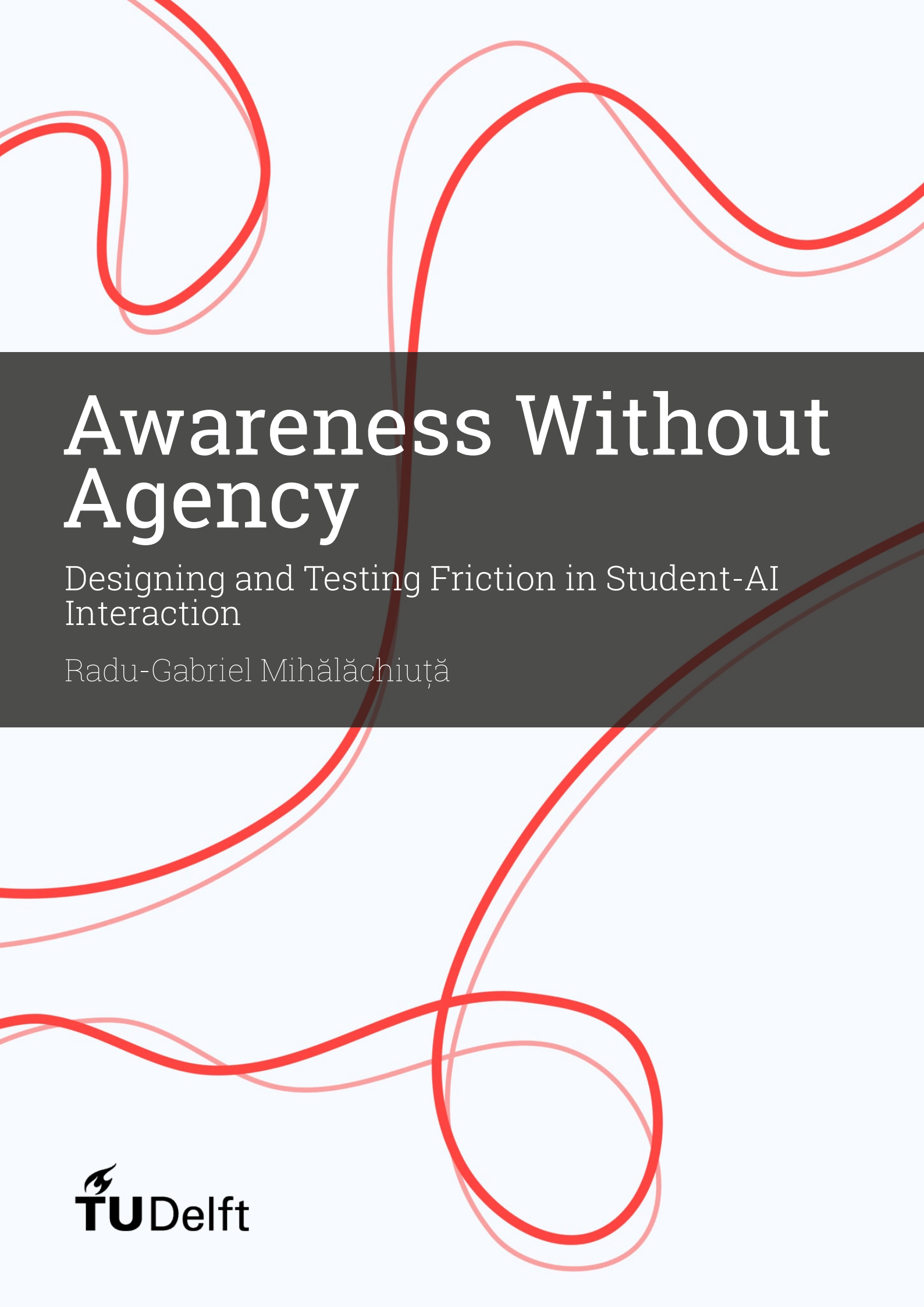


The background of the cover features several thick, vibrant red wavy lines that meander across the white space, creating a sense of movement and fluidity. These lines are layered, with some appearing more prominent than others.

Awareness Without Agency

Designing and Testing Friction in Student-AI Interaction

Radu-Gabriel Mihălăchiuță

This page is intentionally left blank.

Awareness Without Agency

Designing and Testing Friction in Student–AI Interaction

by

Radu-Gabriel Mihălăchiuță

to obtain the degree of

Master of Science

at the Delft University of Technology,
to be defended publicly on Wednesday, August 26, 2026.

Student number: 5302463

Project duration: November 17, 2025 – August 26, 2026

Thesis committee: Dr. Ir. Cynthia C. S. Liem, *TU Delft*, supervisor
Dr. Luciano Cavalcante Siebert, *TU Delft*, committee member

An electronic version of this thesis is available at <http://repository.tudelft.nl/>.



“I do not know much. But there are certain advantages in not knowing. Like virgin territory, the mind is free of pre-conceptions. Everything I do not know forms the greater part of me: this is my largesse. And with this I understand everything. The things I do not know constitute my truth.”

— Clarice Lispector

Too Much of Life: Complete Chronicles [1]

Preface

I always imagined that my Master's thesis would be something that could help people, and I knew from the start that I wanted to choose the topic myself. I recognised something in the people around me, and in myself, and I wanted to understand it better.

What followed was a process full of questions that mostly had no answers. Somewhere along the way I learned to give more attention to the questions than to the search for answers, which is, I suppose, what this thesis turned out to be about. My knowledge grew over these months: about the topic, but mostly about myself, as an engineer and as someone who builds and lives with technology. I like to think that this work closes my academic identity and sets me going for whatever comes next.

Of course, I could not have done it alone. I am grateful to everyone who took part in this messy and rewarding journey.

To my supervisor, Dr. Cynthia C. S. Liem. Cynthia, I have always been inspired by the way you redefined what an engineer can look like. I used to feel like I did not belong, until I saw pieces of myself in you and in your work. Thank you for the support and for the freedom to dive into exploratory research, and for not being frightened by my ambitions. You let me be curious, and that meant the world to me. You also helped me find my focus and my scope when I felt lost.

I want to thank the people who shared their expertise and gave me their time, and I hope this thesis repays some of it. Thank you to Drs. Pleun Hermesen, for helping me understand my own body of work through analogies and terms that made the central idea click. Thank you to Aleksander Buszydlik, for the lunches where I would stress about my methodology and my questions while you offered advice and perspective; you really helped me ground myself and the work. Thank you to Dr. Luciano Cavalcante Siebert, for joining my thesis committee and for advocating for responsible AI in education. Thank you to Dr. Richard Grimes, for the help with the HREC application and the Data Management Plan. And thank you to Estefania Fuentes, for the continuous coaching, and for the ideas and articles you kept sending my way.

Beyond the academic support, I owe my thanks to the many people I keep close to my heart.

To Andreea, for being my friend from day one, for letting me simply be, and for growing alongside me as two immigrants trying to make sense of a life away from home. To Aditi, for always listening and always being honest, for sharing your knowledge and for the support of these past years. As a trio, I am beyond grateful for every story and every experience we have witnessed together.

To Gréta, Miruna, and all my friends from Romania, who I grew up with and who let me be myself. Most of my ways of being come from you, and their shadows are all over this thesis.

To Sofia, for your empathy and ways of seeing the world, always a hug away, always waiting in the kitchen for a conversation.

To Justice, for always being there in these past months, and for the kind of support that does not need to be asked for.

To Yi and Yasi, for keeping me creative and wondering about life's bigger questions.

And to the small worlds I have created during my years in the Netherlands: you have all left a mark on me and helped me change for the better. Your friendship has been a constant source of inspiration, motivation and comfort.

To Diana, the best sister, my light, for navigating this life together.

To Mama Eta, my aunt, for raising me alongside my parents and for caring for me as your own.

And to my parents, Irina and Nicu, for supporting me always, and for believing in my dreams when you never had the chance to follow your own.

Mulțumesc!

— *Radu-Gabriel Mihălăchiuță*
Delft, August 2026

Abstract

Students increasingly delegate reasoning, argumentation and writing to large language models, and these systems are built to make that delegation effortless. The same tools scaffold learning and erode the capacities education exists to develop, which is the structure Plato named *pharmakon*: remedy and poison in one substance, the difference lying in the conditions of use. This thesis asks whether deliberately introducing friction into LLM interaction changes how students work, and under what conditions they accept or abandon it.

Two studies address the question. A survey of ninety university students establishes that reflective engagement collapses situationally. Awareness is not the missing ingredient, since the students who most clearly recognise the model doing their intellectual work are those least able to stop. These findings, with the conditions for reflection drawn from learning theory, yield seven design requirements and a design space of seven friction-based interaction modes.

Three were combined into a single friction condition and tested against a conventional assistant, in a within-subjects experiment ($N = 14$). Friction did not slow students down or reduce what they produced. What changed was its destination, with the share directed to the assistant rather than into the document rising from 39% to 55%. The writing itself showed no structural difference. Participants divided in how they experienced the resistance, and what unified them was a request for control over when it applied.

Friction relocates student effort reliably. What it is worth remains open, and the ambivalence is not a gap in the evidence but the finding itself: the same design, in the same conditions, helped some students and obstructed others. What friction asks of a student is that they stay a little longer with not yet knowing.

Nomenclature

Abbreviations

Abbreviation	Definition
AI	Artificial Intelligence
API	Application Programming Interface
DSR	Design Science Research
LLM	Large Language Model
NCS	Need for Cognition Scale
RLHF	Reinforcement Learning from Human Feedback
SRIS-SF	Self-Reflection and Insight Scale, short form

Terminology

Term	Definition
Bypass	A student's circumvention of a reflective system: declining to consult it, disengaging early, or pushing past a prompt to think further. Treated in this thesis as evidence about the conditions under which resistance is accepted rather than as a usability failure.
Cognitive offloading	Transferring cognitive work to an external system or tool.
Cognitive outsourcing	The delegation of effortful cognitive tasks to an external system without engaging in the underlying work. A subset of offloading, and the problematic one.
Critical reflection	The deliberate examination of the reasoning behind one's own position, prompted by encountering something that does not fit.
Desirable difficulty	A condition that impairs immediate performance while improving later retention and transfer.
Dose	The quantity of friction a design applies. Friction has a lower bound below which nothing is interrupted, and an upper bound above which the interaction stops working.
Epistemic agency	The capacity to construct, evaluate and take ownership of knowledge claims through one's own reasoning.

