to Core but does not establish equal distances between adjacent categories.

For the RQ1 overall and pairwise comparisons and the primary RQ3 conversation-position and cue-context comparisons, the four depth categories were retained

separately. For the primary RQ2 comparison, the ordinal ordering of the four depth levels was used in a Mann-Whitney U test.

In the adjusted RQ1 and RQ2 analyses and the supplementary conversation-level clustered RQ3 analysis, depth was represented as an ordered score from 1 (Superficial) to 4 (Core). This representation treats successive category differences as numerically equal, although equal spacing is not established by the coding scheme. These models were therefore treated as adjusted or sensitivity specifications rather than as equivalent replacements for analyses retaining the categorical or ordinal depth structure.

For the RQ3 conversation quartile $\times$ cue-group interaction, depth was represented by four categorical indicator variables so that differences across the complete depth distribution contributed to the interaction test. Breadth amount remained in its original numerical form throughout the analyses. Self-disclosure volume remained in its original form except in the RQ3 interaction and conversation-level clustered analyses, where $\log_{1p}(\text{volume})$ was used to reduce the influence of the strongly right-skewed distribution while retaining observations with zero self-disclosing words. Here, $\log_{1p}(\text{volume})$ denotes

$$\ln(1 + \text{volume}) \quad (\text{Eq. C1})$$

where $\ln$ is the natural logarithm.

For rank-based procedures, tied observations received average ranks. The implemented Kruskal-Wallis calculations included an explicit correction for tied ranks. The asymptotic Mann-Whitney U procedures used for the RQ1 and RQ2 numerical comparisons relied on a large-sample approximation to the null distribution, with a tie-corrected variance and continuity correction.

Effect sizes were calculated from the observed data rather than from resampled distributions. Rank-biserial correlations for the reported Mann-Whitney U comparisons were calculated as

$$r_{rb} = \frac{2U}{n_1 n_2} - 1 \quad (\text{Eq. C2})$$

where U is the Mann-Whitney U statistic for the first-listed group, and n_1 and n_2 are the sizes of the first- and second-listed groups, respectively. Under this convention, positive values indicate relatively higher ranks in the first-listed group, whereas negative values indicate relatively higher ranks in the second-listed group. The final RQ2 analysis used this rank-biserial measure for ordered depth, breadth amount, and volume.

Cramér's V for categorical comparisons was calculated as

$$V = \sqrt{\frac{\chi^2}{N \min(r-1, c-1)}} \quad (\text{Eq. C3})$$

where N is the analysed sample size and r and c denote the numbers of rows and columns in the contingency table.

Epsilon squared for Kruskal-Wallis comparisons was calculated as

$$\epsilon^2 = \frac{H - k + 1}{N - k} \quad (\text{Eq. C4})$$

where H is the Kruskal-Wallis statistic, k is the number of groups, and N is the analysed sample size (Tomczak & Tomczak, 2014). Throughout this thesis, ϵ^2 refers to the H-based effect-size estimator defined in Eq. C4. This estimator is also denoted η_H^2 in parts of the statistical literature. The notation for Kruskal-Wallis effect-size estimators is not fully consistent across sources. Epsilon-squared estimates were not truncated at zero, so small negative sample estimates could occur.

For the regression-based RQ3 analyses, effect size was expressed as

$$\Delta R^2 = R_{\text{full}}^2 - R_{\text{reduced}}^2 \quad (\text{Eq. C5})$$

where R_{full}^2 and R_{reduced}^2 denote the variance explained by the corresponding full and reduced models. Thus, ΔR^2 represents the additional variance explained by adding the predictor or interaction of interest to the reduced model (Frazier et al., 2004). The specific full and reduced model specifications are described in Sections C.5.2 and C.5.3.

C.3 Multiple-Testing Corrections

Exploratory depth-level analysis. The Holm procedure was used to control for multiple comparisons within predefined families of related tests (Holm, 1979). Corrections were applied to the unrounded p-values generated by the corresponding analyses.

For the exploratory depth-level analysis, Holm correction was applied separately across the six pairwise depth-level comparisons for breadth amount and across the six corresponding comparisons for volume.

RQ1: Recommendation presence and observable uptake. The three overall tests for depth, breadth amount, and volume formed one correction family. Where an overall difference remained statistically supported after correction, the three pairwise recommendation-group comparisons were corrected separately within that outcome. For the adjusted RQ1 analyses, Holm correction was applied separately within each outcome across the three focal recommendation-group contrasts: Recommendation without observable uptake versus No recommendation, Recommendation with

observable uptake versus No recommendation, and Recommendation with observable uptake versus Recommendation without observable uptake. The coefficients for turn number and textual anthropomorphic-cue presence were not included in these correction families.

RQ2: Textual anthropomorphic cues. The three primary comparisons for ordered depth, breadth amount, and volume formed one correction family. For the adjusted analysis, Holm correction was applied across the three focal textual anthropomorphic-cue coefficients for ordered depth, breadth amount, and volume. The coefficients for turn number and recommendation presence were not included in this correction family.

RQ3: Primary conversation-position analysis. The three overall tests for depth, breadth amount, and volume formed one correction family. Where an overall difference remained statistically supported after correction, Holm correction was subsequently applied across the six pairwise conversation-quartile comparisons within that outcome.

RQ3: Conversation-length sensitivity analysis. The three overall outcome tests were corrected separately within each conversation-length stratum. Pairwise comparisons were conducted only for stratum $\times$ outcome combinations whose overall comparison remained statistically supported after correction. Within each eligible stratum $\times$ outcome combination, Holm correction was then applied across the six possible pairwise conversation-quartile comparisons.

RQ3: Cue-context and interaction analyses. The quartile-difference tests were corrected across the four cue groups separately within each outcome. The cue-group-difference tests were corrected across the four conversation quartiles separately within each outcome. The three conversation quartile $\times$ cue-group interaction tests for depth, breadth amount, and volume formed a separate correction family.

RQ3: Conversation-level clustered analysis. Holm correction was applied separately across the three conversation-position tests and across the three conversation quartile $\times$ cue-group interaction tests.

C.4 RQ1 and RQ2 Statistical Implementation

This section provides the technical implementation details for the RQ1 and RQ2 analyses described in Sections 3.7.3 and 3.7.4. It first specifies the primary and pairwise group comparisons, then the adjusted regression models, and finally the conversation-level bootstrap procedure used to evaluate uncertainty in the adjusted estimates.

C.4.1 Primary and Pairwise Comparisons

For the RQ1 primary three-group comparison of depth, the complete distribution across Superficial, Middle, Inner, and Core was compared using a permutation χ^2 test.

Recommendation-group labels were permuted while the observed depth classifications remained fixed, and the observed χ^2 statistic was compared with its Monte Carlo permutation distribution. The RQ1 pairwise depth comparisons used the same permutation procedure within the corresponding two recommendation groups.

For breadth amount and volume, the RQ1 primary three-group comparisons used Kruskal-Wallis tests. Where pairwise follow-up comparisons were applicable under the procedure specified in Section C.3, Mann-Whitney U tests were used to compare the corresponding recommendation-group pairs.

For the RQ2 primary two-group comparison, ordered depth, breadth amount, and volume were each compared between the cue-absence and cue-presence groups using Mann-Whitney U tests.

C.4.2 Adjusted Regression Models

For both RQ1 and RQ2, separate linear regression models were estimated for ordered depth, breadth amount, and volume. Turn number was mean-centred by subtracting its mean across the adjusted analytical dataset before inclusion in the models. No predictor or outcome was otherwise standardised before model estimation. Unstandardised regression coefficients (B) were reported so that the estimated associations remained expressed on the original outcome scales rather than representing the relative importance of predictors (Landis, 2014).

RQ1: Adjusted analysis. Recommendation group was the focal predictor, with No recommendation as the reference category. Turn number and textual anthropomorphic-cue presence were included as additional predictors. Because RQ1 distinguishes recommendations with and without observable uptake, a linear contrast derived from each fitted model additionally compared Recommendation with observable uptake with Recommendation without observable uptake. Because both recommendation-group coefficients were estimated relative to No recommendation, the adjusted direct contrast was calculated as

$$B_{\text{with uptake}} - B_{\text{without uptake}} \quad (\text{Eq. C6})$$

The same contrast was recalculated from the corresponding coefficients in each bootstrap resample.

RQ2: Adjusted analysis. Textual anthropomorphic-cue presence was the focal predictor, with No textual anthropomorphic cue as the reference category. Turn number and recommendation presence were included as additional predictors. Recommendation presence was coded irrespective of observable uptake because recommendation presence was defined from the preceding ChatGPT message, whereas observable uptake was classified from the subsequent user response.

C.4.3 Conversation-Level Bootstrap

Uncertainty in the adjusted RQ1 and RQ2 estimates was evaluated using conversation-level bootstrapping because multiple analytical turn pairs originated from the same conversation. Complete conversations were sampled with replacement, with the number of sampled conversations equal to the number of conversations in the original analytical sample. All turn pairs belonging to a selected conversation remained together within each bootstrap sample, thereby preserving the within-conversation structure during resampling (Huang, 2018). The number of bootstrap resamples and the corresponding random seed are reported in Section C.6.

For each focal bootstrap estimate distribution θ^* , where θ^* represented either a model coefficient or the RQ1 linear contrast defined in Section C.4.2, the two empirical tail probabilities were calculated as

$$p_+ = P(\theta^* \geq 0) \quad (\text{Eq. C7})$$

And

$$p_- = P(\theta^* \leq 0). \quad (\text{Eq. C8})$$

The two-sided raw bootstrap p-value was then calculated as

$$p = \min[1, 2 \min(p_+, p_-)]. \quad (\text{Eq. C9})$$

Zero was included in both empirical tails. No +1 Monte Carlo correction was applied to the bootstrap p-values. The unadjusted 95% percentile bootstrap confidence interval was defined by the 2.5th and 97.5th percentiles of the focal bootstrap estimate distribution.

Holm correction was subsequently applied to the unrounded focal bootstrap p-values according to the correction families specified in Section C.3. Statistical support was evaluated using the Holm-adjusted p-values at $\alpha = .05$. The percentile confidence intervals remained unadjusted measures of estimation uncertainty. Consequently, an unadjusted confidence interval could exclude zero while the corresponding Holm-adjusted p-value exceeded .05. In such cases, the association was not interpreted as statistically supported after multiple-testing correction.

C.5 RQ3 Statistical Implementation

This section provides the technical implementation details for the RQ3 analyses described in Section 3.7.5. It first specifies the permutation-based group comparisons used for the primary conversation-position, conversation-length sensitivity, and cue-context analyses. It then describes the conversation quartile $\times$ cue-group interaction procedure and, finally, the conversation-level clustered analysis used to account for dependence among turn pairs originating from the same conversation.

C.5.1 Permutation-Based Group Comparisons

The RQ3 primary conversation-position analysis, conversation-length sensitivity analysis, and cue-context analysis used permutation-based inference for the specified group comparisons.

For categorical depth comparisons, depth-category labels were permuted relative to the fixed conversation-quartile or cue-group labels. The χ^2 statistic was recalculated for each permutation and compared with the observed χ^2 statistic. This procedure was used for the permutation χ^2 tests of depth.

For numerical Kruskal-Wallis comparisons, the relevant conversation-quartile or cue-group labels were permuted while the numerical outcome values remained fixed. Tied observations retained average ranks, and the Kruskal-Wallis statistic incorporated the corresponding tie correction.

For pairwise numerical comparisons, observations from the two groups were pooled and randomly shuffled before being reassigned according to the original group sizes. The Mann-Whitney U statistic was then recalculated for the permuted groups. For these two-sided pairwise permutation tests, extremeness was defined by the absolute distance of the Mann-Whitney U statistic from its null expectation:

$$T_U = \left| U - \frac{n_1 n_2}{2} \right| \quad (\text{Eq. C10})$$

Where U is the Mann-Whitney U statistic and n_1 and n_2 are the two group sizes. The observed and permuted values of T_U were then compared using the Monte Carlo permutation procedure described below.

Monte Carlo permutation p-values used the +1 correction recommended for randomly sampled permutations (Phipson & Smyth, 2010):

$$p = \frac{\#(T_{\text{perm}} \geq T_{\text{obs}}) + 1}{B + 1} \quad (\text{Eq. C11})$$

where T_{obs} is the observed analysis-specific comparison statistic, T_{perm} is the corresponding statistic obtained under permutation, and B is the number of permutations. The same +1 Monte Carlo correction was used for the RQ1 permutation χ^2 tests. The number of permutations and random seeds used for the resampling procedures are reported in Section C.6.

