Preparing to Quit: A Thematic Analysis of Smokers' engagement with Conversational Agent-Guided Activities in Online Cessation Interventions

Jason Miao
Supervisor: Willem-Paul Brinkman[1]
[1]EEMCS, Delft University of Technology, The Netherlands

2025

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 18, 2025

Name of the student: Jason Miao
Final project course: CSE3000 Research Project
Thesis committee: Willem-Paul Brinkman, Inald Lagendijk

Abstract

Smoking cessation remains a persistent public health challenge, with digital interventions increasingly adopted to support behavior change. This study explores how smokers plan and engage with preparatory activities suggested by conversational agents in online cessation programs. It addresses three main questions: the factors that influence engagement with these activities, the ability of large language models to identify smokers' articulated plans, and the role of conditional "if-then" formulations in expressing coping strategies. A thematic analysis was conducted on qualitative user responses using both manual coding and automated labeling by local large language models. The findings show that smokers create context-sensitive plans shaped by emotional states, routines, and perceived usefulness of the suggestions, often expressed through conditional intentions. While manual analysis produced consistent and detailed themes, the large language model showed low agreement, highlighting limitations in current AI-driven qualitative analysis. These results inform the design of more adaptive digital cessation tools and contribute to the understanding of AI's role in supporting thematic research.

1 Introduction

Despite global public health efforts, smoking remains a leading cause of preventable death which contributes to a range of chronic diseases and economic burdens[4]. Digital health interventions, especially those that utilize conversational agents, have surfaced as scalable, cost-efficient tools to support personalized support on top of encouragement and behavior change strategies for users [11]. Among the techniques employed by such agents are *preparatory activities*. These activities can be, but are not limited to, setting a quit date, identifying triggers, or establishing social support, which are known to increase chances of a successful quit attempt. Although the effectiveness of conversational agents in smoking cessation is increasingly documented, **the interpretation, acceptance, and intended implementation of these preparatory activities by users remain largely unexplored** [19]. Gaining understanding of these behavioral intentions is essential for improving the design and impact of future digital interventions.

To address this, this study investigates the following research questions.

While conversational agents are increasingly used in smoking cessation programs to suggest preparatory activities like setting quit dates or identifying triggers, studying users' interpretation and planning to carry out these suggestions is an aspect that can further aid the improvement of such agents. Understanding how smokers translate these prompts into concrete intentions or actions is crucial for designing more effective, personalized interventions that align with users' real-world contexts and behavioral readiness. This indicates to the construction of the **main research question**:

How do smokers plan to do the proposed preparatory activities by conversational agents as part of online smoking cessation intervention?

Additionally, with the increasing development and usage of AI in research workflows, it is crucial to examine whether large language models can support or even scale up thematic analysis. Evaluating their ability to interpret nuanced user input can inform their role as tools in qualitative research and intervention development. The **first research subquestion** is constructed on the support of large language models:

How effectively can large language models identify and categorize smokers' articulated plans in response to preparatory activity suggestions within online cessation interventions?

Moreover, implementation intentions, often expressed in "if-then" format, have been shown to enhance behavior change and self-regulation. Investigating how smokers naturally use this structure in response to preparatory suggestions from conversational agents can reveal how users internalize and operationalize coping strategies. This, in turn, provides insight into the depth of cognitive engagement and readiness for behavior change within digital cessation interventions, which is reflected in the **second research sub-question**:

How do smokers use implementation intentions formulations to express conditional intentions or coping strategies when responding to preparatory suggestions from conversational agents?

Thematic analysis is the main method used in this research. This method is used to analyze qualitative data, like open-ended survey responses or interview transcripts. It involves looking for patterns or themes in what people say. By conducting a thematic analysis of responses from smokers interacting with an online cessation conversational agents, this research sheds light on the motivational and cognitive factors that shape participants' intentions. These findings show that users' engagement with preparatory activities are shaped by the current emotional readiness, prior quitting experience, and the perceived realism or feasibility of the tasks. This work contributes to the field by offering a nuanced understanding of how users navigate and internalize digital behavior change strategies, paving the way for more adaptive and personal interventions.

The remainder of this paper is organized as follows. Section 2 outlines the methodological approach and experimental setup, including the data collection and thematic analysis procedures, and the formal problem description. Section 3 states the results of the comparison and data analysis between self, peers, and large language models. Section 4 evaluates the results, with discussions on the limitations. Section 5 discusses responsible research and the reproducibility of this method. Section 6 considers the potential directions for future work. Finally, Section 7 conveys the summarized results and the answers to the research questions and sub-questions.

1.1 Related Works

Conversational agents have been increasingly integrated into digital health interventions, particularly in the domain of behavior change. Several studies demonstrate their potential to enhance personalization and user engagement. Marler et al. (2019) investigated the general effectiveness of digital tools in tobacco cessation, emphasizing the potential of these technologies for personalization and user engagement [16]. Their findings suggest that tailored content and user-centered design notably improve intervention outcomes.

Earlier work by Bickmore et al. (2003) introduced relational agents, which are virtual coaches designed to establish trust and long-term engagement with users. This study demonstrated that users often develop social bonds with such agents, which can enhance motivation and adherence to behavioral goals [6]. These findings underscore the value of affective and relational components in digital interventions, particularly in sustained behavior change efforts.

Building on this foundation, Perski et al. (2017) conducted a systematic review of digital smoking cessation interventions, identifying key design features, such as interactivity, feedback, and personalization, that affect user engagement [18]. However, the review primarily focused on intervention characteristics and their correlation with effectiveness, rather than delving into users' subjective responses or cognitive strategies.