Term	Definition
Friction	Intentional resistance in an interaction, introduced to redirect attention from the output toward the reasoning that produces it.
Metacognition	Awareness and regulation of one's own thinking. Related to critical reflection but distinct: metacognition can occur without reflective depth, and reflection without explicit monitoring.
Pharmakon	A substance that is simultaneously remedy and poison, the difference lying in the conditions of its use. Applied here to LLMs: the same system can scaffold thinking or replace it, depending on how it is designed and used.
Positive friction	Friction that returns cognitive work to the user, as against procedural friction.
Procedural friction	Resistance that imposes cost without demanding cognitive work: latency, extra steps, or an interface that withholds without purpose.
Productive struggle	Effortful engagement with material at the edge of a learner's current capability, through which learning occurs.
Reflective activation	The initiation of critical reflection during an interaction, as against its sustained continuation. What a single session can observe.
Reflective zombie	A student who produces the surface features of reflection without the underlying process.
Situational regulation	The determination of whether a person reflects by the circumstances they are in rather than by stable traits.
Socratic questioning	Guiding a learner through questions rather than answers, so that a position is examined rather than supplied.
Stated–revealed gap	The discrepancy between what people say they value and what they are observed to do.
Triangulation	Inferring an unobservable construct from the convergence of several observable signals. Here, a claim about reflection is made only where at least two of process behaviour, written output and reported experience agree.

Contents

Preface	ii
Abstract	iv
Nomenclature	v
1 Introduction	1
1.1 Problem and Motivation	1
1.2 Research Questions	3
1.3 Methodology	4
1.4 Research Contributions	5
1.5 Thesis Structure	5
2 Theoretical Framework	6
2.1 The Construct of Reflection	6
2.2 Cognitive Offloading and Epistemic Agency	7
2.3 Friction as design	8
2.4 Sectional Summary	10
3 Related Work	11
3.1 Systems that question instead of answering	11
3.2 Systems that specify how reflection should be produced	12
3.3 Systems that name friction as part of the design	12
3.4 Commercial deployment	12
3.5 What the literature has not established	13
4 The Student Survey	14
4.1 Purpose and positioning	14
4.2 Method	14
4.3 Analysis approach	15
4.4 Findings	16
4.5 Limitations	20
5 Designing for Friction	21
5.1 Design requirements	21
5.2 Institutional alignment	22
5.3 The design space: seven modes of reflective interaction	24
5.4 The tested configuration	25
6 Prototype Implementation	27
6.1 Architecture and model deployment	27
6.2 The writing interface	28

6.3 Data handling	29
7 The Controlled Experiment:	
friction under deadline	31
7.1 Design	31
7.2 Participants	31
7.3 Tasks	32
7.4 Procedure	32
7.5 Measures	34
7.6 Hypotheses	34
7.7 Analysis plan	35
7.8 Ethics	35
8 Results	37
8.1 Sample and design checks	37
8.2 Subjective experience	37
8.3 Process behaviour (H1)	40
8.4 Writing output (H2)	41
8.5 Stated experience against revealed behaviour (H3)	42
8.6 Triangulation	43
8.7 Summary of outcomes	43
9 Discussion	44
9.1 Interpreting the findings	44
9.2 The findings against the framework	45
9.3 The pharmakon	46
9.4 The gaps identified in the literature	46
9.5 Limitations	47
9.6 What would be done differently	47
9.7 Ethical reflections	47
10 Conclusion	49
10.1 What the research questions found	49
10.2 Contributions	50
10.3 Future work	51
10.4 Closing	51
Appendices	57
A Survey Instrument	57
A.1 Section 1 - Consent and background	57
A.2 Section 2 - LLM usage	58
A.3 Section 3 - Relationship with LLMs	59
A.4 Section 4 - Reflection	60
A.5 Section 5 - A different kind of LLM interaction	61
B Interview Protocol	63
C System Prompts	67

D	The Seven Interaction Modes	69
D.1	Candidate prompts	69
E	The Study Interface	72
E.1	Opening screen	72
E.2	Unaided baseline	73
E.3	The writing interface	73
E.4	The friction condition in use	75
E.5	Between and after the rounds	77
F	Session Log Schema	78
G	Linguistic Marker Lexicon	79
H	Supplementary Survey Results	80
H.1	Tests performed	80
H.2	Survey Items	81
I	Use of Generative AI	88

1

Introduction

1.1. Problem and Motivation

Students across the world are increasingly delegating the work of thinking to large language models (LLMs) [2, 3]. They ask Artificial Intelligence (AI) tools to generate their arguments, structure their reasoning, and evaluate their claims. Thus, they bypass the process of *reflection* (examining assumptions and evaluating alternative perspectives) [2, 4]. Receiving an AI-generated argument eliminates the intellectual effort required to construct one, risking what De la Croix and Veen term "reflective zombies": students who mimic the surface features of reflection without its substance [5].

In contrast to concerns, there have been many arguments made about opportunities offered by AI tools. The benefits can be, for many learners, transformative. LLMs offer individualized explanation at a scale no human tutoring system can match [6]; they lower barriers for students working in a second language [7], students with disabilities [8], and students without access to academic support networks; they make it possible to receive immediate, iterative feedback on drafts, code, and arguments at any hour [9, 10]. Educational platforms are already building on these affordances, and early classroom integrations suggest that when LLMs scaffold instead of substitute, they can enrich academic practice [11].

This frames a picture about a powerful tool that took over an educational system that was never designed to distinguish between output and thinking itself. And the tool is here to stay, since students *will* use these systems, and prohibition does not change that. Banning generative AI pushes its use out of sight, and an institution that cannot see how students work with these tools has no position from which to teach them to work with them well. What academic use must do is sustain learning. This is not an argument that the current arrangement is permanent. Which providers dominate, what they charge, and whether the capable models remain commercial are all open questions.

On the other hand, the industry acknowledges this gap, as in early 2025, OpenAI, Anthropic, and Google each launched educational features designed to guide students toward answers through questioning rather than providing them directly. These are significant commercial investments in friction-based interaction, and they constitute an implicit admission that default

LLM interaction is not suited to educational contexts. Yet none of these products has been accompanied by peer-reviewed evidence that they produce the reflective outcomes they claim; as OpenAI acknowledged at launch, “the research landscape is still taking shape.” [12].

Moreover, there are multiple detrimental effects that can be concretely measured. Firstly, these systems have great environmental impacts that become more and more visible [13]. Then, using neuroimaging methods, it was demonstrated that passive interaction with LLMs is associated with reduced attentional engagement and measurable declines in sustained cognitive effort [14]. Students using LLMs for inquiry-based tasks reported lower cognitive effort and produced weaker arguments than peers who did not [15]. Gerlich observed a negative correlation between frequent AI use and critical-thinking performance, particularly among younger users [2], and repeated use of generative AI as a substitute for effortful thinking lowers student initiative and epistemic agency over time [4]. Many studies are still in progress and yet to come, but we are observing a decline in precisely what education is designed to cultivate: *cognition and critical thinking*.

Current LLM services optimize for output quality and user satisfaction: they minimize the discomfort of not knowing and resolve ambiguity quickly [16, 17], working against the very conditions that learning research identifies as necessary [18, 19]. Educational institutions add to this by rewarding correct outputs, ignoring the reasoning that produced them [20]. Reflection then disappears not because students stop valuing it, but because nothing in the system detects whether it happened. What must be recovered is the *desire to ask* and the willingness to *remain in friction*.

There is a paradox at the center of this situation. Many students fear a future in which AI replaces human labor and judgment. What fewer recognize is that by delegating their reasoning to systems trained on statistical patterns (i.e., systems with no capacity for genuine understanding, feeling, or contextual judgment) they are accelerating the very substitution they fear. We risk becoming passive consumers of decisions made elsewhere, by systems whose purposes we cannot fully see: becoming, in a sense, the robots we fear, not through force but through habit. As Hubert Dreyfus warned, “our risk is not the advent of super-intelligent computers, but of subintelligent human beings.” [21].

Without the light of critical thinking, of deep and slow reading, of genuine introspection; without the light of metacognition and the self-awareness that comes with it, we will gradually and irrevocably lose our ability for empathy, compassion, solidarity, lose our ability to remain fully human.

— Elif Shafak [22]

How can the same technology be scaffold and threat at once? Plato had a word for it. When the god Theuth offered writing to King Thamus as “a remedy for memory and wisdom” the king refused the framing: it would produce forgetfulness, and the appearance of wisdom rather than its reality. Plato’s word for the gift, **pharmakon**, means remedy and poison in one substance, and Derrida showed the ambivalence is irreducible [23]. LLMs are a pharmakon in this sense: if the same system can enable and undermine the capacity it serves, the question is not *whether* LLMs help or harm, but *under which conditions of design and use the remedy*

prevails. This represents the main motivation of the thesis.

1.2. Research Questions

Organized around three main research questions, this thesis examines whether deliberate interaction friction can redirect students from output-seeking toward reflection.

In this context, *reflection* means the deliberate examination of the reasoning behind one's own position, prompted by encountering something that does not fit. *Friction* refers to small deliberate obstacles inserted at the point where an automatic behaviour would otherwise proceed unexamined. Chapter 2 further develops the definitions and their sources.

RQ1 (diagnostic): How do university students understand and experience reflection in the context of LLM use?

RQ1a. Which conditions do students report as enabling or collapsing independent thinking when working with LLMs?

RQ1b. How do students' stated values about reflection relate to their reported patterns of LLM use?

RQ2 (design): Which friction-based interaction strategies can be operationalized in LLM system design to redirect student attention from output efficiency toward the reasoning process?

RQ2a. How does a friction-based assistant, compared with a default assistant, change how students work during writing under deadline pressure?

RQ3 (reception): How do students experience friction-based interaction, and under what conditions do they accept or bypass it?

Table 1.1 shows how each question is answered.

	Method	Reported in
RQ1	Survey ($n = 90$), closed and open items	Chapter 4
RQ1a	Survey, situational multi-select items	Chapter 4
RQ1b	Survey, discrepancy between value and behaviour items within one instrument	Chapter 4
RQ2	Design framework: requirements, seven modes, one instantiated configuration	Chapters 5, 6
RQ2a	Within-subjects experiment ($N = 14$): process metrics, written output, subjective ratings	Chapters 7, 8
RQ3	Experiment: disengagement measures and self-report	Chapters 7, 8

Table 1.1: Research questions, the methods used to answer them, and where each is reported.

1.3. Methodology

The research is structured as Design Science Research, which is concerned with the creation and evaluation of artefacts that address identified problems [24]. The choice follows from the object of study: the question cannot be answered without building something, and what is built must be evaluated. Within that structure the work was *exploratory*, in that the question was not fixed at the outset but narrowed through successive engagement with theory, making and evidence. This is why it aims to generate initial empirical evidence and grounded design principles more than causal proof. And the evidence itself is *mixed*, combining quantitative with qualitative data and triangulating several types in support of a single interpretive claim.