C.5.2 Conversation Quartile × Cue-Group Interaction Analysis

The RQ3 conversation quartile × cue-group interaction analysis tested whether the pattern of self-disclosure across conversation quartiles differed among the four cue contexts.

For each self-disclosure outcome, a reduced model containing the main effects of conversation quartile and cue group was compared with a full model that additionally contained all conversation quartile $\times$ cue-group interaction terms. Model comparison used the Freedman-Lane residual-permutation procedure (Winkler et al., 2014).

The reduced model was first fitted to obtain fitted values and residuals. For each permutation, the residual rows were randomly permuted and added to the fitted values from the reduced model:

$$Y_{\text{perm}} = \hat{Y}_{\text{reduced}} + e_{\text{reduced,perm}} \quad (\text{Eq. C12})$$

where Y_{perm} is the permuted outcome, $\hat{Y}_{\text{reduced}}$ contains the fitted values from the reduced model, and $e_{\text{reduced,perm}}$ contains the permuted residuals from that model. The reduced and full models were then refitted to the permuted outcome. The model-comparison statistic was calculated as

$$F = \frac{(RSS_{\text{reduced}} - RSS_{\text{full}})/(p_{\text{full}} - p_{\text{reduced}})}{RSS_{\text{full}}/(N - p_{\text{full}})} \quad (\text{Eq. C13})$$

where RSS_{reduced} and RSS_{full} are the residual sums of squares from the reduced and full models, p_{reduced} and p_{full} are their respective numbers of estimated model parameters, and N is the number of analysed observations. For categorical depth, the residual sums of squares were calculated jointly across the four depth-indicator columns. The resulting permuted F statistic was then compared with the observed F statistic.

For depth, the four categorical depth indicators formed a multivariate outcome. The complete residual row was therefore permuted jointly so that the relationship among the four depth indicators within each observation was retained. Breadth amount remained in its original numerical form, whereas volume was analysed as $\log_{10}(\text{volume})$, consistent with the outcome representations specified in Section C.2.

The interaction effect was expressed as ΔR^2 , using the definition provided in Section C.2, Eq. C5. For this analysis, the reduced model contained the main effects of conversation quartile and cue group, whereas the full model additionally contained the conversation quartile $\times$ cue-group interaction. Accordingly, ΔR^2 represented the additional variance explained by the interaction beyond the corresponding main-effects model.

The Monte Carlo permutation p-value was calculated as

$$p = \frac{\#(F_{\text{perm}} \geq F_{\text{obs}}) + 1}{B + 1} \quad (\text{Eq. C14})$$

where F_{obs} is the observed model-comparison statistic, F_{perm} is the corresponding statistic obtained from a permutation, $\#\{\cdot\}$ denotes the number of permutations satisfying the stated condition, and $B = 10,000$ is the number of permutations. The +1 correction was applied to the randomly sampled permutation distribution (Phipson

& Smyth, 2010). Holm correction was subsequently applied across the three interaction tests according to the correction family specified in Section C.3. The number of permutation iterations and the outcome-specific random seeds are reported in Section C.6.

C.5.3 Conversation-Level Clustered Analysis

The supplementary conversation-level clustered analysis re-examined the central RQ3 associations while accounting for dependence between turn pairs originating from the same conversation. All 2,639 analytical turn pairs nested within 69 conversations were retained.

Each conversation identifier was converted to an integer cluster code, and all observations originating from the same conversation shared that code. Separate ordinary least-squares models were estimated for ordered depth, breadth amount, and log1p-transformed volume using the outcome representations specified in Section C.2.

Conversation-position analysis. Conversation quartile was the focal predictor. Effect size was expressed as ΔR^2 , using the definition provided in Section C.2, Eq. C5. The reduced model contained only the intercept, whereas the full model additionally contained conversation quartile. Thus, ΔR^2 represented the additional variance explained by conversation position relative to the intercept-only model.

Quartile $\times$ cue-group interaction analysis. The reduced model contained the main effects of conversation quartile and cue group, whereas the full model additionally contained their interaction. Using the same ΔR^2 definition from Section C.2, Eq. C5, the effect size represented the additional variance explained by the interaction beyond these main effects.

Implementation of clustered inference. This was implemented in statsmodels using the cluster covariance option. A finite-sample covariance correction and a degrees-of-freedom correction for the number of conversation clusters were applied, and coefficient-level inference was based on the resulting t distribution rather than on the normal distribution. The corresponding implementation settings were `cov_type="cluster"`, `use_correction=True`, `df_correction=True`, and `use_t=True`.

The cluster-robust covariance estimator therefore used the finite-sample correction factor

$$\frac{G}{G-1} \times \frac{N-1}{N-k} \quad (\text{Eq. C15})$$

where G is the number of conversation clusters, N is the number of observations, and k is the number of estimated model parameters. With 69 conversation clusters, the inferential degrees of freedom were based on

$$G - 1 = 68. \quad (\text{Eq. C16})$$

Conversation quartile and the conversation quartile $\times$ cue-group interaction are multi-parameter terms. They were therefore evaluated using omnibus Wald tests, which jointly tested whether the coefficients associated with each multi-parameter term differed from zero. The tests used `use_f=True`, producing F tests with denominator degrees of freedom equal to 68 (Pustejovsky & Tipton, 2018). Holm correction was applied according to the correction families specified in Section C.3.

The clustered depth specification used ordered depth rather than the categorical depth indicators used in the primary interaction analysis. The ordered-depth specification treats successive depth categories as numerically equally spaced. However, such spacing is not established by the coding scheme. The clustered depth analysis was therefore interpreted as a supplementary sensitivity specification rather than as an identical replication of the categorical-depth interaction analysis.

C.6 Resampling and Reproducibility Settings

Table C2 summarises the resampling settings used across the permutation and bootstrap analyses. All reported resampling procedures used 10,000 iterations. Fixed random seeds were used to ensure computational reproducibility. A base seed of 42 was selected for the analyses. For the three separate Freedman-Lane interaction tests, the implementation used this base seed for depth and incremented it by one and two for breadth amount and volume, resulting in seeds 42, 43, and 44, respectively. This was an implementation convention for assigning distinct reproducible random-number sequences to the three outcome-specific tests rather than a statistical requirement. The specific seed values themselves have no substantive statistical meaning.

Table C2
Resampling Settings Used in the Statistical Analyses

Procedure	Random seed
RQ1 permutation χ^2 tests	42
RQ1 adjusted analysis: conversation-level bootstrap	42
RQ2 adjusted analysis: conversation-level bootstrap	42
RQ3 primary conversation-position, conversation-length sensitivity, and cue-context permutation analyses	42
RQ3 Freedman-Lane interaction: depth	42
RQ3 Freedman-Lane interaction: breadth amount	43
RQ3 Freedman-Lane interaction: volume	44

The permutation procedures used the +1 Monte Carlo correction described in Sections C.5.1 and C.5.2 (Phipson & Smyth, 2010). The conversation-level bootstrap

procedures instead used the empirical two-sided coefficient-distribution procedure described in Section C.4.3 and did not apply a +1 correction.

C.7 Computational Environment and Reporting Conventions

The final computational workflow was executed in Python through JupyterLab. Table C3 reports the principal software and Python packages used during data preparation, statistical analysis, and generation of analytical outputs. Version information is reported to document the computational environment in which the final analyses were conducted.

Table C3

Principal Software and Python Packages Used in the Computational Analysis

Software or package	Version	Main use
Python	3.12.3	Execution of the computational workflow
JupyterLab	4.5.6	Execution and inspection of Python analyses
NumPy	2.4.3	Numerical operations
pandas	3.0.1	Data preparation, transformation, and tabulation
SciPy	1.18.1	Statistical testing and numerical procedures
statsmodels	0.15.0	Statistical modelling and inferential analyses
Matplotlib	3.10.8	Generation of statistical figures
langdetect	1.0.9	Automated language identification

Note. Only principal software and packages directly relevant to the computational workflow are reported. Supporting dependencies installed in the computational environment are not reported.

The principal rank-based and permutation procedures underlying the reported RQ1, RQ2, and RQ3 analyses were implemented directly in the analytical code rather than relying on default SciPy implementations. This allowed the rank handling, tie corrections, permutation procedures, and effect-size calculations specified in Sections C.2–C.5 to be implemented consistently with the documented analytical procedures. Regression modelling and conversation-clustered inference were implemented using statsmodels as specified in Sections C.4.2 and C.5.3, respectively.

Intermediate outputs, analytical logic, formulas, model specifications, and sample counts were checked before the corresponding results were incorporated into the thesis. Statistical figures were generated from the analytical outputs using Matplotlib, while statistical tables were subsequently formatted in Microsoft Word using the numerical results generated in Python.

Results were reported using the statistics applicable to each analytical procedure. Depending on the analysis, these included the analysed sample or group size, test statistic, degrees of freedom where applicable, raw and/or Holm-adjusted p-values,

and the corresponding effect-size estimate. Adjusted regression analyses additionally reported unstandardised coefficients (B) and unadjusted 95% percentile bootstrap confidence intervals.

Main-text tables containing only statistically supported follow-up comparisons report Holm-adjusted p-values, whereas Appendix D provides the corresponding complete results. Statistics that were not applicable to a particular procedure were omitted. Unless otherwise indicated, p-values and effect-size estimates were reported to three decimal places, with p-values below .001 reported as $p < .001$.

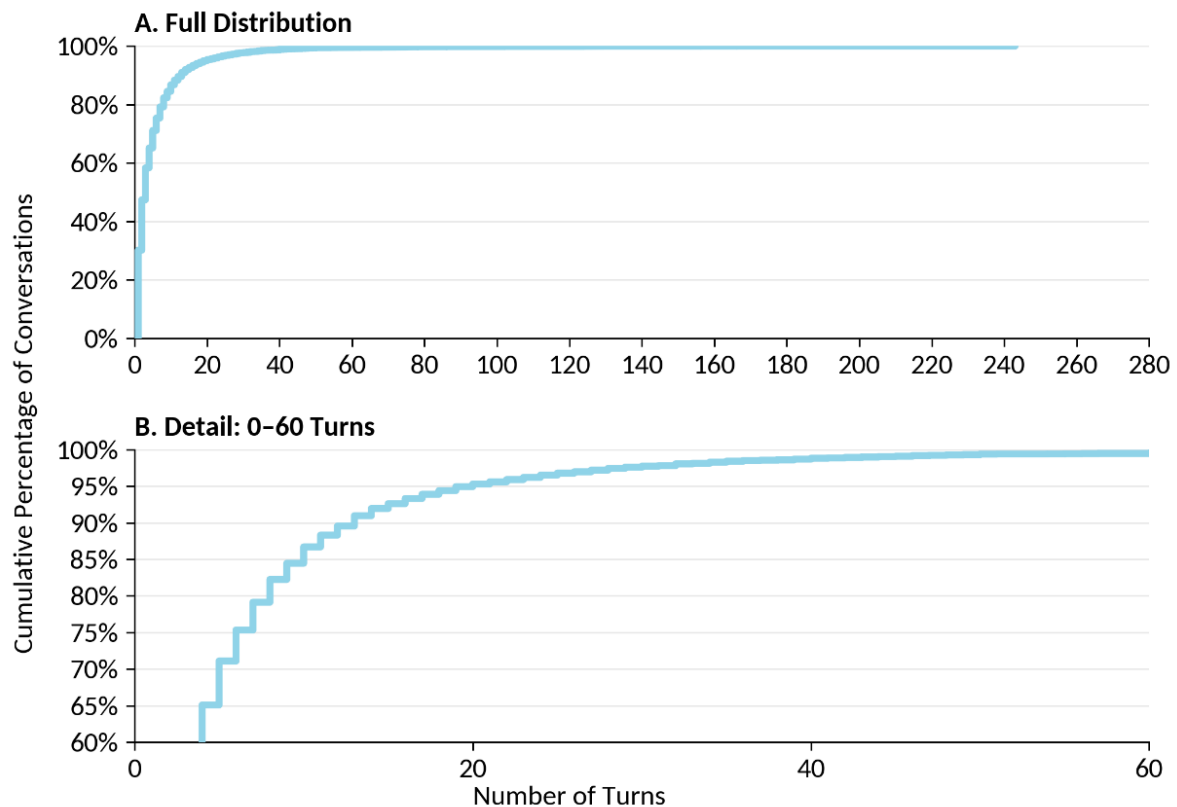
Appendix D:

Complete Results Tables and Figures

This appendix provides supplementary descriptive and statistical results supporting Chapter 4. It contains detailed distributions, complete pairwise comparisons, sensitivity analyses, and supplementary type- and cue-context analyses that are referenced in the main text but not reported there in full. The appendix provides transparency regarding the analytical results while allowing Chapter 4 to remain focused on the findings required to answer the research questions. Interpretation of the findings is therefore reserved for the main Results and Discussion chapters.