One notable approach to bridging intention and action is implementation intentions, of-

ten expressed through "if-then" reasoning. Armitage (2008) tested a volitional help sheet encouraging users to plan ahead using if-then statements (e.g., "If I feel stressed, then I will go for a walk"). The study showed that participants who formed such plans had considerable higher quit rates compared to controls, reinforcing the value of concrete, situational strategies in behavior change [2].

While these studies collectively affirm the potential of conversational agents and digital tools in supporting health-related behavior change, they often stop short of analyzing user-generated content in depth. Specifically, little attention has been paid to the thematic patterns underlying smokers' intentions, coping mechanisms, and psychological framing when responding to preparatory suggestions. There is limited research that thematically dissects how users interpret and plan to implement the advice offered by these agents, particularly when it comes to conditional or "if-then" reasoning.

Furthermore, few studies explore the application of large language models in supporting thematic analysis of qualitative health data. As AI tools become increasingly accessible, understanding their potential and limitations in qualitative coding becomes essential, especially when compared to manual human analysis.

This thesis addresses these apertures by thematically analyzing smokers' responses to preparatory activities suggested by conversational agents and evaluating the capacity of local large language models to replicate or assist with this process. In doing so, it contributes both to the behavioral science of smoking cessation and to the emerging intersection between qualitative research and AI-assisted methods.

1.2 Use of Secondary Data

This research involves secondary data analysis of an existing dataset collected from an online smoking cessation intervention that utilizes CAs conducted by Albers (2017) [1]. The dataset comprises anonymized, open-text responses from smokers engaging with preparatory activities suggested by the CA system. These responses include behavioral reflections, plans, concerns, and motivations, making them ideal for qualitative inquiry.

Secondary data analysis was chosen for several reasons. First, it allows for a cost-effective and time-efficient examination of naturally occurring user behavior in a real-world intervention context [12]. Second, using pre-existing data eliminates ethical concerns associated with recruiting vulnerable populations (e.g., smokers attempting to quit) while still enabling insights into behavioral planning and agent interaction. Lastly, this approach supports ecological validity, as the responses reflect actual user engagement rather than artificial or prompted survey answers.

Although the data were not originally collected for thematic analysis, it is highly compatible with the research aims. This aligns with Heaton's (2004) typology of "supplementary" secondary analysis, where new research questions are applied to an existing dataset for additional theoretical insights.

2 Methodology

This study investigates how smokers plan to undertake preparatory activities suggested by conversational agents (CAs) in an online smoking cessation context. To answer the research questions, particularly those concerning underlying themes, planning intentions, and the linguistic structures used by smokers: a qualitative thematic analysis was conducted, followed

by an evaluation of automated coding using local large language models (LLMs). Additional validation was performed through inter-coder reliability analysis and peer triangulation.

Inter-coder reliability refers to the degree of agreement among different researchers (coders) when they independently code or categorize qualitative data, such as interview transcripts or open-ended survey responses. It serves as a measure of consistency in the application of a coding scheme across multiple analysts. A study by MacPhail et al. (2015) discusses the importance of establishing inter-coder reliability in qualitative research. The authors highlight that inter-coder reliability is crucial for assessing the agreement among multiple coders and for identifying weaknesses such as imprecise code definitions or overlapping meanings in the coding scheme [15].

Triangulation is considered a robust strategy for validating qualitative research findings. By converging information from different sources, researchers can test the consistency and reliability of their data. For instance, Carter et al. (2014) discuss how triangulation serves as a qualitative research strategy to test validity through the convergence of information from different sources [8].

2.1 Rationale for Thematic Analysis

Thematic analysis was selected as the primary method due to its flexibility and effectiveness in analyzing open-text qualitative data [7]. Unlike grounded theory or content analysis, which may require more rigid assumptions about theory generation or frequency counting, thematic analysis is especially well-suited for identifying, interpreting, and organizing patterns of meaning across rich textual responses. Given that user inputs in this study are highly variable in form, tone, and content, thematic analysis allowed for capturing both surface-level content (semantic themes) and deeper intent or logic (latent themes), such as implementation intentions expressed through "if-then" statements [10].

2.2 Manual Thematic Analysis Procedure

The manual analysis followed Braun and Clarke's (2006) six-phase framework [7]. Firstly, the dataset was thoroughly read multiple times to develop an intuitive understanding of the content. For the next three steps, initial coding, code revision, and theme review, an initial coding scheme was developed using a peer-coding approach to reduce individual bias and enhance reproducibility.

A random sample of 150 responses was selected and independently analyzed by two researchers. Each researcher identified recurring patterns and proposed a preliminary set of themes. The independently developed coding schemes were then compared and discussed collaboratively. Discrepancies were resolved through discussion, and overlapping concepts were merged. This process led to the creation of a finalized coding scheme intended to increase the consistency and trustworthiness of the subsequent analysis.

The manual coding served as the ground truth for comparison with LLM-based thematic labeling. The following steps of the six-phase framework continues after the LLM comparisons.

2.3 Integration of Local LLMs

To explore the use of automated thematic analysis, a comparison was conducted using several open-source local LLMs, chosen based on current hardware compatibility and performance benchmarks: QwQ 32B, DeepSeek-R1, Gemma-1.1 2B, Qwen 2.5-3B, and LLaMA-3-8B.