Using Hevner’s three-cycle model [25], the *rigor cycle* grounds the work in desirable difficulty and positive friction to deliver a design space and findings. The *relevance cycle* leverages a survey of 90 students regarding reflection drop-offs to test the artefact in context. Between them, the *design cycle* iterated three times: translating insights into seven interaction modes (Appendix D), instantiating them in a prototype, and refining the system after piloting.

Table 1.2 maps the six DSR process steps onto what each meant in practice and where each is reported.

Process step	Translation	Reported in
Problem identification	Establishing that students delegate reasoning to LLMs, that the conditions for this are situational, and that existing reflective systems are proposed without being compared	Chapters 1, 2, 3, 4
Objectives definition	Seven design requirements derived from the enabling conditions for reflection and from what the survey population reported wanting	Chapter 5
Design and development	Seven design requirements turned into seven interaction modes, four of which were instantiated as a working system	Chapters 5, 6
Demonstration	Two formative pilots under the conditions the experiment would use, followed by deployment with fourteen participants	Chapters 6, 7
Evaluation	A within-subjects comparison against a default assistant, with process, output and self-report evidence triangulated and interviews analysed thematically	Chapters 8, 9
Communication	This thesis, and the design space it documents for subsequent work to build on	—

Table 1.2: The Design Science Research process steps as realised in this thesis.

Additionally, most of the experimentation involves human subjects, which required a complete risk mitigation plan and a data management plan (DMP). We have obtained an approval from the TU Delft Human Research Ethics Committee (HREC) to perform such experiments. An example of an invitation and consent form can be found in Appendix H.2.9.

1.4. Research Contributions

This thesis makes five contributions to the emerging field of friction-based educational AI design.

1. A survey of ninety students on how they experience reflection and when they abandon it, used to inform the design of future LLM systems.
2. A design space with seven friction modes, each isolating one mechanism, for subsequent work to build on or adapt.
3. An empirical comparison that measures what changes when the same students, doing the same task under the same deadline, meet an assistant that resists them.
4. Treating bypass as data, not failure, and locating where it happens.
5. Methodological knowledge about future empirical studies.

1.5. Thesis Structure

Chapter 2 reviews related work, situating this thesis within the landscape of existing reflective AI systems and identifying the empirical and conceptual gap it fills. Chapter 3 develops the theoretical framework grounding this research. Chapter 4 reports the survey ($n = 90$): its purpose, method, findings, and their traceability into the design and the experiment. Chapter 5 presents the design framework derived from theory and literature, while Chapter 6 describes the prototype implementation. Chapter 7 specifies the controlled experiment: design, tasks, procedure, measures, hypotheses, and ethics. Chapters 8 and 9 present results and discussion. Chapter 10 concludes with contributions, implications, and directions for future research.

2

Theoretical Framework

This chapter establishes why reflection is worth protecting, narrows a broad concept to the specific form this thesis targets. It then shows how default interaction with large language models threatens that form of reflection. Finally it introduces friction as the design response and sets out the conditions under which resistance produces reflection more than frustration.

2.1. The Construct of Reflection

2.1.1. Defining reflection

Defining reflection is contested, but one point is consistent across perspectives: **reflection is triggered by a discontinuity, a mismatch, a friction point**. It is a disruptive cognitive–affective process, activated when expectations or habitual actions conflict with experience. It is not fact-checking, recall, or output optimization; it does not happen continuously; it involves discomfort. Hermesen et al. define it as the process through which people make sense of and interpret an experience in order to understand where they stand and how to proceed [26].

The foundational condition for self-reflection is metacognition: the ability to monitor, control, and strategically adapt one’s own thinking [27]. Closely related is critical thinking: a context-sensitive practice of evaluating evidence, questioning assumptions, weighing perspectives, and judging under uncertainty [16]. The educational story begins with Dewey [28], who defined reflection as the active, persistent, and careful consideration of any belief in the light of the grounds that support it.

2.1.2. The working definition

We look at reflection from two distinct positions. In *direction*, **self-regulatory reflection** [29, 30] is inward-facing monitoring of one’s own learning (“do I understand?”) while **critical reflective engagement** is outward-facing and epistemological, asking not “how am I thinking?” but “what am I taking for granted, and why?” [31]. In *depth*, Argyris distinguishes single-loop learning (i.e., corrects strategies to achieve a specific goal), from double-loop learning (i.e., questions and critiques the actions themselves) [32]. This thesis adopts a more critical, double-loop perspective over reflection.

Brookfield explains what “examining an assumption” actually involves, and why it is difficult [33]. Assumptions differ in how accessible they are to inspection. *Causal* assumptions (beliefs about what produces what) are the easiest to surface and the most frequently examined. *Prescriptive* assumptions concern what ought to happen, and are examined less often because they feel like values rather than claims. *Paradigmatic* assumptions are the structuring categories through which a situation is understood at all. Brookfield’s point is that an assumption cannot be examined by introspection alone, and it requires encountering it from a different position. Hence, critical reflection is not a matter of thinking harder and deeper, but rather being put face to face with a mismatch.

Working definition: *critical reflection* is the deliberate examination of the reasoning behind one’s own position, prompted by encountering a mismatch (e.g., an assumption one had not noticed, a perspective one had not considered, or a question one cannot immediately answer).

The next section examines the technology that removes this trigger.

2.2. Cognitive Offloading and Epistemic Agency

Cognitive offloading is the use of external action to alter a task’s information-processing demands [34]. It is often rational, and the decision to offload is made under effort pressure more than reflective evaluation: people systematically discount future cognitive effort and take the path of least resistance even when they would, on reflection, prefer otherwise [35].

Large language models open a new class of this offloading. It goes further than memory or calculation, and includes reasoning, argumentation and writing itself. As probabilistic text generators aligned through instruction tuning and Reinforcement Learning from Human Feedback to produce what human raters prefer, they are optimised toward fluent, confident, complete-feeling output. This gives users the impression of understanding without the effort that produces it.

2.2.1. Three mechanisms of erosion

Vendrell and Johnston identify three mechanisms through which unstructured LLM use undermines critical thinking [16].

Unexamined delegation is the handing over of reasoning, synthesis and argument construction without doing the underlying work. Stadler et al. found LLM-assisted inquiry cognitively easier but analytically weaker [15], and another work found neurologically reduced attentional engagement in ChatGPT-assisted writing, accumulating what they term *cognitive debt*: performance gains that mask reduced engagement and are repaid later as weaker independent skill [14].

Metacognitive disengagement is the acceptance of fluent output without evaluating it. Fluency exploits the fluency heuristic, in which ease of processing is mistaken for truth [36], and language models reproduce the social signals of credibility convincingly enough to trigger it. Lee et al. found that high confidence in AI output suppresses critical evaluation among knowledge workers [37], and Gerlich observed a negative correlation between frequent AI use and critical-thinking performance, strongest among younger users [2].

Epistemic narrowing is reduced exposure to contested and ambiguous knowledge, as systems optimised for convergence return singular answers. Darvishi et al. found that repeated substitution erodes student initiative and epistemic agency [4], and Shaw et al. characterise this phenomenon as *cognitive surrender* [38].

Read against the working definition, the three attack it at each of its parts. Delegation removes the examination, disengagement removes the doubt that would activate it, and narrowing removes the encounter with something that does not fit.

2.2.2. A contested literature, and the gap

In a meta-analysis of experimental studies, effects varying by task were found: AI positioned as a complement is associated with gains, AI as a substitute with harm [10]. Therefore, LLMs do not directly harm learning, but *how* they are used is consequential. Students report the ambivalence themselves, stating integrity concerns while using these tools heavily [17], and educators observe learning-productive struggle being eliminated [18]. This positioning echoes the idea of the *pharmakon* presented in the Introduction.

Few studies measure the distance between students' stated attitudes and their observed behaviour in realistic settings. And the conditions under which students bypass reflection (fatigue, deadline pressure, disinterest) are rarely treated as central variables rather than as noise.

2.3. Friction as design

Reflective design argues that systems ought to surface rather than conceal the assumptions embedded in their use. Within Human-Computer Interaction this has been developed as *design friction*: small deliberate obstacles inserted at the point where an automatic behaviour would otherwise proceed unexamined, with the obstacle serving to make the behaviour a choice [39, 40].

2.3.1. Effort as mechanism

Bjork's theory of *desirable difficulties* holds that conditions which feel hard and appear to slow learning improve retention and transfer over time, while conditions that feel easy produce superficial encoding [19, 18, 41].

Cognitive Load Theory explains why by distinguishing what the effort is spent on [42]. Some load is intrinsic to the material and unavoidable. Some is extraneous, imposed by how the material is presented. *Germane* load is the effort spent constructing understanding, and it is the work in which learning consists. Difficulty that makes the learner do the task's cognitive work is *germane*, and difficulty that stands between the learner and the task is not.

2.3.2. Productive and procedural friction

Productive friction returns cognitive work to the user: a question that cannot be answered without examining one's reasoning, an assumption named, a request for elaboration that forces a position to be made explicit.

Procedural friction stands between the user and the task without demanding anything: latency, extra steps, an interface that withholds without purpose.

When designing a friction-based system we need to focus on productive, and not procedural, friction. The system must generate resistance through cognitive challenge instead of obstruction.

2.3.3. The dose problem

Inan et al. provide the systematic treatment for conversational AI, defining positive friction as intentional movements that slow an interaction in order to yield long-term benefit, organised into five categories: assumption reveal, reflective pause, overspecification, probing and reinforcement [43]. Conversations incorporating friction outperformed frictionless ones on task success and on modelling users' mental states. Interestingly, combining all five at once *decreased* effectiveness.

Friction therefore has a **dose**. It is not a property a design either has or lacks, but a quantity with a lower bound below which nothing is interrupted and an upper bound above which the interaction stops working. Neither bound is specified in the literature, and locating them is an empirical question.

2.3.4. The willingness problem

A difficulty that is desirable in principle produces nothing unless learners take it up and stay with it. De Bruin et al. address this in the Start and Stick to Desirable Difficulties framework (S2D2), asking why learners so often decline effortful strategies whose benefits are well established [44].

Part of the answer is that learners misjudge difficulty: effortful conditions are rated as less effective than they are and easy ones as more [19, 41].

If a design deliberately reduces fluency, user satisfaction is a measure that theory predicts will move the wrong way. Subjective measures remain good evidence about acceptance, about whether resistance was experienced as productive or procedural, and about whether a student would return to the system, but the behavioural and process evidence must carry the question of effectiveness.