D.1 Cumulative Distribution of Conversation Lengths in the Sampling Frame

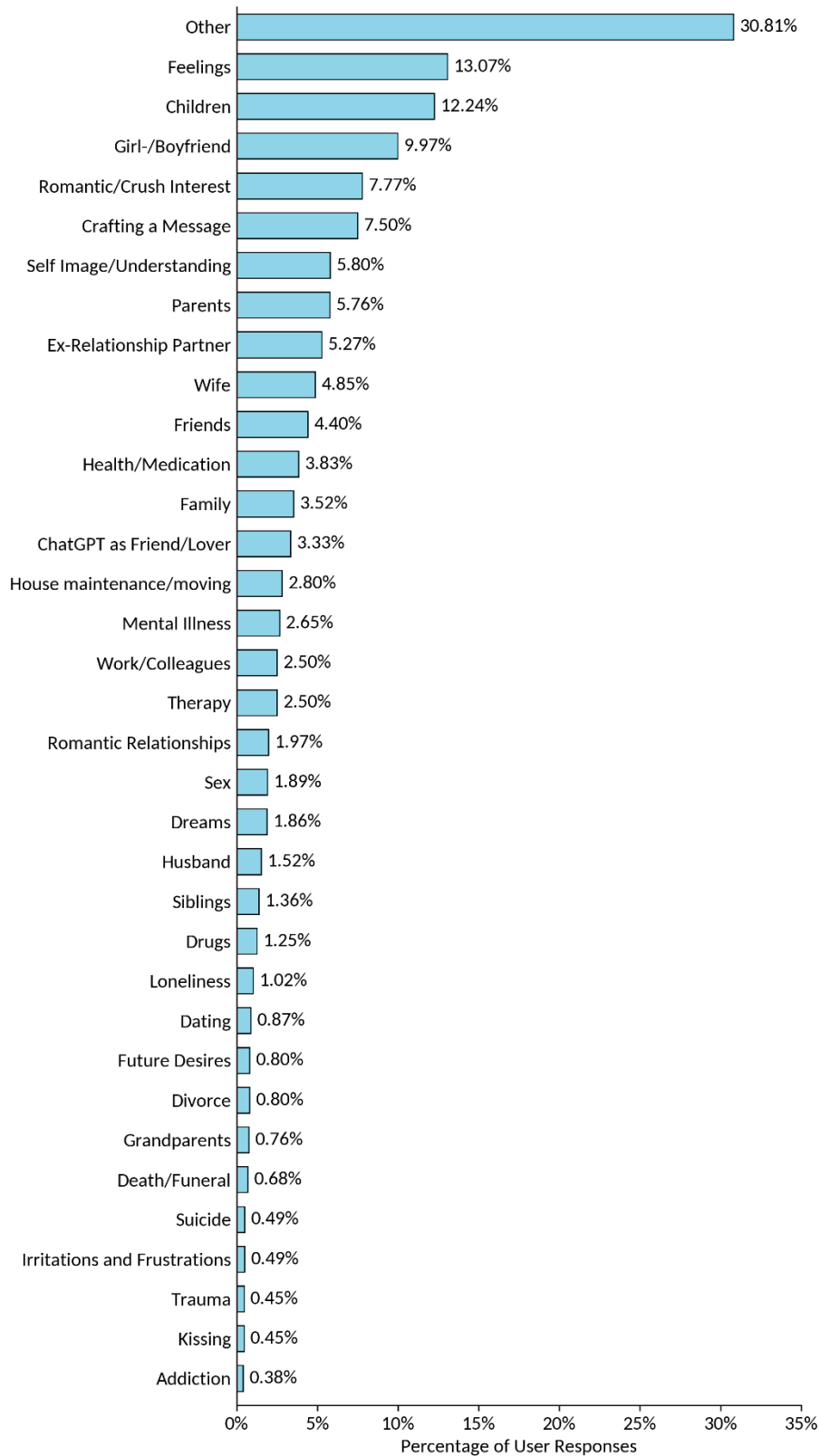
Figure D1 supplements the conversation-length distribution reported in Section 3.4.1 by presenting the cumulative distribution across all 4,978 English-language mental-health conversations in the sampling frame. Conversation length, defined as the number of original user-ChatGPT turns within a conversation, ranged up to 272 turns. The figure provides a more detailed view of the concentration of shorter conversations. It also shows that progressively fewer conversations extend to greater lengths than is visible from the six conversation-length strata used in the Methodology chapter.

Figure D1*Cumulative Distribution of Conversation Lengths in the Sampling Frame*

Note. Panel A presents the full cumulative distribution of conversation length. Panel B provides an enlarged view of the 0–60-turn range, with the y-axis restricted to 60%–100% to make variation within this part of the distribution more visible.

D.2 Breadth Type Frequencies

Figure D2 presents the complete distribution of coded breadth types across the analysed user responses. Section 4.1.1 displays only breadth types occurring in at least 5% of user responses to maintain readability, whereas Figure D2 includes all coded breadth-type categories. Because multiple breadth types could be assigned within the same response, the percentages represent the proportion of user responses in which each breadth type was identified and therefore do not sum to 100%.

Figure D2*Complete Distribution of Coded Breadth Types Across User Responses*

Note. Percentages represent the proportion of analysed user responses in which each breadth type was identified. *Other* represents personal content areas that did not fit any predefined breadth-type category.

D.3 Pairwise Comparisons Across Depth Levels

Section 4.1.4 reports the overall and pairwise differences in breadth amount and volume across the four depth levels. Table D1 provides the complete results for all six pairwise comparisons between Superficial, Middle, Inner, and Core. Holm correction was applied separately across the six comparisons within each outcome.

Table D1

Pairwise Comparisons of Breadth Amount and Volume Across Depth Levels

Depth-level comparison	<i>U</i>	Holm-adjusted <i>p</i>	Rank-biserial <i>r</i>
Panel A. Breadth Amount			
Superficial–Middle	282,660.0	< .001	–.207
Superficial–Inner	160,575.0	< .001	–.494
Superficial–Core	43,629.5	< .001	–.632
Middle–Inner	185,141.0	< .001	–.302
Middle–Core	52,393.5	< .001	–.472
Inner–Core	69,920.5	< .001	–.208
Panel B. Volume			
Superficial–Middle	257,414.5	< .001	–.277
Superficial–Inner	125,657.5	< .001	–.604
Superficial–Core	17,134.0	< .001	–.856
Middle–Inner	166,475.0	< .001	–.372
Middle–Core	27,137.5	< .001	–.726
Inner–Core	47,512.5	< .001	–.462

Note. Holm correction was applied separately across the six pairwise comparisons within each outcome. Rank-biserial *r* is reported as the effect-size measure. Negative values indicate that the second-listed, higher depth level had greater breadth amount or volume than the first-listed, lower depth level.

D.4 Pairwise Comparisons Between Recommendation Groups

Section 4.2.2 reports the recommendation-group comparisons that remained statistically supported after Holm correction. Table D2 provides the complete set of pairwise comparisons between the three mutually exclusive RQ1 groups: No recommendation, Recommendation without observable uptake, and Recommendation with observable uptake. Results are reported for depth distribution, breadth amount, and volume, including comparisons that were not statistically supported.

Table D2*Complete Pairwise Comparisons Between Recommendation Groups*

Group comparison	Test statistic	Holm-adjusted p	Effect size
Panel A. Depth distribution			
Recommendation without observable uptake – Recommendation with observable uptake	$\chi^2(3) = 23.793$	$< .001$	Cramér's $V = .103$
Recommendation without observable uptake – No recommendation	$\chi^2(3) = 48.554$	$< .001$	Cramér's $V = .167$
Recommendation with observable uptake – No recommendation	$\chi^2(3) = 20.072$	$< .001$	Cramér's $V = .126$
Panel B. Breadth amount			
Recommendation without observable uptake – Recommendation with observable uptake	$U = 679,467.50$	$< .001$	$r = .112$
Recommendation without observable uptake – No recommendation	$U = 280,879.50$	$.004$	$r = .088$
Recommendation with observable uptake – No recommendation	$U = 165,372.50$	$.408$	$r = -.022$
Panel C. Volume			
Recommendation without observable uptake – Recommendation with observable uptake	$U = 650,716.50$	$.009$	$r = .065$
Recommendation without observable uptake – No recommendation	$U = 274,651.00$	$.062$	$r = .064$
Recommendation with observable uptake – No recommendation	$U = 168,823.00$	$.947$	$r = -.002$

Note. Holm correction was applied separately across the three pairwise comparisons within each outcome. Depth distributions were compared using permutation chi-square tests with Cramér's V as the effect-size measure. Breadth amount and volume were compared using Mann-Whitney U tests with effect-size r .

D.5 Supplementary Recommendation-Type Analysis

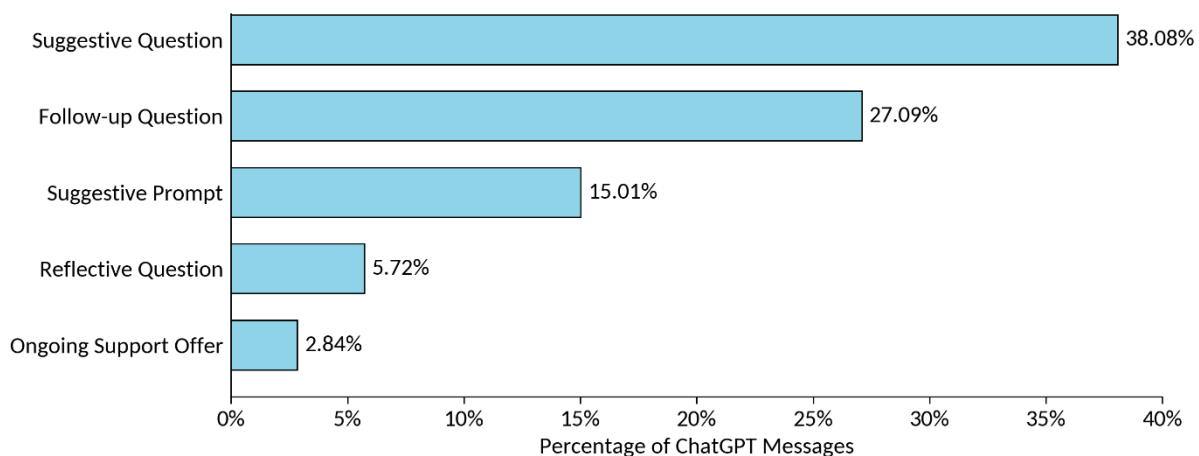
This section provides the complete descriptive recommendation-type results supporting Section 4.2.4. The five coded recommendation types were examined to describe variation in observable uptake and subsequent self-disclosure across recommendation types. Because recommendation types could co-occur within the same ChatGPT message and differed substantially in frequency, these results were treated as descriptive rather than as independent inferential comparisons.

Across all 2,639 analysed ChatGPT messages, the mean number of coded recommendations was 0.98 per message, while the mean number of recommendations with observable uptake was 0.35. Among the 2,261 messages containing at least one recommendation, the corresponding means were 1.14 recommendations and 0.41 recommendations with observable uptake per message.

Figure D3 presents the prevalence of the five coded recommendation types across the analytical dataset. Suggestive Questions occurred most frequently, followed by Follow-up Questions and Suggestive Prompts, whereas Reflective Questions and Ongoing Support Offers occurred less frequently.