These models were selected for their support for efficient local inference and the ability to generalize across text classification tasks.

Each LLM was prompted with structured instructions to identify themes from user responses, such as: Prompt Example: *"You are a qualitative research assistant. Read the following smoking cessation response and identify up to two key behavioral themes. Output format: [Theme 1], [Theme 2]. Text: <user input>"* The models were run on a consistent subset of the dataset, with outputs collected for comparison against the human-coded labels.

The complete prompt that was used for each local LLM is stated in the Appendix A.

2.4 Evaluation of Model Agreement with Human Coding

To identify and assess the agreement between LLMs and manual coding, Cohen's Kappa coefficient k was used. Cohen's Kappa is an accepted measure of inter-rater reliability for categorical data, as it accounts for the likelihood of agreement occurring by chance, unlike raw accuracy scores [17]. Alternative statistical methods, such as Pearson's correlation or power analysis, were considered but deemed inappropriate for nominal, non-ordinal thematic categories.

The use of Cohen's Kappa is especially pertinent given that we are not measuring continuous variation but rather categorical thematic classifications. A k score closer to 1 indicates high agreement, while a score closer to 0 suggests no better than random labeling.

2.5 Triangulation Through Peer Analysis

Peer triangulation was employed to strengthen the validity of the thematic structure. A second round of peer testing was conducted using a new random sample of 150 responses. Prior to this, the peer coder underwent a structured training process involving 50 separate responses. The first 25 responses were coded jointly, allowing for clarification and feedback. The remaining 25 were completed independently to assess consistency. The training was deemed sufficient after this phase.

Following the training, both researchers independently applied the finalized coding scheme to the new sample. Inter-coder reliability was then assessed using Cohen's Kappa, **which yielded a value of 0.74**. According to McHugh's (2012) interpretation, this indicates a moderate level of agreement, suggesting that the coding scheme was sufficiently clear and interpretable by others.

Finally, the validated coding scheme was subsequently used to code the full dataset. As analysis progressed, the scheme was iteratively refined. This involved making minor adjustments, introducing new themes where necessary, and consolidating existing categories to better reflect the data.

2.6 Application of Thematic Rubric via LLMs

To bridge the gap between manual thematic coding and automated analysis, this study included a crucial intermediate step: assessing how well LLM could apply a predefined thematic rubric to previously unlabeled user responses. This rubric is the same set of themes and subthemes developed during the triangulation phase, relating to smokers' articulated plans and strategies in response to preparatory suggestions.

Once the final coding scheme was established, each LLM (QwQ 32B, DeepSeek-R1, Gemma 1.1 2B, Qwen 2.5-3B, and LLaMA-3-8B) was prompted to assign themes to each textual response. Prompt templates were structured to simulate a coding instruction, such

as: *Given the following text: "[User response]" Label the text according to the following rubric categories: [Theme 1], [Theme 2].*

The complete prompt that was used is stated in the Appendix B.

This allowed for a direct one-to-one comparison between how LLMs applied these themes versus how human coders (including the primary researcher and peer annotators) did.

To measure the reliability of the LLM-applied coding, Cohen's Kappa coefficient was once again used to calculate inter-rater agreement between the LLMs and human coders for each theme across a random sample of annotated responses.

2.7 Summary of Methodological Workflow

To summarize, the methodology combined:

- Manual thematic analysis for establishing ground truth;
- Application of LLMs for automated theme generation;
- Inter-rater agreement evaluation via Cohen's Kappa;
- Triangulation through trained peer analysis to validate theme clarity;
- Application of thematic rubric via LLMs and comparison with manual labeling

This multi-method approach not only addresses the primary research questions regarding how smokers respond to preparatory suggestions but also explores the feasibility of using LLMs in qualitative coding tasks, providing both empirical and methodological contributions to digital health intervention research.

3 Results

This section presents the outcomes of the thematic analysis conducted on user responses regarding their experiences with preparatory activities proposed by CAs in the context of online smoking cessation. The themes were derived through a structured coding process (see Section 2.7) and reflect the dominant patterns in participants' open-ended responses.

3.1 Overview of Themes

The analysis yielded a set of overarching themes, each supported by multiple sub-themes. These themes encapsulate users' perceptions, intentions, challenges, and engagement levels with the proposed activities. The final themes are organized in the following conceptual structure, displayed in figure 1:

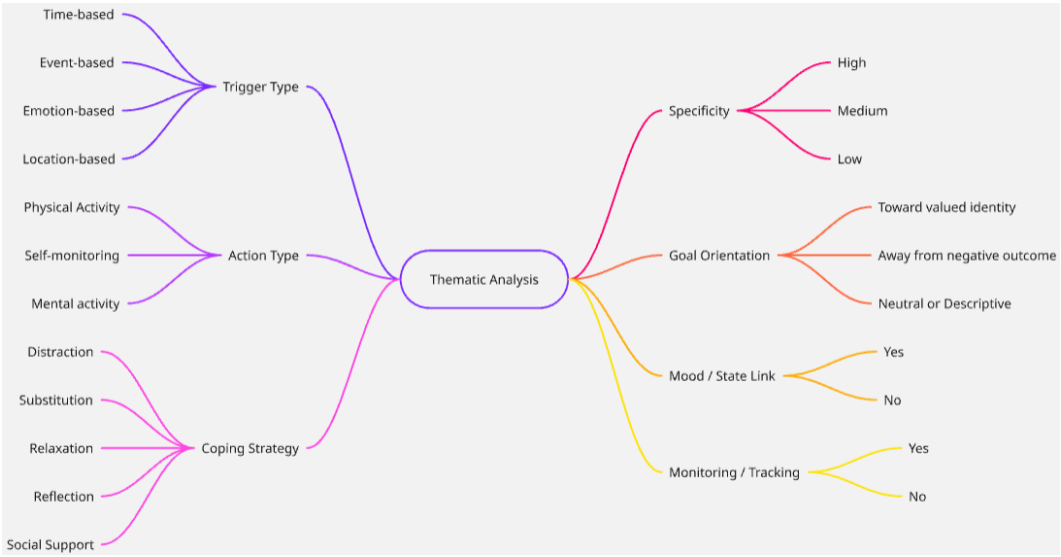

Figure 1: Mindmap of Manual Themes

The table 2 in Appendix C helps explain each of the themes with each of their guiding questions, respectively.

3.2 Example of Data and Thematic Categorization

To demonstrate how participant responses were categorized, the table 3 in Appendix D presents the final themes and sub-themes alongside raw excerpts from the dataset. This helps to clarify the interpretative process that underpinned the coding structure.

For instance, the theme "Trigger Type" was divided into four distinct sub-themes. A common sub-theme involved time-based triggers, where participants referenced specific times or routines that prompt reflection or action (e.g., P2, P32, P96). One such response reads:

If tomorrow i have to wait for my lesson then i will identify the reason why i want to become more phisycally active and quit smoking. (Session data with reflections action plans anonym, P2)

Another recurring theme involved participants' coping strategies in response to cravings or stressful situations. One sub-theme within this category was substitution, where participants planned to replace smoking with alternative activities (e.g., P154, P294). For example:

If i get streesed, i will distract myself instead of smiking, I'll do an actuvity such as cheeing gum or breathing control (Session data with reflections action plans anonym, P294)

These examples illustrate how raw data was organized within the thematic framework, offering insight into how user responses aligned with the identified themes and subthemes.

Importantly, these responses also shed light on how smokers interpreted and internalized the preparatory activities suggested by the CA. Many participants responded in concrete, actionable terms which indicates a general willingness to engage with the suggestions and adapt them to their personal context. For example, linking a quit strategy to "waiting for a lesson" or anticipating stress and planning a distraction response suggests users were thinking ahead and embedding new routines into their day-to-day structure. This demonstrates

a positive reception of the CA's guidance, especially when the advice was seen as realistic, customizable, and relevant to their immediate lives.

However, the variation in specificity also highlights differing levels of cognitive engagement. Some participants offered vague or non-committal responses (e.g., "I'll try something else"), while others detailed precise plans and coping techniques. This suggests that while some users may have fully embraced the CA's suggestions, others remained uncertain or disengaged. It could have been caused by low confidence, lack of clarity, or a mismatch between the suggestion and their lived experience.

Overall, these patterns reveal a complex but largely constructive view of the CA-recommended activities. Smokers generally recognized the potential usefulness of planning and reflection tasks but varied in how deeply they personalized the advice. Thematic analysis thus not only captures surface-level responses but also reveals how users are negotiating behavior change within their psychological and situational constraints.

3.3 Inter-LLM Agreement

As part of the analysis, various LLMs were used (see section 2.3) to assist in theme generation and preliminary coding. For instance, DeepSeek-R1 identified themes such as "Routine Disruption as a Motivator" and "Social Identity Shifts", both of which appeared consistently across a range of user reflections.

To evaluate the consistency of theme application across models, inter-coder reliability was assessed using Cohen's Kappa. This was calculated for each LLM compared to the human-coded baseline. Results indicated low agreement across all tested models, including Qwen 3, DeepSeek-R1, Gemma-1.1 2B, Qwen 3B, and LLaMA 3 8B. These low kappa scores suggest that while LLMs can assist with initial theme detection, human interpretation remains essential for reliable coding and refinement of qualitative data. The table below shows the local LLM model and its kappa score.

Model Name	Cohen's Kappa score
QwQ 32B	0.32
DeepSeek-R1	0.21
Gemma-1.1 2B	0.13
Qwen 2.5-3B	0.24
LLaMA-3-8B	0.17
Manual Inter-Rater	0.74

Table 1: Local LLM model name and the Cohen's Kappa score

The Cohen's Kappa scores presented in Table 1 reveal substantial differences in agreement levels between the manual coding and the local LLM-based automated coding approaches. The manual inter-rater reliability score of 0.74 indicates a strong level of agreement between human coders, demonstrating that the coding scheme is well-defined and can be consistently applied by trained researchers.

In contrast, all tested local LLMs yielded considerably lower Cohen's Kappa scores, ranging from 0.13 (Gemma-1.1 2B) to 0.32 (QwQ 32B). These values fall within the range commonly interpreted as "slight" to "fair" agreement (McHugh, 2012), suggesting that the LLMs struggle to reliably replicate the manual thematic coding scheme. As also seen in Table 1, the average k across all models is very low, highlighting the limitations of these models in accurately interpreting and categorizing complex qualitative data in this domain.

4 Discussion

4.1 Analysis of Manual Thematic Analysis

The manual thematic analysis identified a rich and nuanced set of themes related to smokers' planning and coping strategies in response to preparatory activities suggested by conversational agents. The **Trigger Type** theme encompasses multiple dimensions, including time-, event-, emotion-, and location-based triggers, highlighting the complex and contextual nature of smoking cues that smokers recognize and respond to.