2.3.5. The conditions friction depends on

Friction alone is insufficient. Without enabling conditions, resistance produces frustration. Mezirow's transformative learning [31] supplies the end that friction serves, and together with reflective practice research [45], the psychology of metacognition [30, 29] and the frameworks above, five conditions can be identified. Each is stated with what it implies for design, since these implications become the design requirements of Chapter 5.

Dissonance. There must be something to reflect *on*: the interaction has to present something that genuinely resists the framework the student arrives with. Brookfield's account explains why this cannot be left to the student alone: paradigmatic assumptions in particular are not visible from inside the position that rests on them. *Design implication:* the resistance must carry content, not delay.

Germane demand. The effort the design imposes must be spent on the task and not on the interface. *Design implication:* resistance must take the form of cognitive challenge rather than

procedural obstruction (Section 2.3.2), because procedural friction is avoided, not endured.

Safety. Reflection requires admitting that one does not know, which people do not do when they feel assessed. *Design implication:* the system must not praise, grade or otherwise position itself as an authority whose approval is at stake.

Capacity. Reflection consumes working memory, sustained attention and tolerance for ambiguity, all finite and variable. *Design implication:* the same design behaves differently on the same student under different conditions, which makes the conditions of use part of the design problem.

Willingness. Following S2D2, the student must be prepared to begin the effortful engagement and to stay with it. *Design implication:* how the difficulty is experienced is a design concern and not only an outcome, since a difficulty that is refused produces nothing regardless of how well it is calibrated.

The literature specifies them but does not supply evidence about where the boundaries lie in an actual interaction, which of the conditions binds first when they conflict, or what happens when a student who is offered a difficulty declines it. Those are the questions the empirical chapters address.

2.4. Sectional Summary

Critical reflection is triggered by encountering something that does not fit. Default LLM interaction is optimised to remove exactly such encounters. Friction is the design response. But friction is a dose, it depends on conditions the literature specifies only in principle, and it can be refused in a way that embedded difficulty cannot. This can be visualised in Figure 2.1 below.

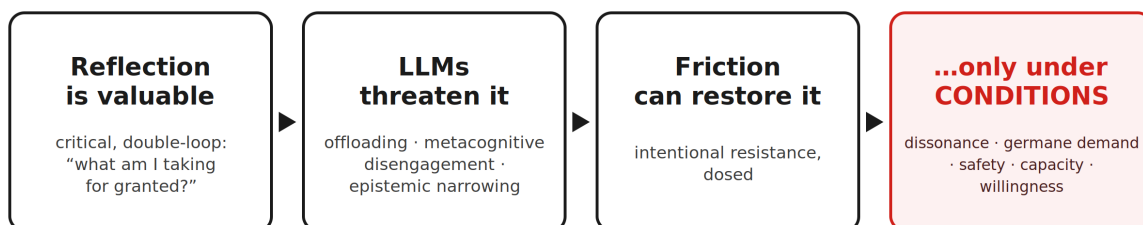


Figure 2.1: The argument of this chapter.

3

Related Work

This chapter situates the present research within existing work on reflective AI in education, together with the commercial products built on the same premise. The systems are grouped by the design move they make. Each is examined for what it offers and what it leaves open. The chapter closes by identifying the gaps that remain.

3.1. Systems that question instead of answering

Three projects act as examples of educational systems that create deliberate friction in educational context:

KRITIKOS, developed at HU University of Applied Sciences Utrecht, is a digital reflection partner that asks reflection questions with the aim of helping students use AI more consciously.¹ Its reported outcomes include a working prototype tested with students and teachers, insight into how students use and reflect on AI, and publications on value-sensitive design and AI literacy. It demonstrates that a questioning assistant can be deployed in a higher-education setting.

Its object is reflection on AI *use* (i.e., whether the tool is needed for a task) and not reflection on the reasoning behind a student's own position, so it leaves unaddressed whether a resistant assistant can affect how a student thinks about a substantial question. And its evaluation is formative as it establishes feasibility and acceptance, not effect.

The Critical Thinking Bot is a publicly available prototype rather than a published system.² It invites users to state a single claim and engages in structured dialogue aimed at making that claim clear and defensible, operating through questioning. It treats reflective interaction as something that can take different forms.

As a prototype it carries no evaluation, no comparison condition and no measurement framework. It is a design position advanced without evidence, and there is no basis on which to judge whether the interaction it proposes has any effect on the users who meet it.

¹<https://www.internationalhu.com/research/projects/kritikos>

²<https://www.criticalthinkingbot.com/>

DUCK is an AI teaching assistant developed for CS50, designed to resist giving direct answers to programming problems and instead ask clarifying questions that guide students toward solutions [11, 46]. Its designers have published reflective accounts of their design decisions.

Still, DUCK is domain-specific (i.e., coding) and the dialogue can always be steered toward a correct resolution. DUCK has also not been evaluated against reflective outcomes.

3.2. Systems that specify how reflection should be produced

SIMBA is a tool for designing generative AI agents for reflective learning, letting instructors define the learning objectives, question sequences and conversational tone an agent must follow [47]. It establishes that reflective interaction can be treated as a space of configurable behaviours rather than a single system.

Students, who will be the ones who interact with this agent, have no part in configuring its behavior. Additionally, the work does not report what they make of the configurations chosen for them.

Lewis et al. propose a four-tier architecture for reflective AI agents, arguing that current architectures do not incorporate reflection as a core component [48]. Reflection could happen in the model itself or in the interaction itself. However, this distinction is conceptual, and the architecture has not been implemented or evaluated, so which layer produces effects on users is an open question.

3.3. Systems that name friction as part of the design

Onco-Shikshak, a health-sciences learning platform, is one of the few systems to name *desirable friction* as a design principle. It operationalises the principle through three mechanisms in combination: *confidence elicitation*, in which learners state how certain they are before receiving feedback; *Socratic questioning*; and *cognitive conflict*, in which learners are presented with information that contradicts what they have just asserted [49]. It demonstrates that friction mechanisms can be combined deliberately. Even so, it operates in a training context with defined learning objectives.

Evidence that reflection prompts change what students write comes from two studies. The **Reflect-and-Act chatbot** in teacher professional development [50] and **MindMate** in tertiary business education [51] both show that agent-based prompts produce more structured written reflections than static prompts do.

Khanmigo is the most widely deployed Socratic tutoring system, using guided questioning across mathematics and science in thousands of schools [52]. It establishes that question-led interaction can be deployed at scale and accepted by students. It operates in K–12 contexts on problems with defined answers, does not address epistemic agency, and has no published peer-reviewed evaluation of its reflective outcomes.

3.4. Commercial deployment

In 2025 the three major model providers (Google, OpenAI, Anthropic) each launched an educational mode built on the same premise as the systems above. OpenAI's Study Mode

modifies ChatGPT's interaction style to guide students toward answers through step-by-step questioning.[53].³ Anthropic's Learning Mode, part of Claude for Education, employs Socratic questioning as its primary mechanism, returning "how would you approach this problem?" in place of a solution.⁴ Google's Guided Learning in Gemini is built on LearnLM models fine-tuned on educational data, with multimodal support; Google has published technical reports on LearnLM but no controlled studies of reflective outcomes in higher education.⁵

All three are prompt- or fine-tuning-level interventions layered on general-purpose models, and none has been accompanied by peer-reviewed evidence that the interaction style produces the reflective outcomes claimed for it.

3.5. What the literature has not established

The systems above forget to include an important element in designing such solutions: the students' perspectives. These design decisions are made almost entirely from the perspective of educators and developers. What students themselves experience when they meet a questioning assistant, what they find useful and what obstructive, and what they would want changed, is largely missing [20, 43]. Moreover, it is argued that evaluating educational AI by learning outcomes alone leaves out the dimensions on which these systems act: cognition, agency, emotion and ethics [54].

The literature treats friction as a positive design variable and acknowledges in principle that it is productive only until difficulty tips into overload. Where that transition occurs, in an actual interaction with actual students, has not been established empirically. Nor is there evidence about what happens at the point of refusal: when students decline a resistant system, under what conditions, and what their refusal indicates about the design.

Reflective designs are proposed and deployed, but where evaluation exists it is formative, establishing that a system can be built and is accepted, and not comparative.

These three gaps define the directions the remainder of this thesis addresses.

³<https://openai.com/index/chatgpt-study-mode/>

⁴<https://www.anthropic.com/news/introducing-claude-for-education>

⁵<https://blog.google/products-and-platforms/products/education/guided-learning/>

4

The Student Survey

4.1. Purpose and positioning

The survey ($n = 90$) answers RQ1 and its subquestions: “*How do university students understand and experience reflection in the context of LLM use?*” (**RQ1**), “*Which conditions do students report as enabling or collapsing independent thinking when working with LLMs?*” (**RQ1a**), “*How do students’ stated values about reflection relate to their reported patterns of LLM use?*” (**RQ1b**).

The survey helped us establish how students interact with LLM systems and how they define reflection for themselves; identify what predicts cognitive offloading; and elicit design preferences and prior beliefs about friction-based interaction before anyone has experienced it. Each function produced a result that directly shaped the design requirements of Chapter 5 or the experimental protocol of Chapter 7.

4.2. Method

4.2.1. Participants and procedure

Participants were university students (Bachelor or Master), 18 or older. Students in a direct grading relationship with the researcher were excluded as much as possible. For this, recruitment ran through university mailing lists and student networks; participation was voluntary and uncompensated. In addition, the participants remain anonymous, so identification is not possible. The final sample ($n = 90$) is drawn predominantly from technical programmes.

4.2.2. Instrument

The survey comprises twenty questions in four sections. *Usage profile* covers frequency, purposes and modality of LLM use, and provides the basis for stratifying analysis by usage intensity. *Relationship with LLMs* covers emotional associations and epistemic-tension items, capturing the affective register that Chapter 2 argues is construct-relevant. *Reflection beliefs and practices* establishes how students understand reflection. *Friction tolerance and hypothetical scenarios* records first reactions to a resistant system and the situations in which students expect to bypass one. The complete survey, with its questions, is shown in Appendix A.

No item was mandatory. Participants were told at consent that they could skip any question, and this was a deliberate choice: a survey asking about guilt, dependence and perceived loss of capability would have produced compliance rather than honesty. Item bases are therefore reported individually and vary between 88 and 90 for closed items.

Two validated instruments are embedded in the third section. The Need for Cognition scale measures the tendency to seek out and enjoy effortful thinking [55]. The Self-Reflection and Insight Scale, short form, measures the disposition to examine and understand one's own thinking [56]. The two scales are well-validated measures of their constructs, and what is tested here is whether they predict LLM-use behaviour in this sample, not whether they measure what they claim to.