Figure D3

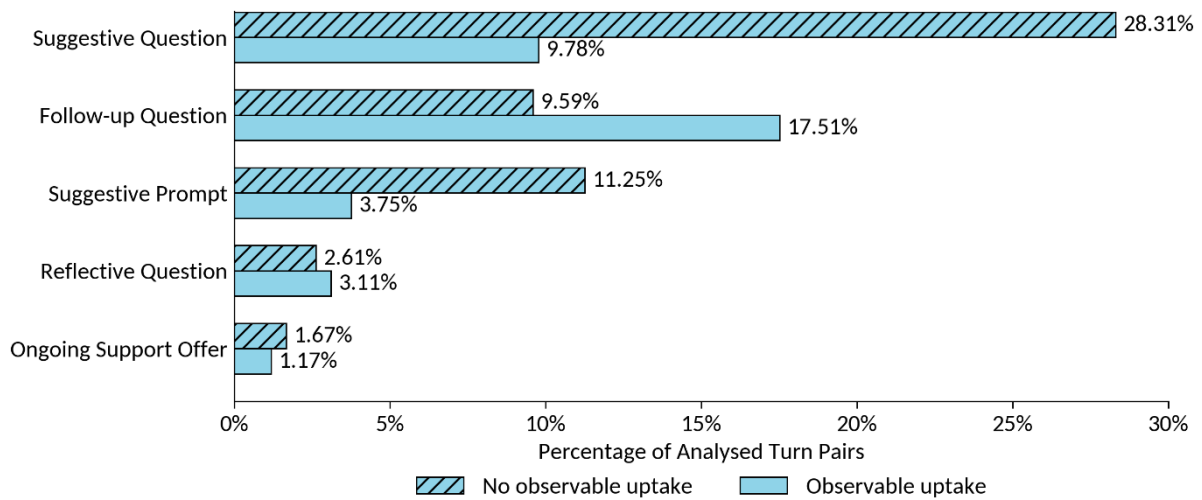
Prevalence of Recommendation Types Across ChatGPT Messages



Note. Percentages represent the proportion of analysed ChatGPT messages containing each recommendation type at least once. Multiple recommendation types could be coded within the same message; percentages therefore do not sum to 100%.

Observable Uptake by Recommendation Type

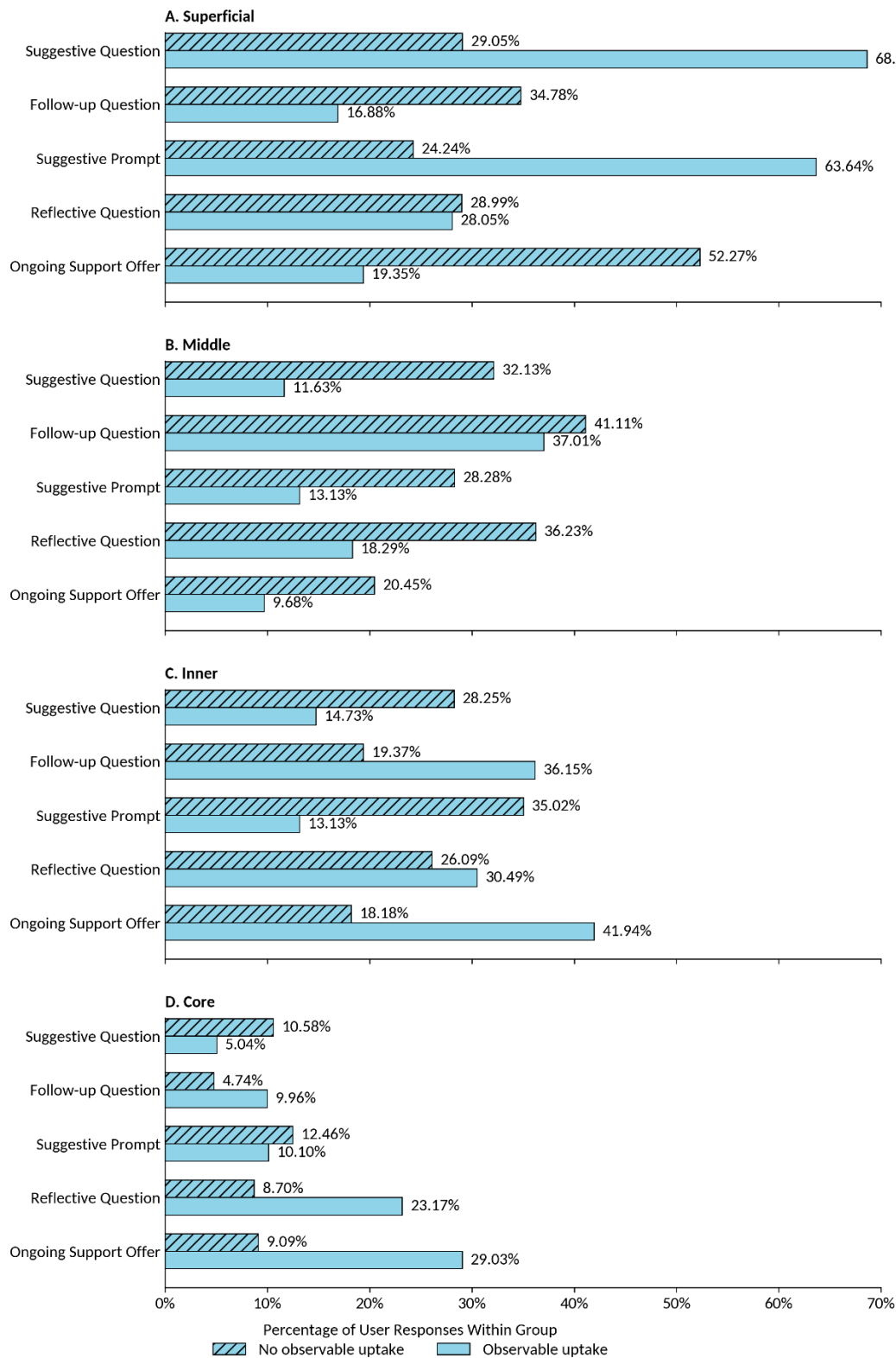
Figure D4 presents observable uptake across the five recommendation types. Suggestive Questions and Suggestive Prompts occurred more frequently without observable uptake than with observable uptake. Follow-up Questions showed the opposite descriptive pattern and occurred more frequently with observable uptake. Reflective Questions showed a smaller difference in favour of observable uptake, whereas Ongoing Support Offers showed a more balanced distribution.

Figure D4*Frequency of Recommendation Types Across ChatGPT Messages*

Note. Percentages represent the proportion of all analysed turn pairs belonging to each recommendation-type × observable-uptake combination. Recommendation types could co-occur within the same ChatGPT message; a turn pair could therefore contribute to more than one recommendation type. Observable uptake refers only to uptake identified in the immediately following user response.

Depth Distribution by Recommendation Type and Observable Uptake

Figure D5 presents the complete four-category depth distributions within each recommendation-type × observable-uptake combination. For Suggestive Questions and Suggestive Prompts, responses with observable uptake contained a larger proportion of Superficial responses than those without observable uptake. Follow-up Questions showed a different descriptive pattern, with a larger proportion of Inner responses when observable uptake was identified. Reflective Questions and Ongoing Support Offers also showed larger proportions of Inner and/or Core responses when observable uptake was identified.

Figure D5*Depth Distribution by Recommendation Type and Observable Uptake*

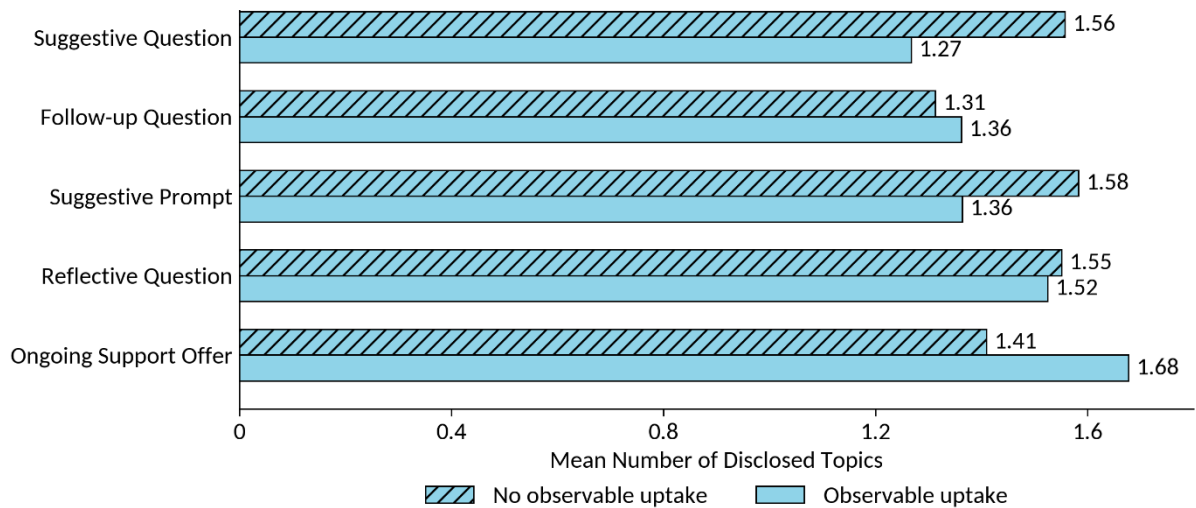
Note. Percentages were calculated separately within each recommendation-type × observable-uptake combination and sum to approximately 100% across the four depth levels due to rounding. Recommendation types could co-occur within the same ChatGPT message. Observable uptake refers only to uptake identified in the immediately following user response.

Breadth Amount by Recommendation Type and Observable Uptake

Figure D6 presents mean breadth amount within each recommendation-type × observable-uptake combination. Mean breadth amount was higher with observable uptake for Follow-up Questions and Ongoing Support Offers. For Suggestive Questions, Suggestive Prompts, and Reflective Questions, mean breadth amount was equal to or lower with observable uptake than without observable uptake. The No recommendation group is included as a descriptive reference.

Figure D6

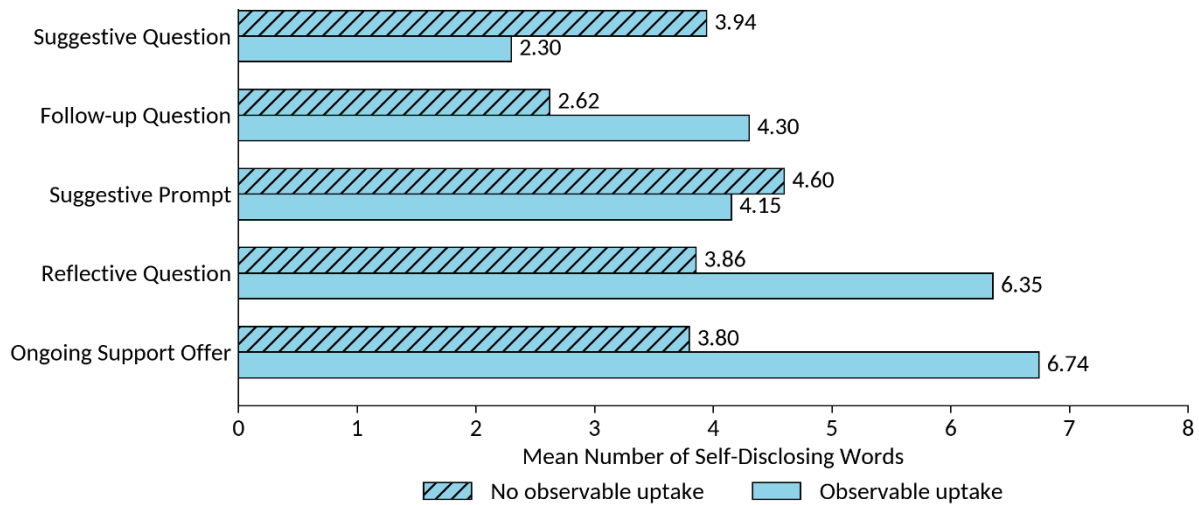
Mean Breadth Amount by Recommendation Type and Observable Uptake



Note. Bars show mean breadth amount, measured in personal content areas. The No recommendation group is included as a descriptive reference. Recommendation types could co-occur within the same ChatGPT message.

Volume by Recommendation Type and Observable Uptake

Figure D7 presents mean volume within each recommendation-type × observable-uptake combination. Mean volume was higher with observable uptake for Follow-up Questions, Reflective Questions, and Ongoing Support Offers. For Suggestive Questions and Suggestive Prompts, mean volume was lower with observable uptake than without observable uptake.