Within **Action Type**, participants described a variety of behavioral responses such as physical activity, smoking-related behaviors, self-monitoring, and mental activities, reflecting diverse strategies to manage cravings or adhere to cessation plans. The **Coping Strategy** theme further revealed a wide repertoire including distraction, substitution, visualization, relaxation, reflection, and seeking social support, underscoring the multifaceted approach smokers take to maintain cessation efforts.

Themes related to **Specificity** and **Goal Orientation** provided insight into the precision of plans and motivational direction-whether smokers' intentions were aimed toward a valued identity (e.g., becoming a nonsmoker) or away from negative outcomes (e.g., avoiding health consequences). The linkage between **Mood/State** and the reported plans demonstrated that emotional states often influenced the activation of coping behaviors, while the presence or absence of **Monitoring/Tracking** suggested varied levels of self-regulatory engagement.

Overall, the manual thematic analysis captured detailed and interpretable patterns that offer meaningful insights into smokers' preparatory behaviors and mental frameworks.

4.2 Analysis of LLM Thematic Analysis

The themes generated by the local LLMs, although fewer and less granular, introduced novel but plausible categories such as **Routine Disruption as a Motivator** and **Social Identity Shifts**. These themes reflect important psychological constructs; for example, routine disruption aligns with the theory that changing habitual contexts can facilitate behavior change [14], and social identity shifts resonate with the importance of identity transformation in cessation [21].

However, the LLM-derived themes tended to be broader and less specific than the manual themes. For instance, while the manual analysis distinguished between multiple trigger types and coping strategies, the LLMs grouped some concepts into more generalized categories. This reflects a limitation in the LLMs' ability to detect detailed distinctions and capture contextual subtleties in smokers' responses.

4.3 Comparison of Manual and LLM Thematic Applications

The manual coding scheme was highly detailed and consistent, as demonstrated by the strong inter-rater reliability. In contrast, the LLM thematic labeling showed much lower agreement (Table 1), indicating challenges in automated interpretation. Human coders demonstrated greater precision in distinguishing themes such as specific coping strategies and the emotional underpinnings of plans, while the LLMs often conflated related but distinct themes or missed latent meanings.

Human coders excelled at capturing subtle distinctions and latent meanings, consistent with prior research emphasizing the importance of contextual understanding in thematic analysis. For example, Wachinger et al. (2024) found that while AI can identify descriptive

themes, it struggles with interpretative richness and nuance. They stated that ChatGPT predominantly identified rather descriptive themes, as well as supplemented by a few themes leaning toward more interpretative engagement [22].

Similarly, Prescott et al. (2024) concluded that LLMs can function well in initial coding phases but need human oversight for nuanced thematic interpretation. They indicated that ChatGPT demonstrated its ability to code, generate themes, but has its limitations, which necessitates the involvement of human researches [20].

Where the AI analysis added value was in highlighting broader, potentially overlooked themes like "Social Identity Shifts", suggesting that LLMs might assist in uncovering emergent or higher-level patterns that manual coding may not initially prioritize. Conversely, AI-generated labels sometimes introduced ambiguity due to inconsistent theme application or incomplete contextual understanding, which could misrepresent nuanced user intentions.

4.4 Possible Reasons for Disagreements Between Methods

Several factors likely contributed to discrepancies between manual and AI analyses. Manual coding benefits from domain knowledge, contextual interpretation, and the ability to resolve ambiguous responses through reflective judgment. LLMs, despite their language understanding capacity, can struggle with domain-specific jargon, implied meanings, and complex conditional statements typical in smoking cessation discourse.

Moreover, the local LLMs used here had limited fine-tuning for thematic coding and qualitative data interpretation, which constrained their ability to replicate human-level coding accuracy. The training data and prompts used may also have influenced LLM performance, highlighting the need for better model adaptation to domain-specific tasks.

4.5 Limitations

Several limitations affect the extent to which the conclusions of this study can be generalized beyond the present context. The primary research question, how smokers plan to engage with preparatory activities proposed by CAs, was investigated using a dataset collected from individuals already enrolled in an online cessation intervention. These participants are likely more motivated, self-selecting, and possibly more digitally literate than the general population of smokers. As a result, their planning behaviors, engagement with digital prompts, and openness to preparatory activities may not represent the broader spectrum of smokers, especially those who are less technologically experienced or earlier in their quitting journey.

Regarding the first research sub-question, which explores how effectively LLMs can identify and categorize smokers' articulated plans, the study was constrained by the use of local LLMs chosen based on available hardware. These models lacked domain-specific fine-tuning and exhibited relatively low agreement with human-coded themes. While this highlights current limitations of LLMs in complex qualitative coding tasks, the findings may not hold for more advanced or fine-tuned models. Therefore, any conclusions about LLM performance should be viewed as a snapshot under specific technical constraints and may not generalize to broader or future deployments where more sophisticated models are used.

As for the second research sub-question, which examines how smokers use implementation intentions ("if-then" reasoning), the thematic presence of such formulations was notable in the data. However, since participants were explicitly instructed to use an if-then formulation, this may have artificially increased the presence or clarity of implementation intentions in the dataset. In real-world applications where such prompts are absent or more loosely framed,

smokers may not spontaneously articulate conditional plans with the same structure or depth, limiting the ecological validity of these findings.

In sum, while this study offers valuable insights into smokers' planning behaviors and the potential utility of LLMs in supporting qualitative research, its conclusions should be interpreted within the boundaries of its context, population, and technical setup.