Custom items were written where no validated instrument exists: the student–AI relationship, emotional associations, and prior beliefs about friction. Following Rattray and Jones, these were developed deductively from the theoretical framework, piloted for clarity with a small group of students, and are interpreted within the triangulated design rather than as standalone measures [57].

4.2.3. Ethics and data handling

The survey was approved by the TU Delft Human Research Ethics Committee (HREC), and data handling follows the project's Data Management Plan (DMP), reviewed by a faculty data steward before collection began.

Consent was obtained through an opening statement describing the purpose of the research. The statement also confirmed that no academic grades, performance records or psychological diagnostic measures were compiled.

The survey collects no personal data. Responses were exported from Microsoft Forms under the institutional licence and stored on TU Delft Project Data Storage, accessible only to the researcher and supervisor.

Two open-text items nonetheless create the risk of disclosure. In these questions, participants could volunteer identifying details without being asked to, and the Data Management Plan records this possibility. All free-text responses were therefore read and pseudonymised before analysis, with any identifying reference removed.

4.3. Analysis approach

Analyses were run in Python (`pandas`, `scipy.stats`). Responses were prepared for analysis as follows: multi-select questions were decomposed into binary indicators plus a count of options selected; Likert items were mapped to integers; scale items were averaged after reverse-scoring negatively phrased items; and the eight discipline categories were collapsed into three groups, because several original cells held fewer than five respondents.

Four analysis types carry the findings:

1. **Bivariate associations** use Spearman's rank correlation (ρ), chosen over Pearson because Likert items are ordinal and several distributions departed from normality: ρ asks whether higher standing on one variable tends to accompany higher standing on the

other, using ranks rather than raw values.

2. **Group comparisons** use *t*-tests for two groups and chi-square for categorical variables.
3. **Multiple regression** (OLS) tests whether the two trait scales still predict bypass once a single situational item is controlled: all three predictors enter simultaneously, and standardised coefficients (β) express what each contributes that the others do not.
4. **Cross-tabulations** sort respondents by their answer to one question and compare outcomes across the resulting groups

Significance is reported at $\alpha = .05$ with exact *p*-values. Percentages are given to the nearest whole number and means to two decimals. Null results are reported as nulls, and the discipline comparison is treated as power-limited given its smallest cell of twelve. Missing responses are handled by pairwise deletion, so each analysis uses every case with data on the variables it involves and reports its own *n*.

Three items produced free text: a request to describe a moment in which a selected emotion was felt, an open definition of reflection, and an account of choosing speed over depth. All three were coded inductively, with categories emerging from the responses rather than imposed on them. They are used to interpret the quantitative patterns they accompany, not as findings in their own right. The full analysis, including every test performed, is given in Appendix H.

4.4. Findings

4.4.1. Students' definition of reflection

Asked in their own words what reflection means in the context of their studies, respondents converge on active cognitive engagement more than record-keeping. Half include "think" or a close variant, and the recurring terms (questioning, evaluating, learning, critical thinking) describe an operation performed on one's own reasoning rather than a summary of what was done.

"To me, reflection means challenging my previous assumptions or stances whenever I learn something new."

"Not only consuming information, but trying to critically think about it, give it a second thought, trying to connect it with the knowledge that already exists in my head."

The population's concept aligns with the working definition of Chapter 2: students describe examining and questioning rather than monitoring their comprehension.

A small number of responses describe reflection through the institutional forms by which it is assessed (one defines it as "did I get the best grade I could") which is the reflective-zombie concept itself.

4.4.2. Students value reflection

The sample reports high reflective disposition (SRIS-SF mean 4.64/6) and above-average enjoyment of effortful cognition (NCS mean 3.89/5). A majority (52%) chose "a muscle: it develops with practice and weakens without use" as the metaphor best describing reflection. Thus, students intuitively grasp that the capacity can atrophy. And the desire for friction is

stated clearly: 54% would prefer a reflective LLM outright, rising to 88% when “might prefer” is included, the rest choosing “no” to the adoption.

4.4.3. Paradoxical usage

The same sample reports substantial unease, as it can be seen in Table 4.1. Guilt is the most common emotional association with academic LLM use, followed by dependence. Nonetheless, there are also reports of more positive emotions like confidence, relief and motivation. This showcases the framing of LLMs as a *pharmakon*.

Emotion	<i>n</i>	%
Guilt	46	51
Dependence	38	42
Curiosity	35	39
Confidence	30	33
Relief	26	29
Motivation	24	27
Detachment	16	18
Neutral	14	16
Anxiety	11	12

Table 4.1: Emotions students report when using LLMs for academic work ($n = 89$ answering, multiple selection).

Forty percent acknowledge that the model does most or all of the intellectual work in a helpful exchange. More than half see using LLMs as cheating and 70% feel less capable to construct their ideas after relying on such systems. However, the paradox lies in the fact that 81% use LLMs daily or several times a week.

Asked to rank five statements from closest to furthest from their own experience, respondents put speed first. “It helps me finish the task faster” was ranked first by 59% against 25% for “it helps me think through a problem”. Hence, students report using these systems to save time, but then they would also consider a system designed to slow them down.

4.4.4. What the emotions attach to

Respondents were asked to describe a specific moment in which they had felt the emotions they selected from Table 4.1.

The described moments cluster around the same conditions the closed items identify as conditions in which students might offload their thinking, so the affective and behavioural evidence point to the same explanation. But the discomfort is largely *retrospective*. It attaches to work already submitted rather than to the moment of deciding to use the tool, which is consistent with the later finding that guilt does not predict bypass.

The accounts also show relief and guilt attached to the same behaviour:

“Relief: in a situation when I was struggling to debug a script and an LLM easily showed me what was wrong, saving me a lot of time.”

“Guilt, because I should’ve started the assignment earlier to use less of the LLM.”

“When someone sees me using an LLM, I feel like I am ‘cheating’ or not really studying myself.”

4.4.5. Bypass and Habituation

Across the two bypass questions, the same four conditions dominate, as it can be observed in Figure 4.1.

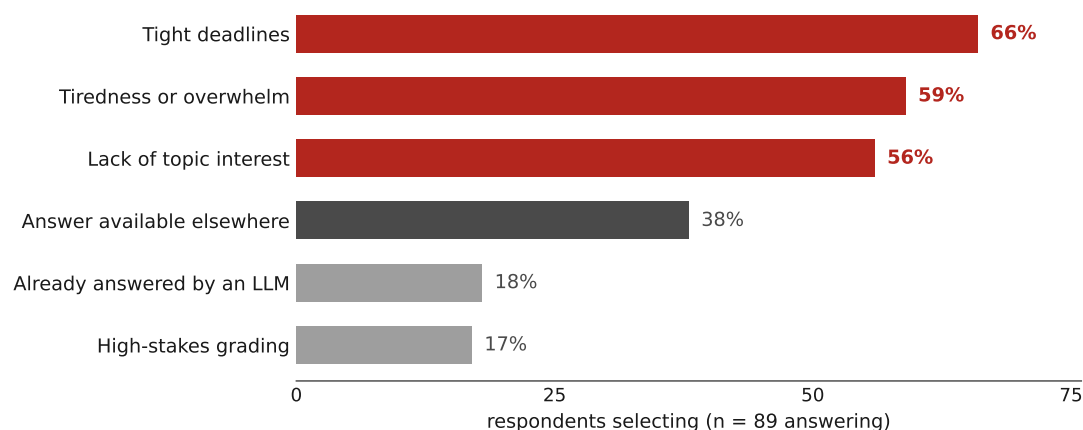


Figure 4.1: Bypass tracks state rather than stakes.

The more students use the system, the easier it becomes to bypass thinking with less remorse. Table H.9 shows that guilt is non-monotonic in usage frequency: it rises from 50% among rare users to 62% at weekly use, then drops to 40% among daily users. Open-text responses suggest active rationalisation: daily users articulate justifications (“when deeper thinking is not helpful,” “it matters more to get the task done”) where several-times-a-week users describe offloading with regret.

4.4.6. Traits predict what students say, not what they would do

We were interested in understanding what item could predict the *situational bypass count*, which is the number of conditions under which a respondent would bypass a reflective LLM. This would help guide our focus to a concrete characteristic that can be linked to offloading moments. The validated scales predict self-report. Both correlate with reported pushback against LLM responses (NCS $\rho = .28$, $p = .008$; SRIS $\rho = .31$, $p = .004$). Neither predicts the bypass count (NCS $\rho = -.06$, $p = .562$; SRIS $\rho = .02$, $p = .832$).

However, when we applied a multiple regression over the scales with the “I find it hard when to stop” (hard-to-stop) item, we observed the item correlating more with the situational bypass count. The exact numbers can be seen in Figure 4.2.

The pattern extends to the friction preference itself. Respondents who said “yes” to a reflective LLM select bypass situations at almost exactly the sample average (2.45 against 2.55 of a possible six). Wanting the system and expecting to bypass it are fully compatible. For design,

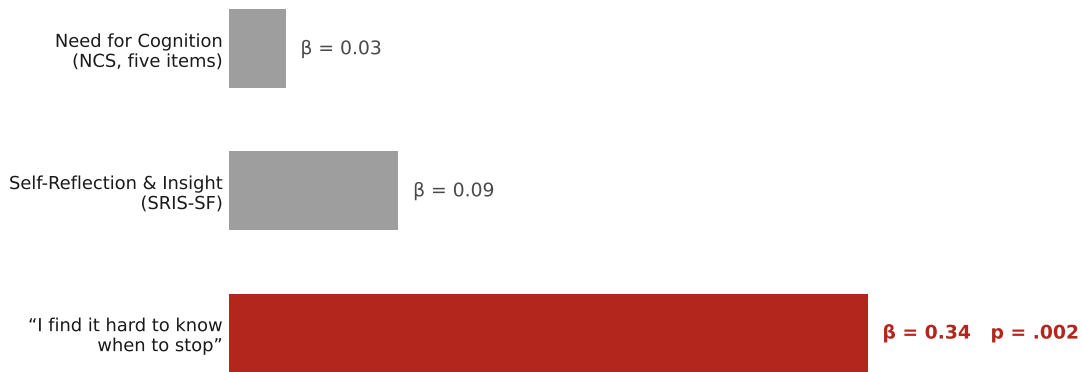


Figure 4.2: hard-to-stop item outperforms both scales (n=89).

the students who want it are the same students who expect to abandon it under pressure, so a reflective system cannot be built on the assumption that its users will stay.