Figure D7*Mean Volume by Recommendation Type and Observable Uptake*

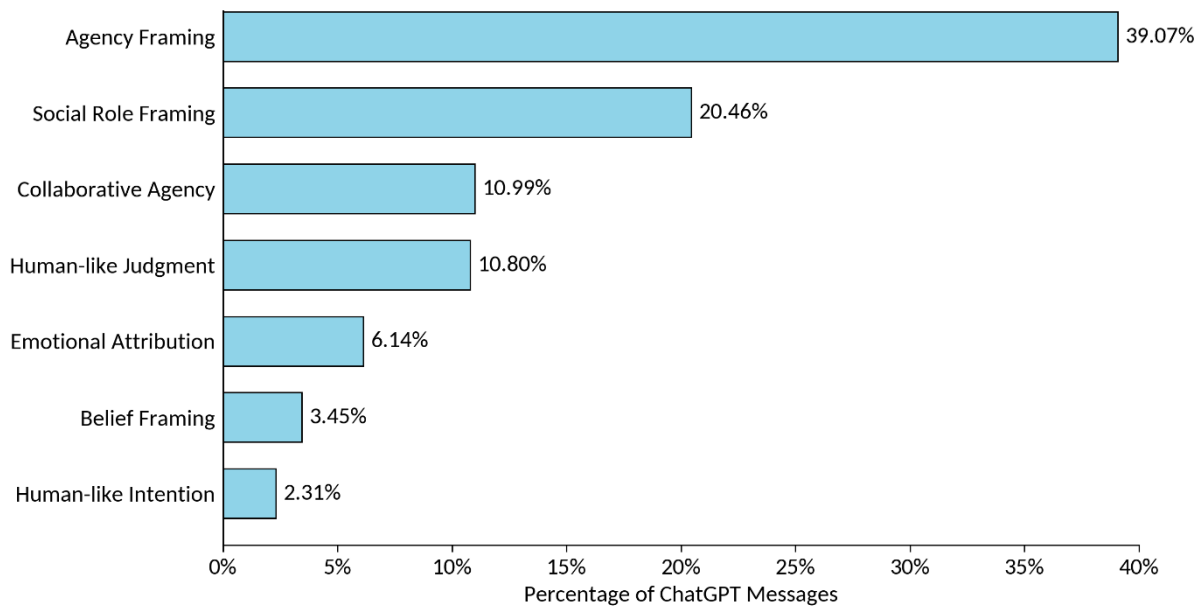
Note. Bars show mean volume, measured in self-disclosing words in the subsequent user response. Recommendation types could co-occur within the same ChatGPT message. Observable uptake refers only to uptake identified in the immediately following user response.

D.6 Descriptive Textual Anthropomorphic-Cue Analysis

This section provides the supplementary descriptive results for textual anthropomorphic cues supporting Sections 4.1.3 and 4.3.3. Beyond the overall cue-presence results reported in Section 4.1.3, this section reports the number of coded cue instances per ChatGPT message and the frequency of the seven textual anthropomorphic-cue types. The type-level analysis subsequently describes depth distribution, breadth amount, and volume for responses following each cue type. Because multiple cue types could occur within the same ChatGPT message and differed substantially in frequency, these results were treated as descriptive rather than as independent inferential comparisons.

Across all 2,639 analysed ChatGPT messages, the mean number of coded textual anthropomorphic-cue instances was 1.10 per message. The median was one cue instance, with observed values ranging from 0 to 10. In total, 70.94% of messages contained zero or one coded cue instance, while 3.03% contained four or more.

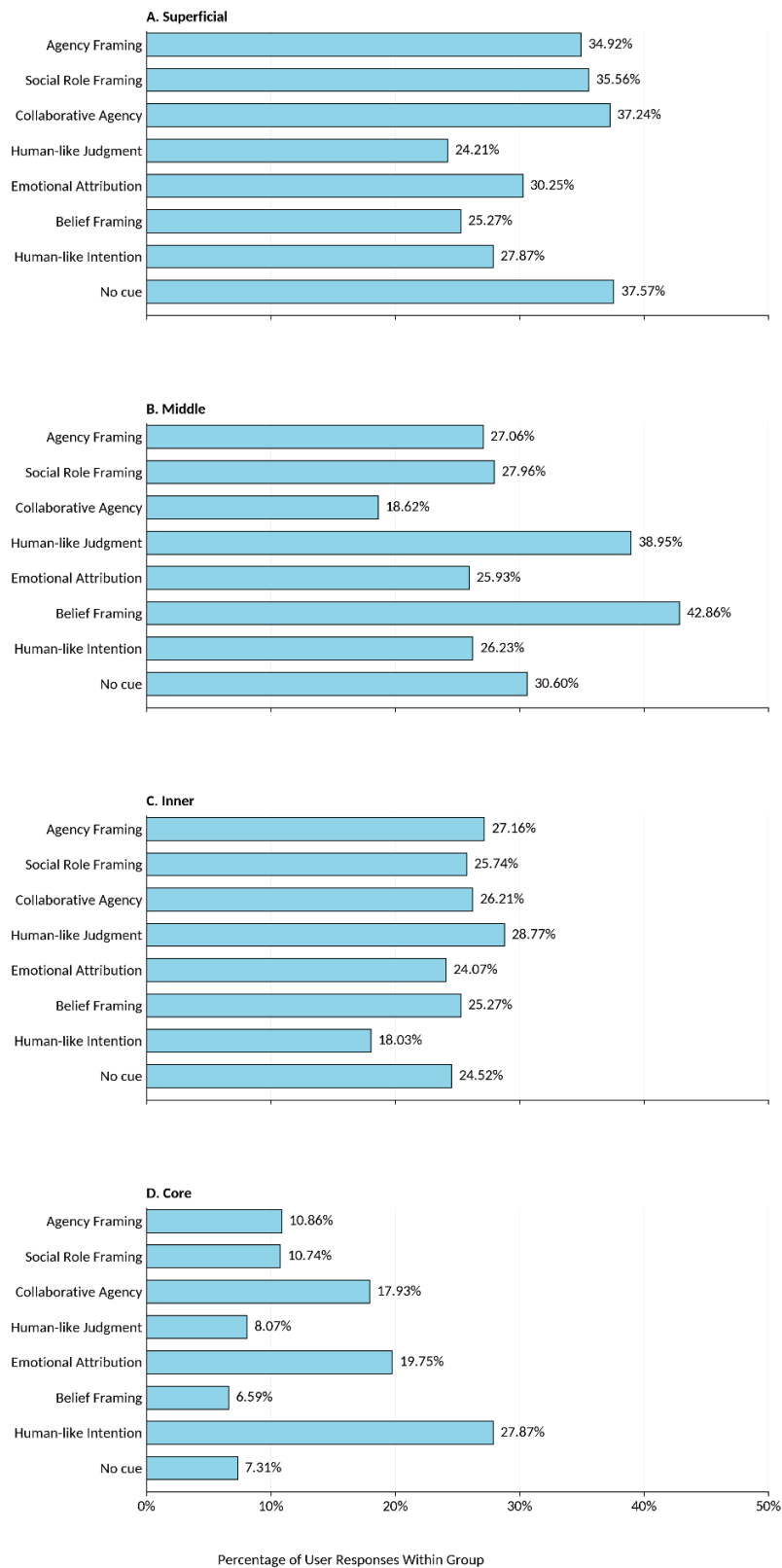
Figure D8 presents the frequency of the seven coded textual anthropomorphic-cue types. Agency Framing occurred most frequently, followed by Social Role Framing. Collaborative Agency and Human-like Judgment occurred at similar frequencies, while Emotional Attribution, Belief Framing, and Human-like Intention occurred less frequently.

Figure D8*Frequency of Textual Anthropomorphic-Cue Types Across ChatGPT Messages*

Note. Percentages represent the proportion of analysed ChatGPT messages containing each textual anthropomorphic-cue type at least once. Multiple cue types could occur within the same message; percentages therefore do not sum to 100%.

Depth Distribution by Textual Anthropomorphic-Cue Type

Figure D9 presents the complete four-category depth distribution for responses following each textual anthropomorphic-cue type. Responses following messages containing no textual anthropomorphic cue are included as a descriptive reference. Agency Framing and Social Role Framing showed broadly similar depth distributions. Human-like Judgment and Belief Framing contained larger proportions of responses classified as Middle, whereas Collaborative Agency, Emotional Attribution, and Human-like Intention contained larger proportions classified as Core. Human-like Intention had the largest Core proportion and was identified in 61 ChatGPT messages.

Figure D9*Depth Distribution by Textual Anthropomorphic-Cue Type*

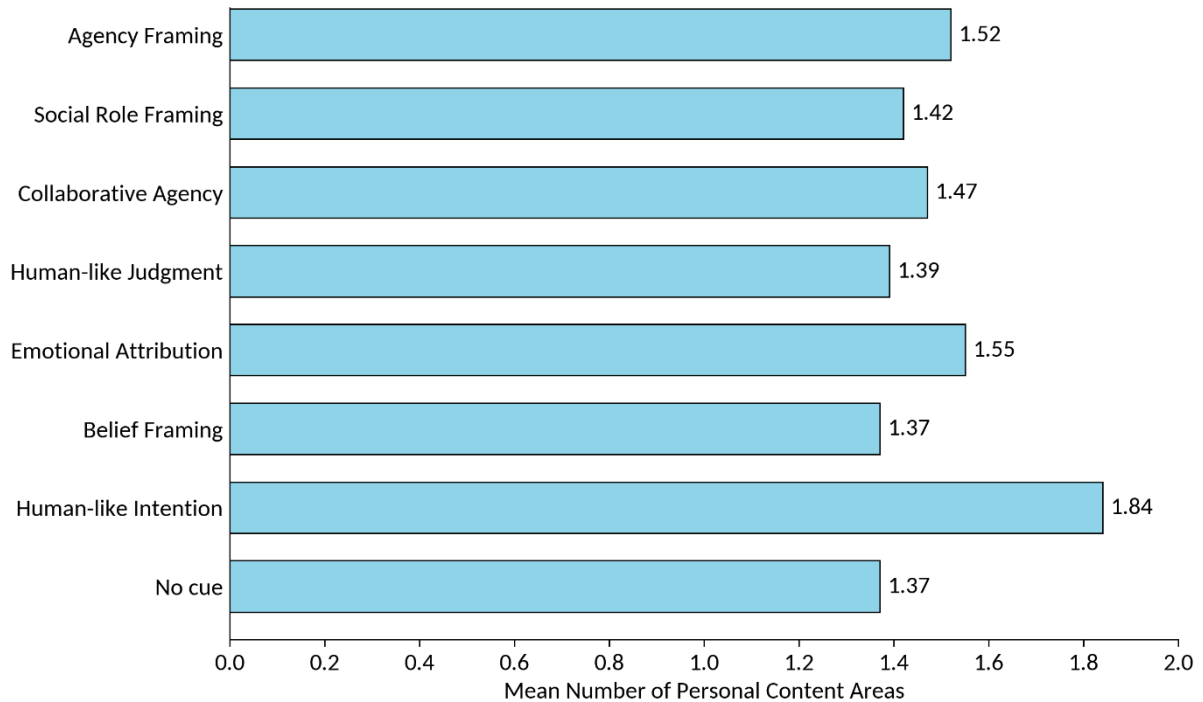
Note. Percentages were calculated separately within each textual anthropomorphic-cue type and within the No cue reference group; within each group, percentages sum to approximately 100% across the four depth levels due to rounding. Cue types could co-occur within the same ChatGPT message.

Breadth Amount by Textual Anthropomorphic-Cue Type

Figure D10 presents mean breadth amount for responses following each textual anthropomorphic-cue type. Human-like Intention showed the highest mean breadth amount followed by Emotional Attribution. Agency Framing and Collaborative Agency also showed higher mean breadth amount than the No cue reference group, whereas Social Role Framing, Human-like Judgment, and Belief Framing were closer to the No cue reference value.

Figure D10

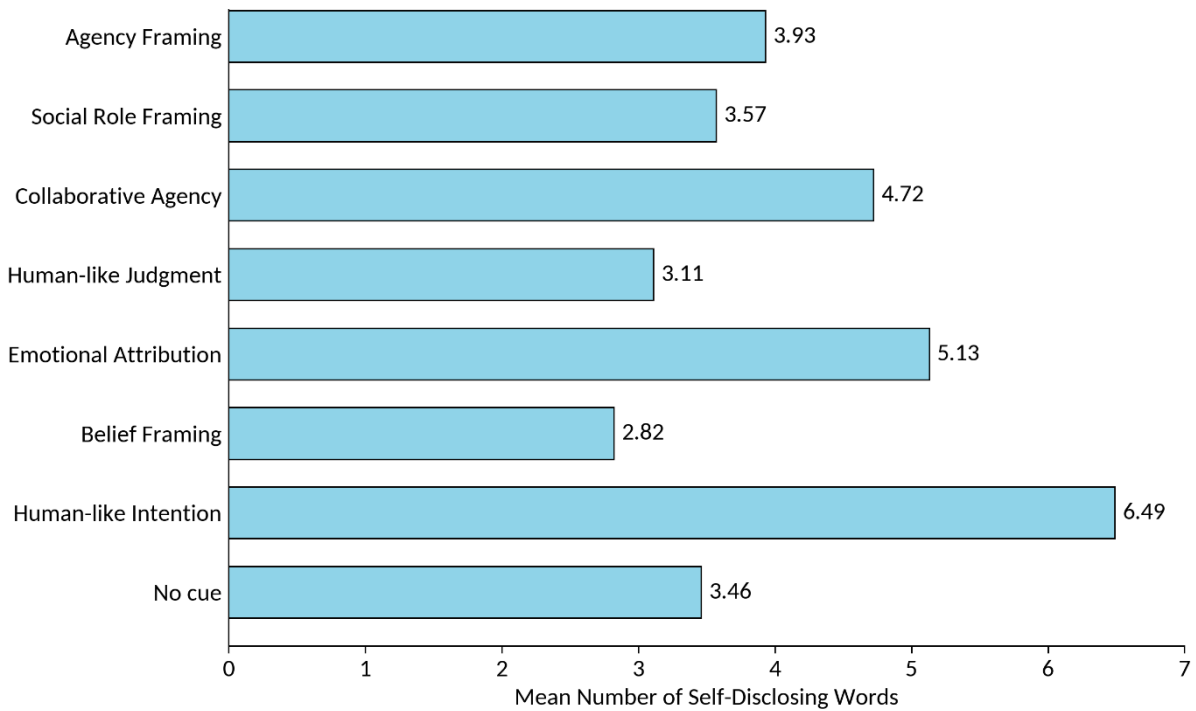
Mean Breadth Amount by Textual Anthropomorphic-Cue Type



Note. Bars show mean breadth amount, measured in personal content areas in the subsequent user response. No cue is included as a descriptive reference group. Cue types could co-occur within the same ChatGPT message.