4.6 Implications for Practice and Research

Despite these limitations, the study provides valuable insights into how smokers conceptualize and communicate preparatory activities, informing the design of more personalized and context-aware conversational agents. The demonstrated challenges with LLM-based thematic analysis underscore the importance of careful integration of AI tools in qualitative research, where human judgment remains crucial.

5 Responsible Research

Throughout this study, we made a conscious decision to use local large language models instead of public, cloud based alternatives. This is due to the concerns of commitment to data privacy and security, which ensures that the content generated participant generated was not transmitted to external servers. By maintaining control over the computational environment, we could better manage ethical and technical considerations around sensitive behavioral and health-related data.

5.1 Bias and Reliability

To reduce bias and increase the reliability of the thematic analysis, a manual peer analysis was conducted alongside AI-assisted analysis. A second analyst first independently reviewed the dataset to recognize their own patterns and realizations from the dataset. In addition, discussions were made to help identify inconsistencies and further refine the coding scheme. Providing another peer in this process furthermore enhances the transparency of thematic choices, which supports the reproducibility of results, and additionally strengthens the triangulation of findings across different methods.

Although these efforts attempt to increase responsibility, it remains important to be conscious of the inherent subjectivity and potential for bias in thematic analysis, particularly when involving LLMs. Both manual and automated approaches can be shaped by personal interpretations, contextual assumptions, and a tendency to focus on certain themes over others. Even with the inclusion of peer validation, complete objectivity in qualitative research is not always achievable, so findings and discoveries should be approached with careful interpretation and contextual sensitivity.

5.2 Use of LLMs

When using LLMs, prompt design plays a pivotal role in shaping their output, and this introduces another layer of variability. The public dataset, derived from a specific context and intervention, also carries limitations that may affect reproducibility. Specifically on the various sets of participants, they have other own language styles, various language proficiency, contrasting ways and abilities to express oneself, response length, and interpretation

of the tasks. These factors could influence how well future models or researchers replicate the thematic framework.

An important consideration is that locally deployed LLMs may have been trained on datasets that include publicly available user-generated content, potentially even similar thematic analysis studies. This raises concerns about data contamination, where models might internalize and reproduce patterns that closely mirror our results, undermining the originality of new analyses. Research has documented that closed source LLMs, such as GPT-3.5 and GPT-4, have absorbed millions of benchmark examples, complicating the reproducibility of novel findings [5]. In the context of privacy and anonymity, additional risks arise when models inadvertently memorize identifiable or sensitive data from their training sources [9].

Additionally, the dynamic nature of LLM training, especially in continuously updated systems, raises further concerns. As models evolve, their outputs may change, making it difficult to replicate exact analyses over time. This instability, combined with potential privacy risks from memorized content researched by Ippolito et al. (2023), underscores the importance of transparent reporting. Researchers must document which model versions were used, how outputs were verified, and where human oversight occurred to ensure responsible use of LLMs in qualitative analysis [13].

5.3 Ethics of AI

Ethically, integrating AI into qualitative research demands rigorous oversight. While LLMs may offer speed and appear precise, their outputs can conceal biases, perpetuate stereotypes, or omit nuanced contextual meaning, risks that are especially critical in behavioral health studies. Ashwin et al. (2023) demonstrate that LLMs can introduce serious annotation biases in qualitative data, skewing interpretations based on demographic or contextual factors [3].

Finally, the broader societal implications of AI-assisted qualitative research should be considered. While AI can democratize access to analytical tools and speed up processes, it also risks reducing complex human experiences to oversimplified patterns if misapplied. Responsible research must remain rounded in human interpretation, ethical scrutiny, and methodological transparency, ensuring that technological advances enhance rather than compromise the depth and rigor of social science inquiry.

6 Future Work

Future research could benefit from integrating active learning techniques into the thematic analysis pipeline. Active learning allows models to iteratively improve by selecting the most informative samples for manual labeling, which can reduce human workload while maintaining high annotation quality. Prior studies in qualitative text analysis have shown that active learning can effectively bridge the gap between automated theme detection and nuanced human interpretation, particularly in health behavior research.

Additionally, expanding the dataset would allow for more robust theme discovery and better generalization across different user groups. With a larger corpus, it would also be feasible to run comparisons across multiple local LLMs to identify which models best support theme development, and under what conditions. This could further highlight the strengths and limitations of different model architectures in qualitative settings.

Future work may also explore alternative machine learning methods for identifying latent behavioral patterns, such as clustering techniques, topic modeling (e.g., BERTopic or LDA),

or hybrid human-AI co-coding systems. These tools could support a more data-driven but still interpretable analysis pipeline, which remains critical in health research domains.

Finally, establishing frameworks that combine manual coding with LLM assistance in a structured, reproducible manner could help define best practices for the responsible use of AI in qualitative research.

7 Conclusion

This study investigated how smokers plan to engage with preparatory activities proposed by conversational agents within online smoking cessation interventions. The primary research question focused on understanding smokers' planning processes in response to these suggestions, while three sub-questions explored specific influencing factors, the performance of large language models in thematic analysis, and the use of conditional "if-then" formulations in smokers' responses.

In addressing the main research question: How do smokers plan to do the proposed preparatory activities by conversational agents as part of online smoking cessation interventions?, the manual thematic analysis revealed that smokers develop complex, context-dependent plans involving various triggers, such as time, events, emotions, and locations, that influence their preparatory behaviors. Smokers employ diverse coping strategies, including distraction, substitution, relaxation, and social support, often tailoring their actions to fit personal goals oriented either toward a valued identity or away from negative outcomes. This demonstrates that smokers actively use preparatory activities as part of their cessation journey, adjusting plans based on internal states and environmental cues.