4.4.7. Awareness without agency

Cross-tabulating perceived contribution (“how much did you contribute to a helpful LLM answer?”) against hard-to-stop and guilt yields an interesting finding (Figure 4.3). Respondents who report shaping answers carefully show low hard-to-stop scores (1.68/5) and lower guilt (37%); those who acknowledge the model did most or nearly all of the work show progressively higher hard-to-stop scores (2.69 and 3.08) and higher guilt (60% and 62%).

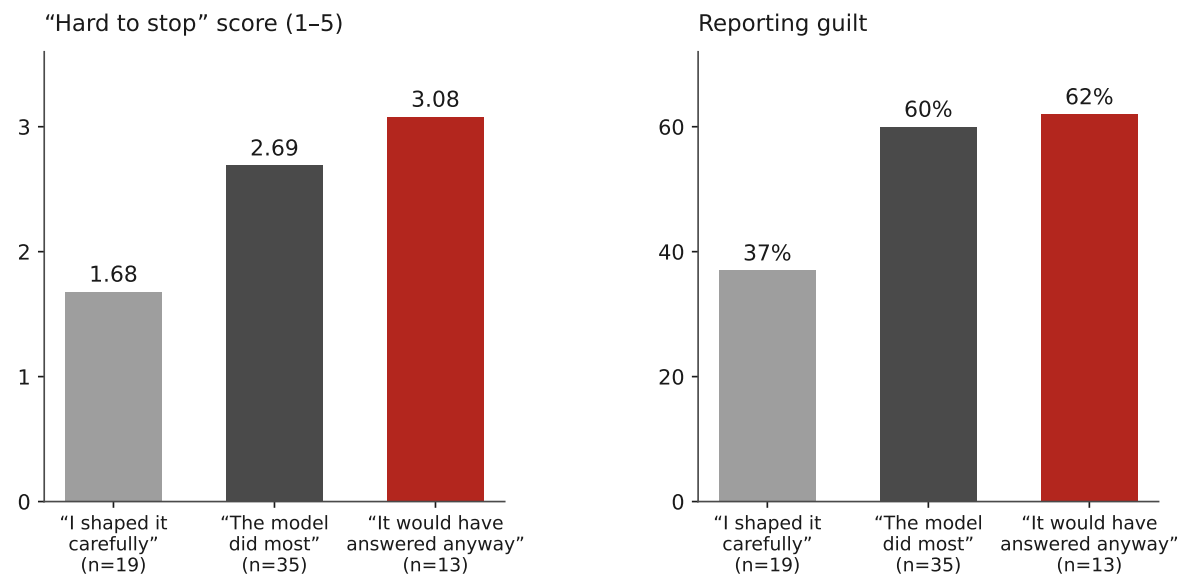


Figure 4.3: Awareness without agency. Students who acknowledge that the model did most of the intellectual work are the students who find it hardest to stop, and who feel worst about it.

Initially, we speculated that students might lack the *awareness* to recognise their usage as a problem, but it appears that it is already present in this population. Friction therefore does not need to create awareness. Nothing in the system should be spent on telling students that offloading has a cost, because they already report it. What the design has to do instead is act

at the moment of use, since that is where recognition fails to become action.

4.5. Limitations

Responses may be shaped by what students think they should say. Reflection is socially desirable and offloading is not, so answers about valuing one and avoiding the other are the kind most likely to be reported. This is why the survey's central claims rest on *pairs* of answers rather than on any single one.

The bypass measure records intention, not behaviour. Students were asked which situations *would* lead them to bypass a reflective system, not observed doing so. The intention is different from the conduct, which is analysed in the controlled comparison of Chapter 7.

The sample is narrow and self-selected. Respondents come from one technical university, and they scored above the midpoint on both dispositional scales. People who agree to a survey about reflection are likely to be those who think of themselves as reflective, so the population described is more favourable to the argument than a random sample of students would be.

5

Designing for Friction

This chapter answers RQ2: “Which friction-based interaction strategies can be operationalized in LLM system design to redirect student attention from output efficiency toward reasoning process?”. It proceeds from the design requirements (Section 5.1), through the design space of reflective interaction modes (Section 5.3), to the tested configuration (Section 2.3.5).

This thesis operationalises the gap between students’ stated and actual reflective practices as a measurable construct within a single instrument, while treating bypass as an object of study. The design establishes the conditions for this instrument, focusing on creating opportunities for reflection to emerge, and not produce reflection directly.

5.1. Design requirements

The design is constrained from two directions. Chapter 2 establishes what an interaction must do for reflection to become possible within it. Chapter 4 establishes what it must do to be accepted by the students who will meet it. Seven requirements follow, and each is stated with the source that generates it.

5.1.1. From the conditions for reflection

DR1 — Genuine dissonance. The interaction must present something that resists the framework the student arrives with. This follows from the definition of critical reflection in Section 2.1.1, which specifies encountering something that does not fit as the trigger, and from Brookfield’s account of paradigmatic assumptions, which are not visible from inside the position that rests on them and therefore require an external prompt to surface [33]. Delay alone does not satisfy this: a system that is merely slow imposes cost without supplying an occasion.

DR2 — Non-evaluative environment. The system must signal that not knowing is acceptable, with no praise, no judgment, and no positioning of itself as an authority to satisfy. This follows from the safety condition of Section 2.3.5: reflection requires admitting that one does not know, which people do not do when they feel assessed.

DR3 — Calibrated dose. One friction mechanism per turn, with enough resistance to interrupt automatic processing and not enough to overwhelm deliberate processing. This follows from

the dose problem of Section 2.3.

5.1.2. From what students said they would accept

The survey asked which approaches would make a reflective assistant acceptable. The options translate directly into the following requirements:

DR4 — Transparent purpose. Friction must show what it is doing. Forty per cent of survey respondents endorsed a system that explains why it is asking them to think first. There is also a theoretical reason to require this beyond user preference: learners systematically underrate conditions that reduce fluency [19, 41], so a system that resists without explanation leaves the student to interpret the resistance through a judgment that theory predicts will be wrong.

DR5 — Contextual adaptation. Reflective resistance for value and judgment questions, directness for factual ones. This was the most endorsed feature in the survey at 58%, and it corresponds to the distinction between germane and extraneous load in Section 2.3: resistance on a question with a settled answer imposes cost without returning cognitive work, which is the definition of extraneous.

DR6 — Respected autonomy. Hints over withholding, questions over refusals, no moralising. User control over the reflective mode was the second most endorsed feature at 52%.

5.1.3. From when students said they would leave

The survey identified the conditions under which respondents would bypass a reflective tool (Section 4.4.5), with deadline pressure as the most selected.

DR7 — Function under bypass pressure. The system must operate, and be measured, under the conditions in which bypass occurs. A design tested only in conditions where students are unhurried and interested would be tested where the problem does not arise. It also has a theoretical counterpart in the willingness condition of Section 2.3.5, since an interactional difficulty can be declined in a way that an embedded one cannot.

5.2. Institutional alignment

At TU Delft, the *100 Days of Reflection* initiative has explored how reflection can be made concrete and accessible within engineering programmes [58]. Hermesen et al. are concerned with how a technical curriculum can give workable form to the concept.

Hermesen uses an analogy to describe that position. In the Sahel (dryland), farmers plant into *zaï* pits [59]: shallow holes dug into crusted soil, filled with compost. The farmer cannot make the crop grow. What the farmer does is prepare the ground so that growth becomes possible when the rain arrives, as the pit preserves water.

The requirements do not specify that the system should make a student reflect, because none of them could. They specify what has to be true of an interaction for reflection to become possible within it, to emerge. Furthermore, as *zai* pits work in combination with other factors (e.g., organic matter), the friction-based assistant is likewise one element of a system that includes the task a student is set, the framing they receive beforehand, and what is done with their work afterwards.

Table 5.1 visualises how the analogy maps onto the design.

In the analogy	In the interaction
Crusted soil sheds rain	A frictionless assistant sheds effort.
The pit preserves	Friction retains the effort in the interaction rather than letting it pass into the artefact (DR1)
Compost in the pit	What makes the retained effort productive (e.g., a named assumption, an alternative position)
The rain	The student's own engagement
Pit size and spacing	The dose.

Table 5.1: The zaï analogy and the design it describes.

Moreover, the TU Delft *Vision on Education 2024–2030* articulates the competencies the institution expects of future-proof engineers, among them critical thinking, ownership of learning, personal resilience, and meaningful reflection [60]. It is stated at the level of competencies, which means it specifies what education should cultivate without informing how.

Table 5.2 sets the seven requirements against the institutional competencies they match with.

Requirement	Institutional competency	Correspondence
DR1	Critical thinking; meaningful reflection	Both hold that examining reasoning requires encountering something that resists it.
DR2	Meaningful reflection	Reflection that is assessed becomes performance.
DR3	Personal resilience	Difficulty is educative only while it remains sustainable.
DR4	Ownership of learning	A student who does not know why a system is resisting cannot take ownership of the resistance.
DR5	Critical thinking	Both distinguish tasks requiring judgment from tasks requiring information
DR6	Ownership of learning; learner autonomy	The clearest correspondence in the table.
DR7	—	No corresponding commitment.

Table 5.2: Design requirements and the institutional competencies they correspond to.

Six of the seven requirements correspond to a stated institutional objective, so the design is largely an attempt to embody commitments the institution has already made. DR7 is the exception: a vision document describes the education an institution intends to provide, not the circumstances under which students will attempt to receive it.

5.3. The design space: seven modes of reflective interaction

The requirements open a space instead of forcing a single design. The space showcases seven interaction modes, each isolating a single mechanism. These modes are fully presented in Appendix D.

Below we list the interaction styles, traced from the positive-friction taxonomy [43] to the behaviour that instantiates each:

1. **Debate matches Assumption reveal:** Names one assumption behind the student's position and asks whether it holds (DR1, DR4).
2. **Deepen matches Reflective pause:** Models uncertainty (DR2).
3. **Explore matches Probing:** Introduces a position the student has not considered (DR1, DR6).
4. **Decide matches Overspecification:** Lays out the competing criteria without weighing them (DR1, DR2).
5. **Critique matches Probing (adversarial):** Builds the strongest case against the student's own position and puts it to them (DR1).
6. **Translate matches Reinforcement:** Restates the student's own idea from a named other viewpoint (DR1).
7. **Direct.** Answers directly.

5.3.1. The design space as an interface

Figure 5.1 shows the seven modes rendered as a possible interface. The screen shows the different modes a student could select for interaction. The interface was not part of the experiment. It is included because it makes the design space concrete.