Volume by Textual Anthropomorphic-Cue Type

Figure D11 presents mean volume for responses following each textual anthropomorphic-cue type. Human-like Intention showed the highest mean volume, followed by Emotional Attribution and Collaborative Agency. Agency Framing and Social Role Framing were closer to the No cue reference value, whereas Human-like Judgment and Belief Framing showed lower mean volume.

Figure D11*Mean Volume by Textual Anthropomorphic-Cue Type*

Note. Bars show mean volume, measured in self-disclosing words in the subsequent user response. No cue is included as a descriptive reference group. Cue types could co-occur within the same ChatGPT message.

D.7 Pairwise Comparisons Between Conversation Quartiles

Section 4.4.1 reports the pairwise conversation-quartile comparisons that remained statistically supported after Holm correction. Table D3 provides the complete set of pairwise comparisons across Quartiles 1–4 for depth distribution, breadth amount, and volume, including comparisons that were not statistically supported.

Table D3
Complete Pairwise Comparisons Between Conversation Quartiles

Quartile comparison	Test statistic	Raw <i>p</i>	Holm-adjusted <i>p</i>	Effect size
Panel A. Depth distribution				
Q1–Q2	$\chi^2 = 8.6850$	.0324	.0972	$V = .0804$
Q1–Q3	$\chi^2 = 27.2929$	.0001	.0006	$V = .1437$
Q1–Q4	$\chi^2 = 25.7838$	.0001	.0006	$V = .1380$
Q2–Q3	$\chi^2 = 8.1074$	.0417	.0972	$V = .0794$
Q2–Q4	$\chi^2 = 10.1627$	.0143	.0572	$V = .0878$
Q3–Q4	$\chi^2 = 3.7132$	.3013	.3013	$V = .0535$
Panel B. Breadth amount				
Q1–Q2	$U = 241,014.00$	.0069	.0414	$r = .0698$
Q1–Q3	$U = 232,434.50$	.0105	.0525	$r = .0660$
Q1–Q4	$U = 240,366.00$	.0606	.2424	$r = .0493$
Q2–Q3	$U = 205,256.00$	.8311	1.0000	$r = -.0053$
Q2–Q4	$U = 212,047.50$	.3911	1.0000	$r = -.0219$
Q3–Q4	$U = 206,304.50$	.5191	1.0000	$r = -.0168$
Panel C. Volume				
Q1–Q2	$U = 242,376.00$	.0070	.0280	$r = .0759$
Q1–Q3	$U = 246,619.00$	.0001	.0006	$r = .1311$
Q1–Q4	$U = 250,578.00$	.0008	.0040	$r = .0938$
Q2–Q3	$U = 217,856.50$	.0413	.1239	$r = .0558$
Q2–Q4	$U = 220,734.00$	.5048	.5048	$r = .0182$
Q3–Q4	$U = 202,107.00$	.1769	.3538	$r = -.0368$

Note. Holm correction was applied separately across the six pairwise conversation-quartile comparisons within each outcome. Depth distributions were compared using permutation chi-square tests with Cramér's V as the effect-size measure. Breadth amount and volume were compared using permutation Mann-Whitney U tests with rank-biserial r as the effect-size measure. For rank-biserial r , positive values indicate higher ranks in the first-listed quartile, whereas negative values indicate higher ranks in the second-listed quartile.

D.8 Conversation-Length Sensitivity Analysis

This section provides the complete results of the conversation-length sensitivity analysis reported in Section 4.4.2. The primary conversation-position analysis was repeated within six conversation-length strata: 2–10, 11–20, 21–30, 31–40, 41–50, and 51+ turns. Table D4, Table D5, Table D6 and Table D7 report the stratum-specific descriptive results, overall quartile tests, and applicable pairwise comparisons.

Table D4 presents the stratum-specific sample sizes and descriptive statistics for breadth amount and volume across conversation quartiles.

Table D4

Breadth Amount and Volume by Conversation-Length Stratum and Conversation Quartile

Conversation-length stratum	Quartile	Analysed turn pairs (<i>n</i>)	Contributing conversations (<i>n</i>)	Breadth amount mean	Breadth amount median	Volume mean	Volume median
2–10	Q1	8	7	1.88	1.0	3.62	0.0
2–10	Q2	4	4	2.00	1.5	6.75	2.0
2–10	Q3	2	2	1.00	1.0	0.00	0.0
2–10	Q4	5	5	1.00	1.0	2.00	0.0
11–20	Q1	55	16	1.78	2.0	9.35	0.0
11–20	Q2	46	16	1.80	1.5	9.70	0.0
11–20	Q3	43	16	1.70	1.0	3.58	0.0
11–20	Q4	47	16	1.72	1.0	6.28	1.0
21–30	Q1	67	11	1.75	1.0	6.60	1.0
21–30	Q2	63	11	1.56	1.0	5.03	0.0
21–30	Q3	60	11	1.52	1.0	4.40	0.0
21–30	Q4	64	11	1.41	1.0	2.94	0.0
31–40	Q1	121	14	1.35	1.0	3.26	0.0
31–40	Q2	113	14	1.43	1.0	2.67	0.0

Conversation-length stratum	Quartile	Analysed turn pairs (<i>n</i>)	Contributing conversations (<i>n</i>)	Breadth amount mean	Breadth amount median	Volume mean	Volume median
31–40	Q3	108	14	1.44	1.0	3.48	0.0
31–40	Q4	116	14	1.52	1.0	3.14	0.0
41–50	Q1	77	7	1.64	1.0	4.83	2.0
41–50	Q2	73	7	1.42	1.0	3.60	2.0
41–50	Q3	71	7	1.39	1.0	3.55	0.0
41–50	Q4	75	7	1.39	1.0	3.60	0.0
51+	Q1	362	14	1.46	1.0	3.48	0.0
51+	Q2	354	14	1.32	1.0	3.49	0.0
51+	Q3	348	14	1.31	1.0	2.87	0.0
51+	Q4	357	14	1.37	1.0	3.11	0.0

Note. Conversation counts indicate the number of conversations contributing at least one analysed turn pair to the respective conversation quartile. Breadth amount is measured in personal content areas; volume is measured in self-disclosing words.

Table D5 presents the corresponding depth distributions across conversation quartiles within each conversation-length stratum.

Table D5

Depth Distributions by Conversation-Length Stratum and Conversation Quartile

Conversation-length stratum	Quartile	Superficial, <i>n</i> (%)	Middle, <i>n</i> (%)	Inner, <i>n</i> (%)	Core, <i>n</i> (%)
2–10	Q1	3 (37.50%)	1 (12.50%)	2 (25.00%)	2 (25.00%)
2–10	Q2	0 (0.00%)	1 (25.00%)	1 (25.00%)	2 (50.00%)
2–10	Q3	2 (100.00%)	0 (0.00%)	0 (0.00%)	0 (0.00%)
2–10	Q4	3 (60.00%)	1 (20.00%)	1 (20.00%)	0 (0.00%)
11–20	Q1	14 (25.45%)	13 (23.64%)	15 (27.27%)	13 (23.64%)
11–20	Q2	6 (13.04%)	11 (23.91%)	19 (41.30%)	10 (21.74%)
11–20	Q3	10 (23.26%)	13 (30.23%)	12 (27.91%)	8 (18.60%)

Conversation-length stratum	Quartile	Superficial, n (%)	Middle, n (%)	Inner, n (%)	Core, n (%)
11–20	Q4	12 (25.53%)	9 (19.15%)	19 (40.43%)	7 (14.89%)
21–30	Q1	19 (28.36%)	16 (23.88%)	20 (29.85%)	12 (17.91%)
21–30	Q2	23 (36.51%)	11 (17.46%)	20 (31.75%)	9 (14.29%)
21–30	Q3	27 (45.00%)	10 (16.67%)	16 (26.67%)	7 (11.67%)
21–30	Q4	26 (40.62%)	10 (15.62%)	25 (39.06%)	3 (4.69%)
31–40	Q1	51 (42.15%)	39 (32.23%)	20 (16.53%)	11 (9.09%)
31–40	Q2	56 (49.56%)	25 (22.12%)	23 (20.35%)	9 (7.96%)
31–40	Q3	50 (46.30%)	26 (24.07%)	24 (22.22%)	8 (7.41%)
31–40	Q4	62 (53.45%)	21 (18.10%)	20 (17.24%)	13 (11.21%)
41–50	Q1	16 (20.78%)	13 (16.88%)	36 (46.75%)	12 (15.58%)
41–50	Q2	22 (30.14%)	13 (17.81%)	25 (34.25%)	13 (17.81%)
41–50	Q3	28 (39.44%)	13 (18.31%)	26 (36.62%)	4 (5.63%)
41–50	Q4	27 (36.00%)	26 (34.67%)	18 (24.00%)	4 (5.33%)
51+	Q1	97 (26.80%)	107 (29.56%)	119 (32.87%)	39 (10.77%)
51+	Q2	111 (31.36%)	140 (39.55%)	73 (20.62%)	30 (8.47%)
51+	Q3	120 (34.48%)	140 (40.23%)	71 (20.40%)	17 (4.89%)
51+	Q4	138 (38.66%)	113 (31.65%)	82 (22.97%)	24 (6.72%)

Note. Percentages were calculated within each conversation-length-stratum × conversation-quartile combination and sum to approximately 100% across the four depth levels within each row due to rounding.

Table D6 reports the overall conversation-quartile tests within each conversation-length stratum. Holm correction was applied across the three outcomes separately within each stratum.

Table D6*Overall Conversation-Quartile Tests Within Conversation-Length Strata*

Conversation-length stratum	Outcome	Contributing conversations (n)	Analysed turn pairs (n)	Test statistic	Raw p	Holm-adjusted p	Effect size
2–10	Depth distribution	7	19	$\chi^2 = 7.9760$	.6251	1.0000	Cramér's V = 0.3741
2–10	Breadth amount	7	19	H = 3.7917	.2789	.8366	$\epsilon^2 = 0.0528$
2–10	Volume	7	19	H = 1.4189	.7429	1.0000	$\epsilon^2 = -0.1054$
11–20	Depth distribution	16	191	$\chi^2 = 7.0138$	.6506	1.0000	Cramér's V = 0.1106
11–20	Breadth amount	16	191	H = 0.4315	.9349	1.0000	$\epsilon^2 = -0.0137$
11–20	Volume	16	191	H = 1.2288	.7521	1.0000	$\epsilon^2 = -0.0095$
21–30	Depth distribution	11	254	$\chi^2 = 10.6842$	.3049	.5977	Cramér's V = 0.1184
21–30	Breadth amount	11	254	H = 3.6896	.2989	.5977	$\epsilon^2 = 0.0028$
21–30	Volume	11	254	H = 7.8134	.0471	.1413	$\epsilon^2 = 0.0193$
31–40	Depth distribution	14	458	$\chi^2 = 9.2565$	.4139	.8277	Cramér's V = 0.0821
31–40	Breadth amount	14	458	H = 4.6471	.1979	.5936	$\epsilon^2 = 0.0036$
31–40	Volume	14	458	H = 0.1974	.9772	.9772	$\epsilon^2 = -0.0062$
41–50	Depth distribution	7	296	$\chi^2 = 26.2654$	.0019	.0057	Cramér's V = 0.1720
41–50	Breadth amount	7	296	H = 3.6795	.3024	.3024	$\epsilon^2 = 0.0023$
41–50	Volume	7	296	H = 8.4398	.0381	.0762	$\epsilon^2 = 0.0186$
51+	Depth distribution	14	1421	$\chi^2 = 41.2540$	.0001	.0003	Cramér's V = 0.0984
51+	Breadth amount	14	1421	H = 12.2307	.0048	.0048	$\epsilon^2 = 0.0065$
51+	Volume	14	1421	H = 19.3233	.0004	.0008	$\epsilon^2 = 0.0115$

Note. Depth distributions were compared using permutation chi-square tests with Cramér's V as the effect-size measure. Breadth amount and volume were compared using permutation Kruskal-Wallis tests with ϵ^2 as the effect-size measure. Holm correction was applied across the three outcomes separately within each conversation-length stratum. Negative ϵ^2 values are untruncated sample estimates and indicate effect-size estimates effectively at zero.