Regarding sub-research question 1: How effectively can large language models identify and categorize smokers' articulated plans in response to preparatory activity suggestions within online cessation interventions?, the evaluation of local large language model showed that while these models could generate some relevant themes and even introduce novel categories such as "Routine Disruption as a Motivator" and "Social Identity Shifts", their overall agreement with manual coding was low. This indicates limited effectiveness in reliably identifying and categorizing smokers' articulated plans. The large language model struggled with fine-grained distinctions and contextual nuances, underscoring current limitations of AI in thematic qualitative analysis.

For sub-research question 2: How do smokers use implementation intentions formulations to express conditional intentions or coping strategies when responding to preparatory suggestions from conversational agents?, the manual analysis highlighted frequent use of "if-then" conditional statements by smokers to express coping strategies and intentions. This "if-then" structure enables smokers to anticipate high-risk situations and prepare adaptive responses, reflecting a strategic and flexible approach to behavior change. It reveals that conditional planning is an important cognitive mechanism facilitating self-regulation during smoking cessation.

The key contributions of this research include deep insights into the nuanced ways smokers plan preparatory activities during online cessation, as well as a critical examination of AI tools' capacity to support qualitative thematic analysis. These findings can inform the design of more personalized conversational agents that better accommodate smokers' contextual needs and advance the integration of human expertise with AI-assisted methodologies.

In summary, this work not only advances understanding of smoker behavior in digital interventions but also highlights the promise and current challenges of employing AI in qualitative health research, laying groundwork for future improvements in both domains.

8 Acknowledgments

I would like to express my sincere gratitude to Prof. Willem-Paul Brinkman for his invaluable guidance, constructive feedback, and continuous support throughout the research process. His insights greatly shaped the direction and clarity of this work.

Additionally, I acknowledge the use of large language models to assist in and refining the grammar, vocabulary, and structure of this thesis for clarity.

References

[1] Nele Albers, Willem-Paul Brinkman, and Mark Neerincx. Preparing for quitting smoking and becoming more physically active with a virtual coach: Reflections for persuasive messages and action plans. *4TU.ResearchData*, 2017.

[2] Armitage and Christopher J. A volitional help sheet to encourage smoking cessation: A randomized exploratory trial. *Health Psychology*, 2008.

[3] Julian Ashwin, Aditya Chhabra, and Vijayendra Rao. Using large language models for qualitative analysis can introduce serious bias. *arxiv*, 2023.

[4] Bridget Balch. Smoking is still the leading cause of preventable death in the u.s. doctors may soon have new tools to help people quit. *AAMC*, 2024.

[5] Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In Yvette Graham and Matthew Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, March 2024.

[6] Timothy Bickmore and Rosalin W. Picard. *Relational agents: effecting change through human-computer relationships*. PhD thesis, MIT, 2003.

[7] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Taylor & Francis Online*, 2006.

[8] Nancy Carter, Denise Bryant-Lukosius, Alba DiCenso, Jennifer Blythe, and Alan J Neville. The use of triangulation in qualitative research. *National Library of Medicine PubMed*, 2014.

[9] Georgios Feretzakis and Vassilios S. Verykios. Trustworthy ai: Securing sensitive data in large language models. *AI*, 5, 2024.

[10] Peter M. Gollwitzer. Implementation intentions: Strong effects of simple plans. *APA PsycNet*, 1999.

[11] Kara Hartnett. Conversational ai in healthcare: Benefits and use cases. *RASA*, 2025.

[12] Janet Heaton. Reworking qualitative data. *SAGE Publications*, 2004.

[13] Daphne Ippolito, Florian Tramer, Milad Nasr, Chiyuan Zhang, Matthew Jagielski, Katherine Lee, Christopher Choquette Choo, and Nicholas Carlini. Preventing generation of verbatim memorization in language models gives a false sense of privacy. In C. Maria Keet, Hung-Yi Lee, and Sina Zarrieß, editors, *Proceedings of the 16th International Natural Language Generation Conference*, pages 28–53, Prague, Czechia, September 2023. Association for Computational Linguistics.

[14] Phillippa Lally, Cornelia H. M. van Jaarsveld, Henry W. W. Potts, and Jane Wardle. How are habits formed: Modelling habit formation in the real world. *Wiley.*, 2009.

[15] Catherine MacPhail, Nomhle Khoza, and Meghna Ranganathan. Process guidelines for establishing intercoder reliability in qualitative studies. *Sage Journals*, 2015.

[16] Jennifer D Marler, Craig A Fujii, David S Utley, Lydia J Tesfamariam, Joseph A Galanko, and Heather Patrick. Initial assessment of a comprehensive digital smoking cessation program that incorporates a mobile app, breath sensor, and coaching: Cohort study. *JMIR Mhealth Uhealth*, 2019.

[17] Marry L. McHugh. Interrater reliability: the kappa statistic. *Biochemia Medica*, 2012.

[18] Olga Perski, Ann Blandford, Robert West, and Susan Michie. Conceptualising engagement with digital behaviour change interventions: a systematic review using principles from critical interpretive synthesis. *Oxford Academic*, 2017.

[19] Olga Perski, David Crane, Emma Beard, and Jamie Brown. Does the addition of a supportive chatbot promote user engagement with a smoking cessation app? an experimental study. *DOAJ*, 2019.