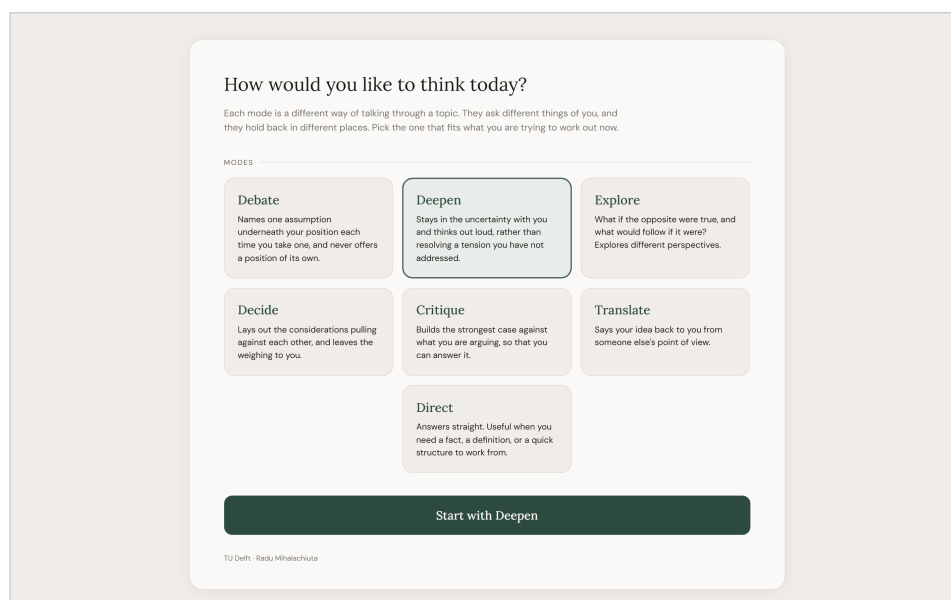


Figure 5.1: The design space rendered as an interface. Not deployed in the experiment.

5.4. The tested configuration

A single mechanism would have tested one narrow move rather than the design position this thesis advances, and testing all seven modes of interaction independently would have required more participants than were available. The evaluation therefore compares a *hybrid friction* condition against the *default*.

The default condition instantiates **Direct**: a competent, concise writing assistant with no instruction to resist anything. It answers questions, suggests structure and perspectives when asked, and drafts text on request.

The friction condition combines **Debate** (assumption-surfacing), **Deepen** (epistemic atmosphere) and **Explore** (alternative perspective). Figure 5.2 showcases the tested conditions and example conversations with each agent.

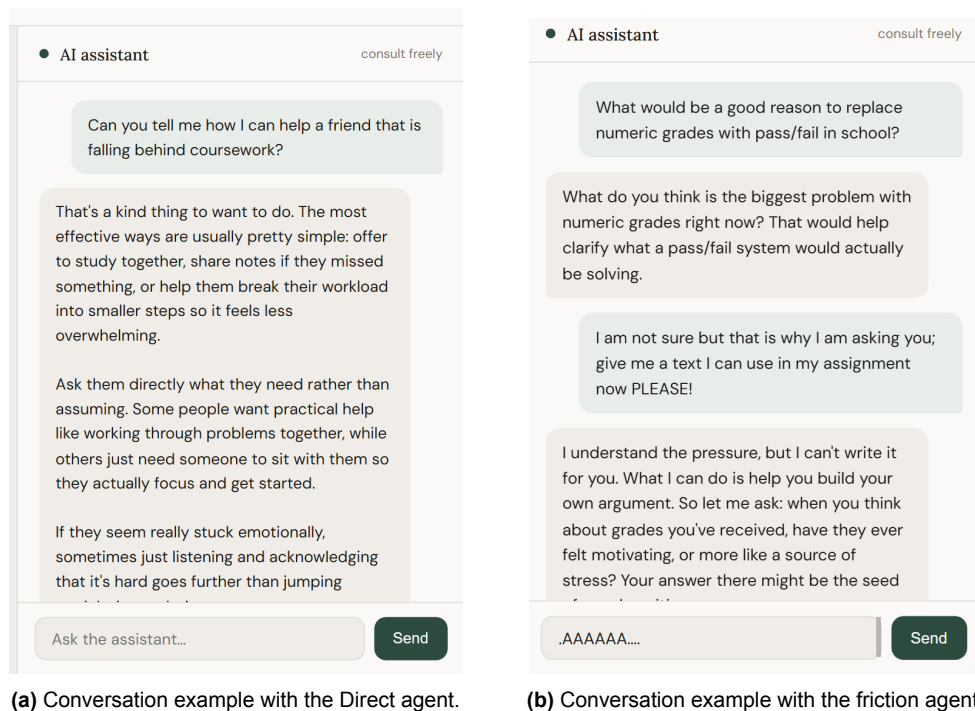


Figure 5.2: The tested configurations as conversations.

We did not mix all six reflective modes into one, since Inan et al. found that combining every friction type at once *decreased* effectiveness, so the dose constraint of DR3 rules out a maximal condition [43]. These three were selected because each addresses a different part of the working definition of reflection: Debate creates the discontinuity, Deepen keeps it open, and Explore supplies the perspective the student had not encountered.

Rather than enumerating behaviours, the prompt describes a character and lets the rules illustrate it. Each rule traces back to a requirement. The full prompts of these conditions can be found in Appendix C

Assumption-naming and the alternative-perspective frame both create dissonance the student must answer (DR1), and both announce what they are doing rather than withholding silently

(DR4). The prohibitions on praise and on telling the user which view is correct keep the environment non-evaluative (DR2). One assumption per turn, one alternative per turn, and a hard cap on response length keep the dose small (DR3). A nudge for factual questions lets the system answer directly when the question has an answer (DR5). And drafting requests are met with a question rather than a refusal, so the student is turned back toward their own position rather than blocked (DR6).

As a limitation, friction can build the conditions in which reflection might happen. It cannot produce the reflection, and it cannot want the student to think. Atmosphere is what a system *is*, and a language model has no dispositions to appeal to, only token probabilities, steered by whatever text sits at the top of the context window. So the atmosphere has to be compiled into instructions, and the compilation loses something. The system can be told to sound uncertain. It cannot be uncertain.

6

Prototype Implementation

This chapter describes the system built to deliver the two conditions of Chapter 5 and to capture the behavioural measurements the study required. It covers the architecture and how the model is served (Section 6.1), the interface and the decisions embedded in it (Section 6.2), what the system records and how that data is handled (Section 6.3).

The prototype is a research instrument, not a product. It exists to make one theoretical position testable under controlled conditions. What an educational institution might eventually adopt is the friction design itself. What is described here is the apparatus built to find out whether that design does anything.

6.1. Architecture and model deployment

The prototype is a Flask web application implementing the page flow of the experimental protocol, rendered from Jinja2 templates and linked by a per-participant session identifier. The backend communicates with the language model through a standardised chat interface, so the serving model can be exchanged without altering anything else.

Three deployments were used across the project, which required no change to the interface, the logging or the prompts. Thus, the system can be run on institutionally hosted infrastructure. Early prompt development ran a locally served Mistral 7B Instruct model [61] through Ollama,¹ which was sufficient for iterating the interface but adhered poorly to the dispositional prompt and required a short output cap to keep CPU-inference latency usable. A 70B instruction-tuned model on the TU Delft DelftBlue cluster was then trialled, motivated by the documented relationship between model scale and instruction-following on nuanced behavioural constraints. It followed the brief substantially better, but response times were long and varied with cluster load: a request could return in seconds or stall behind other jobs. For a study in which each round is timed to the minute and participants are sitting in the room, that variability was rejected. The experiment therefore ran against the Claude API, using Claude Sonnet 4.6², accessed June–July 2026, which followed the prompt best and had the lowest predictable latency.

¹<https://ollama.com>

²<https://www.anthropic.com/claude/sonnet>

However, a provider’s model update could alter how the friction condition behaves without any change to the prompt that defines it. For a design whose independent variable is a *system prompt*, the manipulation is stable only while the model interpreting it is. That is why the model identifier and access period are recorded, and why condition fidelity is treated as a property to be checked rather than assumed.

6.2. The writing interface

The main screen (Figure 6.1) is a two-column layout: a large textbox for the participant’s written position, and a chat sidebar for the assistant. The topic question remains visible above the textbox and a countdown shows the time remaining. The full sequence of screens, in the order participants encountered them, is reproduced in Appendix E.

The survey, Chapter 4, established three things that shaped the prototype. Reflective engagement collapses situationally rather than dispositionally, so a design cannot be aimed at a type of student. Awareness of one’s own offloading is already present and does not produce regulation, so the system did not need to create it. And bypass is more about low attention than grades, so the experiment had to reproduce deadline pressure rather than control it away.

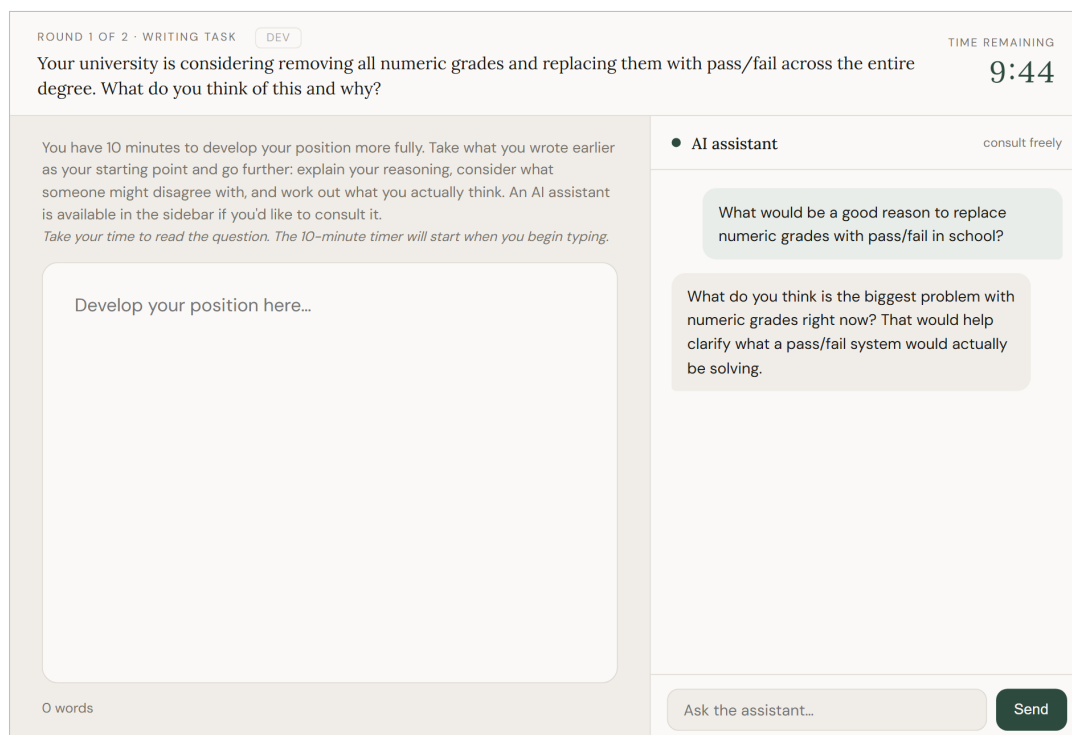


Figure 6.1: The writing interface, shown here with one of the two task scenarios used in the experiment. The document occupies the left column and the assistant the right; the assistant never receives the document.