Pairwise conversation-quartile comparisons were conducted only for stratum $\times$ outcome combinations with a statistically supported overall quartile difference after Holm correction. Pairwise results are therefore reported for depth distribution in the 41–50-turn

stratum and for depth distribution, breadth amount, and volume in the 51+ stratum. Table D7 provides all six pairwise quartile comparisons within each of these combinations, including comparisons that were not statistically supported.

Table D7

Complete Pairwise Conversation-Quartile Comparisons for Supported Conversation-Length Stratum × Outcome Combinations

Conversation-length stratum	Outcome	Quartile comparison	Test statistic	Raw <i>p</i>	Holm-adjusted <i>p</i>	Effect size
41–50	Depth distribution	Q1–Q2	$\chi^2 = 2.8663$	.4107	.4107	Cramér's <i>V</i> = 0.1382
41–50	Depth distribution	Q1–Q3	$\chi^2 = 8.6566$	.0342	.1368	Cramér's <i>V</i> = 0.2418
41–50	Depth distribution	Q1–Q4	$\chi^2 = 17.1239$	.0006	.0036	Cramér's <i>V</i> = 0.3356
41–50	Depth distribution	Q2–Q3	$\chi^2 = 5.4776$	.1433	.3756	Cramér's <i>V</i> = 0.1950
41–50	Depth distribution	Q2–Q4	$\chi^2 = 10.7227$	.0136	.0680	Cramér's <i>V</i> = 0.2692
41–50	Depth distribution	Q3–Q4	$\chi^2 = 5.7008$	.1252	.3756	Cramér's <i>V</i> = 0.1976
51+	Depth distribution	Q1–Q2	$\chi^2 = 17.4588$	.0008	.0032	Cramér's <i>V</i> = 0.1562
51+	Depth distribution	Q1–Q3	$\chi^2 = 27.3504$	.0001	.0006	Cramér's <i>V</i> = 0.1963
51+	Depth distribution	Q1–Q4	$\chi^2 = 17.6653$	.0006	.0030	Cramér's <i>V</i> = 0.1567
51+	Depth distribution	Q2–Q3	$\chi^2 = 3.9232$	.2617	.2617	Cramér's <i>V</i> = 0.0748
51+	Depth distribution	Q2–Q4	$\chi^2 = 6.9858$	.0683	.2049	Cramér's <i>V</i> = 0.0991
51+	Depth distribution	Q3–Q4	$\chi^2 = 6.0093$	.1070	.2140	Cramér's <i>V</i> = 0.0923
51+	Breadth amount	Q1–Q2	<i>U</i> = 70739.5	.0027	.0162	<i>r</i> = 0.1040
51+	Breadth amount	Q1–Q3	<i>U</i> = 69184.0	.0055	.0275	<i>r</i> = 0.0984
51+	Breadth amount	Q1–Q4	<i>U</i> = 68927.5	.0522	.2088	<i>r</i> = 0.0667
51+	Breadth amount	Q2–Q3	<i>U</i> = 61218.0	.8534	.8534	<i>r</i> = -0.0061
51+	Breadth amount	Q2–Q4	<i>U</i> = 60823.5	.2595	.7784	<i>r</i> = -0.0374
51+	Breadth amount	Q3–Q4	<i>U</i> = 60164.0	.3412	.7784	<i>r</i> = -0.0315
51+	Volume	Q1–Q2	<i>U</i> = 70226.0	.0105	.0525	<i>r</i> = 0.0960
51+	Volume	Q1–Q3	<i>U</i> = 73137.5	.0001	.0006	<i>r</i> = 0.1611
51+	Volume	Q1–Q4	<i>U</i> = 69917.5	.0341	.1128	<i>r</i> = 0.0820

Conversation-length stratum	Outcome	Quartile comparison	Test statistic	Raw p	Holm-adjusted p	Effect size
51+	Volume	Q2–Q3	U = 65326.0	.0792	.1584	$r = 0.0606$
51+	Volume	Q2–Q4	U = 62277.5	.6809	.6809	$r = -0.0144$
51+	Volume	Q3–Q4	U = 57404.5	.0282	.1128	$r = -0.0759$

Note. Holm correction was applied separately across the six pairwise conversation-quartile comparisons within each stratum $\times$ outcome combination. Depth distributions were compared using permutation chi-square tests with Cramér's V as the effect-size measure. Breadth amount and volume were compared using permutation Mann-Whitney U tests with rank-biserial r as the effect-size measure. For rank-biserial r , positive values indicate higher ranks in the first-listed quartile, whereas negative values indicate higher ranks in the second-listed quartile.

D.9 Self-Disclosure Depth Distributions by Cue Group and Conversation Quartile

Section 4.4.3 presents the depth distributions across conversation quartiles for the four mutually exclusive cue groups. Table D8 provides the complete numerical distributions underlying Figure 24. Percentages were calculated separately within each cue-group × conversation-quartile combination.

Table D8

Complete Depth Distributions by Cue Group and Conversation Quartile

Conversation quartile	N	Superficial (%)	Middle (%)	Inner (%)	Core (%)
Panel A. No cue					
Quartile 1	53	32.08	26.42	18.87	22.64
Quartile 2	40	62.50	20.00	12.50	5.00
Quartile 3	42	52.38	26.19	19.05	2.38
Quartile 4	97	54.64	22.68	16.49	6.19
Panel B. Recommendation only					
Quartile 1	175	25.71	34.29	32.00	8.00
Quartile 2	183	27.87	39.34	21.86	10.93
Quartile 3	154	42.21	27.92	25.97	3.90
Quartile 4	145	37.24	29.66	28.97	4.14
Panel C. Textual anthropomorphic cue only					
Quartile 1	39	43.59	23.08	20.51	12.82
Quartile 2	28	53.57	35.71	0.00	10.71
Quartile 3	35	54.29	14.29	20.00	11.43
Quartile 4	44	27.27	43.18	18.18	11.36
Panel D. Both cues					
Quartile 1	423	28.61	25.06	32.62	13.71
Quartile 2	402	31.59	27.61	28.86	11.94
Quartile 3	401	32.67	35.66	23.44	8.23
Quartile 4	378	39.42	25.40	26.19	8.99

Note. n indicates the number of analysed turn pairs within each cue-group × conversation-quartile combination. Percentages were calculated within each combination and sum to approximately 100% across the four depth levels due to rounding. No cue indicates that neither a recommendation nor a textual anthropomorphic cue was present in the preceding ChatGPT message; Recommendation only indicates recommendation presence without a textual anthropomorphic cue; Textual anthropomorphic cue only indicates cue presence without a recommendation; Both cues indicates the presence of both features. Recommendation presence was classified irrespective of observable uptake.

D.10 Complete Secondary Cue-Group Comparisons

Section 4.4.3 reports the secondary cue-context comparisons that remained statistically supported after Holm correction. Table D9 and Table D10 provide the complete results, including comparisons that were not statistically supported. Table D9 reports differences across conversation quartiles within each cue group, whereas Table D10 reports differences across cue groups within each conversation quartile.

Table D9

Conversation-Quartile Differences Within Cue Groups

Cue group	Outcome	<i>N</i>	Test statistic	Raw <i>p</i>	Holm-adjusted <i>p</i>	Effect size
No cue	Depth distribution	232	$\chi^2(9) = 20.8072$	.0153	.0306	$V = .1729$
No cue	Breadth amount	232	$H(3) = 1.2425$	.7488	.8855	$\epsilon^2 = -.0077$
No cue	Volume	232	$H(3) = 3.4993$	.3201	.3610	$\epsilon^2 = .0022$
Recommendation only	Depth distribution	657	$\chi^2(9) = 24.8343$	.0030	.0090	$V = .1122$
Recommendation only	Breadth amount	657	$H(3) = 2.6679$	.4428	.8855	$\epsilon^2 = -.0005$
Recommendation only	Volume	657	$H(3) = 6.9163$	.0701	.2103	$\epsilon^2 = .0060$
Textual anthropomorphic cue only	Depth distribution	146	$\chi^2(9) = 16.3521$	.0546	.0546	$V = .1932$
Textual anthropomorphic cue only	Breadth amount	146	$H(3) = 4.5491$	.2053	.6158	$\epsilon^2 = .0109$
Textual anthropomorphic cue only	Volume	146	$H(3) = 4.7781$	.1805	.3610	$\epsilon^2 = .0125$
Both cues	Depth distribution	1,604	$\chi^2(9) = 32.0313$	.0001	.0004	$V = .0816$

Cue group	Outcome	N	Test statistic	Raw <i>p</i>	Holm-adjusted <i>p</i>	Effect size
Both cues	Breadth amount	1,604	$H(3) = 13.4564$	.0040	.0160	$\epsilon^2 = .0065$
Both cues	Volume	1,604	$H(3) = 15.8172$	.0012	.0048	$\epsilon^2 = .0080$

Note. Conversation quartiles were compared separately within each cue group. Depth distributions were compared using permutation chi-square tests with Cramér's V as the effect-size measure. Breadth amount and volume were compared using permutation Kruskal-Wallis tests with ϵ^2 as the effect-size measure. Holm correction was applied across the four cue groups separately within each outcome. Negative ϵ^2 values are untruncated sample estimates and indicate effect-size estimates effectively at zero.

Table D10

Cue-Group Differences Within Conversation Quartiles

Quartile	Outcome	N	Test statistic	Raw <i>p</i>	Holm-adjusted <i>p</i>	Effect size
Q1	Depth distribution	690	$\chi^2(9) = 19.5829$	.0198	.0345	$V = .0973$
Q1	Breadth amount	690	$H(3) = 14.2445$	.0019	.0076	$\epsilon^2 = .0164$
Q1	Volume	690	$H(3) = 0.3521$	.9504	1.0000	$\epsilon^2 = -.0039$
Q2	Depth distribution	653	$\chi^2(9) = 37.3806$	.0001	.0004	$V = .1381$
Q2	Breadth amount	653	$H(3) = 4.7796$	.1862	.5585	$\epsilon^2 = .0027$
Q2	Volume	653	$H(3) = 3.5865$	.3055	1.0000	$\epsilon^2 = .0009$
Q3	Depth distribution	632	$\chi^2(9) = 21.0859$	.0126	.0345	$V = .1055$
Q3	Breadth amount	632	$H(3) = 0.5896$	.8980	1.0000	$\epsilon^2 = -.0038$
Q3	Volume	632	$H(3) = 0.8170$	.8513	1.0000	$\epsilon^2 = -.0035$
Q4	Depth distribution	664	$\chi^2(9) = 21.9028$	.0115	.0345	$V = .1049$
Q4	Breadth amount	664	$H(3) = 1.7604$	.6250	1.0000	$\epsilon^2 = -.0019$
Q4	Volume	664	$H(3) = 3.7980$	.2865	1.0000	$\epsilon^2 = .0012$

Note. Cue groups were compared separately within each conversation quartile. Depth distributions were compared using permutation chi-square tests with Cramér's V as the effect-size measure. Breadth amount and volume were compared using permutation Kruskal-Wallis tests with ϵ^2 as the effect-size measure. Holm correction was applied across the four conversation quartiles separately within each outcome. Negative ϵ^2 values are untruncated sample estimates and indicate effect-size estimates effectively at zero.

Appendix E:

Use of AI and Other Digital Tools

This appendix documents the use of generative AI and other digital tools throughout the research and thesis-writing process. It describes how these tools were used across different parts of the research process. These included writing and revision, literature support, operationalisation, programming and statistical analysis, and the preparation of figures and tables. Particular attention is given to the extent of generative-AI involvement, the verification of AI-assisted outputs, the handling of research data, and the distinction between tool-supported tasks and substantive decisions made by the researcher. The appendix concludes with a reflection on the benefits and limitations of generative AI encountered during the thesis process.