[20] Maximo R Prescott, Samantha Yeager, Lillian Ham, Carlos D Rivera Saldana, Vanessa Serrano, Joey Narez, Dafna Paltin, Jorge Delgado, David J Moore, and Jessica Montoya. Comparing the efficacy and efficiency of human and generative ai: Qualitative thematic analyses. *JMIR AI*, 3, 2024.

[21] Peggy A. Thoits. Role-identity salience, purpose and meaning in life, and well-being among volunteers. *Social Psychology Quarterly*, 2012.

[22] Jonas Wachinger, Kate Barnighausen, Louis N. Schafer, Kerry Scott, and Shannon A. McMahon. Prompts, pearls, imperfections: Comparing chatgpt and a human researcher in qualitative data analysis. *Qualitative Health Research*, 2024.

A Prompt for Local LLM Theme Generation

Context: You are analyzing a set of "if-then" action plans written by smokers participating in an online smoking cessation program. These action plans were created in response to preparatory tasks proposed by a conversational agent (CA). The goal is to understand how participants plan to carry out these suggested tasks.

Research Question: How do smokers plan to do the proposed preparatory activities by conversational agents as part of online smoking cessation intervention?

Your Task:

1. Carefully read the list of 150 "if-then" action plans.
2. Identify patterns in how participants intend to engage with the preparatory activities.
3. Group responses into broad themes based on shared intentions, strategies, or conditions.
4. For each theme, provide:
    - A short title (max 5 words),
    - A brief description (1-2 sentences) summarizing what the theme captures,
    - 1-2 example quotes (copied directly from the dataset) that best illustrate the theme.

Important Instructions:

- Focus on the planning strategies users describe: what they plan to do, under what conditions, and how they will carry it out.
- Identify 5 to 10 major themes that capture the core types of action plans.
- Only include a response in one theme (best-fit).
- Ignore minor spelling or grammar issues, focus on meaning.
- Do not speculate or evaluate user behavior, only extract themes from what is written.

B Prompt for Local LLM Theme Labelling

You are a qualitative researcher assisting with a thematic analysis project. Your task is to read user responses and assign the most appropriate theme(s) from the list below. These responses are written by smokers who are planning to complete activities recommended by conversational agents to prepare for quitting smoking.

IMPORTANT INSTRUCTIONS:

- Focus only on how the user experienced or evaluated the preparatory activities.
- Assign one or more themes from the list below to each response.
- Be strict and conservative in your labeling: only assign a theme if it is clearly supported.
- Do not infer motivations or emotions unless explicitly stated.

Theme Categories and Labels:

- Trigger Type (time-based, event-based, emotion-based, location-based)
- Action Type (physical activity, smoking-related behavior, self-monitoring, mental activity)
- Coping Strategy (distraction, substitution, visualization, relaxation, reflection, social support)
- Specificity (high, medium, low)
- Goal Orientation (toward valued identity, away from negative outcome, neutral/descriptive)
- Mood/State Link (yes, no)
- Monitoring/Tracking (yes, no)

C Themes and Guiding Questions

Theme	**Guiding Question**
Trigger Type	What situation, time, place, feeling, or cue is prompting the planned action?
Action Type	What is the participant planning to do in response to the trigger?
Coping Strategy Type	Is the action a known strategy for managing cravings or preparing for change?
Specificity	How specific and actionable is the plan?
Goal Orientation	Is the plan framed around a positive goal or avoidance of negative outcomes?
Mood/State Link	Does the entry mention a psychological or emotional state tied to action?
Monitoring/Tracking	Is the user planning to track, record, or reflect on behavior?

Table 2: Themes and their Guiding Questions

D Sub-themes and their examples

Theme	**Sub-theme**	**Participant Quote**
Trigger Type	Time-based	*"If tomorrow i have to wait for my lesson..."*
	Event-based	*"If I sit on the couch after a meal..."*
	Emotion-based	*"If I feel guilty..."*
	Location-based	*"If I'm at work..."*
Action Type	Physical Activity	*"... then I'll be able to go out and work out"*
	Self monitoring	*"then i will record my current smoking behaviour."*
	Mental activity	*"then I will visualize smoking as a battle that I win."*
Coping Strategy	Distraction	*"then I will think about activities to keep myself busy when I want to smoke"*
	Substitution	*"then I will eat a gum instead"*
	Relaxation	*"I will learn how to tense and relax areas of my body"*
	Reflection	*"then I will wrote down the importance of being an active person."*
	Social Support	*"If my friend go to play football"*
Specificity	High	*"If I wait ... supermarket this afternoon, then I will lean about ... on dealing with cravings to smoke"*
	Medium	*"If Im done with my chores for the day, Ill think about who I will become once Ill get more physically active"*
	Low	*"If I can then I will"*
Goal Orientation	Toward valued identity	*"... then I will think about who I want to be once I have quit smoking"*
	Away from negative outcome	*"... then i will think about the feared future my kids might see in the future if i fail to become more physically active."*
	Neutral or Descriptive	*"then I will do the activity"*
Mood/State Link	Yes	*"... I will write down my fears about continueing smoking ..."*
	No	*"if I spend more than 4 hours sitting in front of this computer, then I will go out and walk a few minutes"*
Monitoring / Tracking	Yes	*"... I'll write down my current activity routines in the notepad I laid out for that."*
	No	*"then I will go for a walk with him"*

Table 3: Sub-themes and their examples