E.1 Overview of Generative AI and Digital Tool Use

Generative AI and other digital tools supported several stages of the research process. Generative AI was used primarily for programming support, assistance in locating potentially relevant academic literature and clarifying selected concepts or passages, and revision of text that had already been developed by the researcher. Conventional digital tools supported data access and organisation, programming and statistical analysis, reference management, document preparation, and the production and editing of visual materials. Table E1 provides an overview of the tools used, their main purpose, the stage of the thesis in which they were applied, and the extent of their use.

AI outputs were reviewed before being incorporated into the research. Publications identified with AI assistance were independently accessed and read by the researcher before being used in the thesis. AI-assisted Python code was executed and checked in Jupyter Notebook, while proposed textual revisions were manually evaluated and subsequently accepted, modified, or rejected. All coded classifications of the conversation data were made manually by the researcher. No raw conversation data, individual user or ChatGPT messages, or participant-level conversation records from the research dataset were provided to AI. Sections E.2–E.8 provide more detailed descriptions of AI use, verification procedures, data-handling boundaries, and researcher responsibility.

ChatGPT from OpenAI was used between February and September 2026. As OpenAI updated and replaced the underlying models during this period, multiple ChatGPT model versions from the GPT-5.2 through GPT-5.6 families were used. The exact model used for every individual interaction was not systematically recorded.

Table E1
Overview of Generative AI and Digital Tool Use

Tool	Main use	Stage of thesis	Extent of use
ChatGPT (GPT-5.2 – GPT-5.6)	Programming support; literature-search, clarification, and verification support; writing and structuring support; visual refinement and presentation review	Literature review, methodology and analysis, thesis writing and revision, and figure preparation	Extensive for programming and writing support; selective for literature and visual support
Grammarly	Grammar and language checking	Final revision	Selective
Python / Jupyter Notebook on the TU Delft server	Data exploration, sampling-frame construction, data preparation, manual conversation coding, statistical analysis, and generation of analytical outputs and statistical figures	Methodology and empirical analysis	Throughout data preparation, coding, and analysis
Excel	Literature review tracking and development of the preliminary coding scheme	Literature review and coding- scheme development	Throughout literature review and codebook development
Word	Thesis drafting and formatting	Thesis writing	Throughout
draw.io	Diagram creation	Figure preparation	Selective
Mendeley	Reference management	Literature review and writing	Throughout
Windows PowerShell	SSH tunnelling to the research server	Data access	When server access was required
GIMP	Visual editing	Figure preparation	Selective

E.2 Writing and Revision Support

Generative AI was used to support the writing and revision process, primarily through ChatGPT, while Grammarly was used selectively for grammar, sentence-level clarity, and wording. The substantive reasoning, arguments, and thesis text were developed and written by the researcher. ChatGPT was used as a supporting tool to help translate this reasoning into a clearer and more structured academic presentation.

The role of ChatGPT changed over the course of the writing process. In the earlier stages, it was used mainly for brainstorming and structuring support, for example to consider where information and arguments could best be positioned within the thesis. The researcher subsequently wrote the corresponding text independently. As individual sections developed, ChatGPT was also used to suggest clearer section and paragraph structures based on information and arguments that had already been developed by the researcher. These suggested structures served as guidance, after which the researcher wrote and completed the sections.

During the later stages of the thesis, ChatGPT was used more frequently for revision and refinement of existing text. This included improving grammar and sentence clarity, academic and scientific tone, conciseness, transitions, and the organisation and flow of arguments that had already been developed. In some cases, existing sentences or paragraphs were reformulated to communicate the researcher's reasoning more clearly. ChatGPT was also used to identify potential duplication between sections. The researcher independently determined whether this information should be retained, removed, rewritten, or relocated.

After the thesis had been independently drafted, ChatGPT was additionally used as a review tool to compare draft sections and the thesis as a whole with the TPM thesis assessment rubric and relevant MOT lecture slides. Tables and figures were also reviewed against APA 7th edition formatting and presentation conventions. These reviews were used to identify possible weaknesses in structure, clarity, completeness, and presentation. The resulting suggestions were treated as feedback and were independently evaluated by the researcher before being accepted, modified, or rejected. ChatGPT was not used to determine the substantive arguments, methodological choices, interpretation of findings, or conclusions presented in the thesis. The researcher remained responsible for the final content and wording of the submitted thesis.

E.3 Literature and Conceptual Support

Generative AI was used selectively as a supporting and quality-control tool during the literature review. The main literature search was conducted by the researcher, primarily through Web of Science. The researcher formulated the principal search terms, assessed the retrieved publications, and determined which sources were relevant for inclusion. ChatGPT was used selectively to suggest additional search terms, concepts, authors, or literature areas that might warrant further investigation. It was also used as a coverage check to suggest additional literature areas for further consideration. Any resulting searches and decisions regarding relevance and inclusion were subsequently made by the researcher.

For selected publications, full-text papers were provided to ChatGPT to support summarisation and clarification of specific concepts, methods, or findings. These outputs were used as supplementary support rather than as substitutes for reading the original publications. All sources incorporated into the thesis were read and evaluated by the researcher. The researcher first independently identified and documented relationships, similarities, and differences across publications. ChatGPT was then used selectively to review whether these comparisons appeared complete and to improve the organisation of the researcher's written synthesis. The research gap and the interpretation of the literature were determined by the researcher.

The researcher's own literature-review draft was also provided to ChatGPT to support targeted retrieval of previously documented information and to facilitate focused questions about the reviewed literature. Any substantive information retrieved in this way was checked against the original publications before being incorporated into the thesis.

ChatGPT was also used as a quality-control aid for the literature review matrix. The researcher first extracted and recorded the relevant information from the selected publications. ChatGPT was subsequently used to flag potentially incomplete or inconsistent entries, including missing page references and possible discrepancies in reported samples, methods, and findings. It was also used to improve the readability and organisation of the literature matrix by restructuring information that the researcher had already extracted or recorded within individual cells. These outputs were used as organisational and verification support. The researcher subsequently returned to the original publications and independently determined whether factual corrections or changes to the recorded information were required. The researcher also verified whether ChatGPT's suggested revisions were supported by the source material.

The theoretical relationships incorporated into the conceptual framework were determined by the researcher. Generative-AI support for the visual refinement of the conceptual framework and other schematic figures is described in Section E.6. Citation and reference management were conducted using Mendeley, not with generative AI.

E.4 Coding and Operationalisation Support

The coding scheme and its substantive categories were developed by the researcher, as described in Section 3.6.1. ChatGPT was used only to improve the clarity and precision of selected operational definitions for the recommendation and textual anthropomorphic-cue codes. The underlying categories and their substantive meaning had already been determined by the researcher. The clearer formulations supported more consistent application of the established category boundaries during manual coding. ChatGPT was not used to develop the underlying categories or the self-disclosure dimensions.

Generative AI was not involved in coding individual observations. No raw conversation data, individual messages, or ambiguous coding cases were provided to ChatGPT, and ChatGPT did not propose classifications for any <AI, user> turn pair. All 2,639 turn pairs were classified manually by the researcher using the definitions and decision rules documented in Appendix B. Coding was conducted through a custom Python-based interface in Jupyter Notebook on the TU Delft TPM cybersecurity server.

E.5 Programming and Statistical Analysis Support

ChatGPT was used as a programming aid during the data-preparation, manual-coding, and analytical stages of the research. The researcher determined the data-preparation steps, coding workflow, and statistical procedures based on the research questions, variable characteristics, methodological considerations, and supervisor feedback. ChatGPT was then used to translate these decisions into executable Python code in Jupyter Notebook. This included code for data preparation, development of a custom coding interface that facilitated the manual classification of conversation turns, generation or modification of analytical functions and scripts, debugging errors, and explanation of relevant Python packages and functions.

ChatGPT did not determine which statistical tests, adjusted analyses, or sensitivity analyses were conducted. These analytical choices remained with the researcher or followed from supervisor feedback. The researcher independently interpreted the resulting estimates, significance tests, effect sizes, confidence intervals, and substantive patterns.

AI-assisted code was verified before its outputs were used in the thesis. All code was executed in Jupyter Notebook, intermediate outputs were inspected, analytical logic and formulas were checked, and sample counts were manually verified where applicable. No raw conversation data were provided to ChatGPT during programming or statistical analysis.

E.6 Figures, Tables, and Visual Materials

Different digital tools were used depending on the type of visual output. Statistical graphs were generated in Python through Jupyter Notebook using the corresponding analytical outputs. Statistical tables were constructed and formatted manually in Microsoft Word using results produced in Python through Jupyter Notebook. GIMP was used only for minor image preparation, such as cropping unnecessary white space where appropriate.

Schematic and overview figures were initially developed by the researcher in draw.io, including their substantive content, relationships, and overall structure. ChatGPT was subsequently used to improve their visual presentation, for example by refining alignment, spacing, colour use, arrows, and the arrangement of graphical elements. The visually refined versions were then transferred back into draw.io, where the researcher reviewed the content and made the final manual adjustments where needed before inclusion in the thesis. This process was used for all schematic and overview figures.

For statistical tables and figures, ChatGPT was provided only with selected numerical outputs that had already been chosen for reporting. It was used to provide feedback on clarity, presentation, and conformity with APA 7th edition presentation conventions. The underlying results and decisions about what to report remained with the researcher. The front and back cover designs of the thesis were generated using ChatGPT.

E.7 Data Handling, Verification, and Researcher Responsibility

No raw conversation data, individual user or ChatGPT messages, individual coded observations, coded datasets, or derived CSV/Excel data files were shared with ChatGPT. This boundary was particularly important because conversations in the underlying dataset may contain personal information even though participant profile files were excluded from the dataset collection (Karnam et al., 2026). ChatGPT received only thesis text, academic publications, relevant Python code or scripts, and selected aggregated analytical outputs that had already been chosen for reporting in the thesis. Complete Jupyter notebooks containing research data were not uploaded. Grammarly received thesis text for language editing but was not provided with the underlying conversation dataset.

Outputs produced with AI assistance were independently checked before being incorporated into the thesis. Literature-related suggestions and clarifications were verified against the original publications. AI-assisted Python code was executed in Jupyter Notebook, with intermediate outputs, analytical logic, formulas, and sample counts checked by the researcher. All coding classifications were made manually without AI involvement. AI-assisted textual revisions were individually evaluated and accepted, modified, or rejected according to whether they accurately represented the researcher's intended reasoning. Statistical results were generated in Python

through Jupyter Notebook and interpreted by the researcher. Values reported in the final tables and figures were checked against the corresponding Jupyter Notebook outputs.

The final responsibility for the research remained with the researcher. This included the research design, operationalisation of the theoretical constructs, individual coding decisions, selection of statistical procedures, interpretation of the empirical findings, development of the discussion and conclusions, and the final content of the submitted thesis. Generative AI provided specific programming, verification, literature-clarification, visual-presentation, and writing support, but did not independently determine the empirical classifications, analytical strategy, interpretation of findings, or substantive conclusions of the research.

E.8 Reflection on AI Use

The main value of generative AI in this thesis was not the independent creation of research content, but the acceleration and refinement of tasks for which the substantive direction had already been established. ChatGPT saved time when debugging Python code, supported clarification of difficult academic material, and helped improve the visual presentation of figures that had already been developed by the researcher. Its most important contribution, however, was to the writing process. I initially found it difficult to consistently translate complex reasoning into a clear, reader-oriented academic structure. Using ChatGPT iteratively as a writing and structuring aid helped me present arguments more clearly, distinguish main points from supporting details in the written presentation, and communicate reasoning more concisely. Useful revisions generally required the substantive argument and supporting information to be developed first by the researcher before ChatGPT could effectively help refine their presentation.

At the same time, extensive use of ChatGPT made its limitations apparent. It sometimes misunderstood the intended argument, strengthened claims beyond what the evidence supported, removed important nuance during rewriting, or produced text that was overly generic or verbose. It also occasionally suggested incorrect or non-existent references and sometimes generated programming errors. Visual and textual outputs therefore frequently required verification and several iterations before they accurately represented the intended content. The main lesson from this process was that generative AI was most useful as an iterative support tool after the substantive direction had already been established, but unreliable when treated as a substitute for independent academic judgement. Overall, its value lay primarily in helping refine and communicate work that had already been substantively developed by the researcher.



Delft University of Technology
Faculty of Technology, Policy and Management
Management of Technology