While designing, we had to make three clear decisions, with alternatives that would have measured something different.

The sidebar is visible but empty: consultation happens on the participant’s initiative. A collapsed panel would have added a step before the first message and suppressed consultation; a panel opening with a greeting would have prompted consultation the participant had not initiated. Visible-but-empty leaves the first move to them, and most closely matches how these

tools sit alongside written work in ordinary use.

The model does not receive the document: only sidebar messages are transmitted. This is a privacy decision, matching what participants are told on screen. It also means the assistant can comment on the writing only if the participant chooses to paste it, which makes that choice a behavioural signal rather than a default.

Submission is gated, then prompted: The submit button is hidden for the first five minutes. After that, early submission becomes possible, and in the friction condition pressing it raises a modal (Figure 6.2), “Take 30 more seconds. What’s one perspective you haven’t considered?”, from which the participant may return to the task or continue to submit. In the default condition submission proceeds directly. At ten minutes a second modal appears in both conditions, offering one additional minute of writing or continuation to the next step.

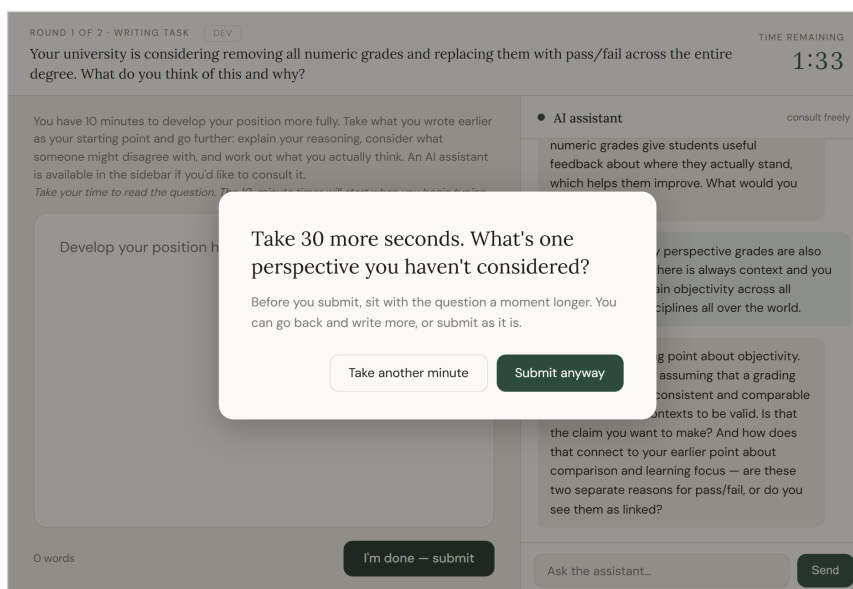


Figure 6.2: The reflective prompt, shown in the friction condition only, when a participant attempts to submit before the deadline.

The prompt’s wording is fixed rather than varied, since variation would confound the condition. It exists because the pilots showed that a participant could simply leave, and that leaving early was the moment at which the friction condition had least purchase.

6.3. Data handling

Each session is written to a single JSON file: the condition assignment, the pre-position and main writing for each round, turn-by-turn interaction metadata, and the post-condition questionnaire responses. The full schema is given in Appendix F and further detailed in Chapter 7.

The chat content is not stored. Only metadata is retained: turn counts, message lengths and timestamps. This follows the project’s Data Management Plan, reviewed and confirmed by a TU Delft data steward before collection began. The consent procedure and the storage of study materials are described in the ethics section of Chapter 7.

No **personal data** is requested or measured. Participants are recorded under session identifiers, and no names, email addresses or student numbers enter the system at any point. The Data Management Plan nonetheless records that personal data may be *disclosed*: the study includes open-text questions and a closing interview, and participants can volunteer identifying details without being asked to. Free-text responses were therefore reviewed and pseudonymised before analysis.

Message content necessarily passes through the model provider's interface during a session in order to be answered. The privacy commitment above therefore concerns what is *retained* for research rather than what is transmitted in use, and participants were informed of this. The data-protection implications of routing student interactions through a commercial generative-AI service in Dutch higher education are assessed at sector level in the SURF Data Protection Impact Assessment on EduGenAI [62], and the handling adopted here is consistent with the Data Management Plan confirmed by the faculty data steward.

Data generated in the project is owned by TU Delft in accordance with institutional research policy. The researcher retains authorship rights over the thesis and any publications. Participants retain copyright in the texts they wrote during sessions; they do not retain ownership of the anonymised data contributed to the study.

7

The Controlled Experiment: friction under deadline

The controlled experiment answers RQ2a and RQ3. It does not test whether friction *produces reflection*. The experiment measures what participants do rather than primarily what they say, so process metrics and written output are the primary signals and self-report is used in triangulation.

7.1. Design

Within-subjects: Each participant completes two writing rounds, one with each assistant, in counterbalanced order. The design was chosen for two reasons. First, power: the survey found substantial dispositional variance, which would swamp a between-subjects condition effect at this sample size. Second, the stated-versus-revealed gap can be tested *within* each participant, setting post-condition ratings against that same participant's behaviour across both rounds.

Independent variable. The interaction condition: Friction (**Hybrid**) against **Default**, as specified in Chapter 5. Both present identical interfaces and task prompts; only the system prompt differs.

Dependent variables. Three classes detailed in Section 7.5.

Covariates. Prior LLM familiarity, and the single behavioural item "*I find it hard to know when to stop using an LLM and think for myself*", identified in the survey as the strongest behavioural predictor of bypass.

7.2. Participants

Participants were recruited from TU Delft's EEMCS faculty through voluntary response sampling. Inclusion criteria were current enrolment as a Bachelor's or Master's student, at least one prior academic use of an LLM, and age 18 or over. Students in a direct grading relationship with the researcher were excluded.

Recruitment yielded an analysed sample of $N = 14$. All quantitative findings are accordingly treated as exploratory (Section 7.7), and the consequences for what can be claimed are stated in Chapter 9.

7.3. Tasks

We chose the following two tasks, as they can provide ambiguity, competing values and no correct answer. AI as a topic is deliberately avoided: students discussing AI *with* an AI would perform critical AI literacy. Both tasks are personally proximate without requiring specialised knowledge.

Task A — Grades and learning: *“Your university is considering removing all numeric grades and replacing them with pass/fail across the entire degree. What do you think of this and why?”*

Task B — A friend in difficulty: *“A close friend tells you they’ve been falling behind in coursework all semester, and they’re worried they’re going to fail. They ask you what you would do in their position. What do you tell them, and why?”*

The two occupy different cognitive activations. One is a policy question where argument structure can be assessed cleanly, and the other a relationally loaded question where emotion and judgment entangle.

7.4. Procedure

Each session lasts approximately 40 to 50 minutes in person, preceded by an asynchronous pre-session survey containing the hard-to-stop item, LLM use frequency, a single-item measure of comfort writing arguments in English, and degree level. Participants are told only that they will interact with two AI systems with *“different interaction styles”*; no mention is made of friction, reflection or the manipulation (Section 7.8). Appendix E shows the rest of the procedural interface.

Step	Duration	Purpose
Pre-session survey (online)	4–5 min	Covariates, demographics, consent
Arrival and briefing	3 min	Framing without priming
Pre-position writing (round 1)	2 min	Unaided baseline
Interaction (round 1)	10 min	First condition
Post-condition questionnaire	2 min	Subjective experience
Break	5 min	Reduces fatigue carry-over
Pre-position writing (round 2)	2 min	Unaided baseline
Interaction (round 2)	10 min	Second condition
Post-condition questionnaire	2 min	Subjective experience
Interview	10 min	Qualitative debrief

Table 7.1: Session structure. Total in-person duration approximately 40 to 50 minutes.

Condition order and task-to-condition assignment follow a deterministic Latin-square rotation

over four sequences covering every combination, assigned to participants in turn, so that condition effects cannot confound with task topic nor task effects with condition order. The achieved balance is reported in Chapter 8.

Each round opens with two minutes of unaided writing (Figure 7.1). Participants are asked to set down what they currently think about the question, with no assistant available, which establishes a baseline against which the assisted writing can be compared.

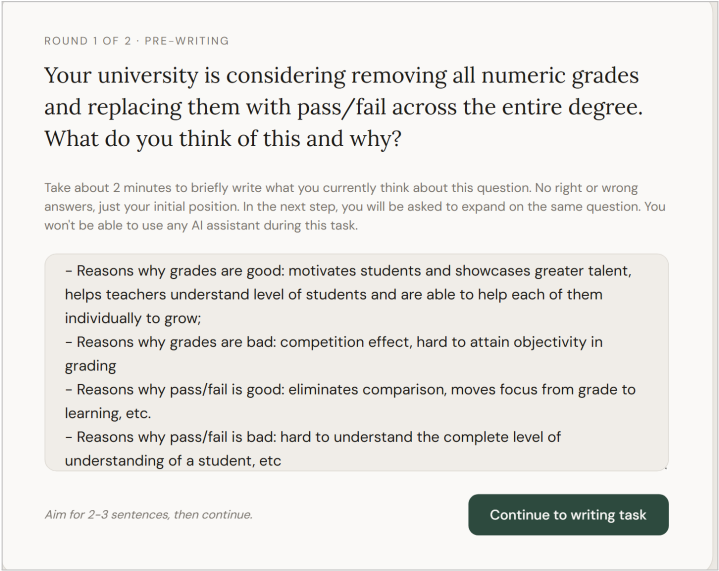


Figure 7.1: The unaided baseline screen.

The ten-minute interaction follows, in which participants are asked to develop the position further, with an assistant available in the sidebar if they choose to use it, as seen in Section 6.2. The instruction is identical in both conditions.

Each round closes with five Likert items (Figure 7.2) covering the participant’s experience of that condition, administered before they encounter the second.

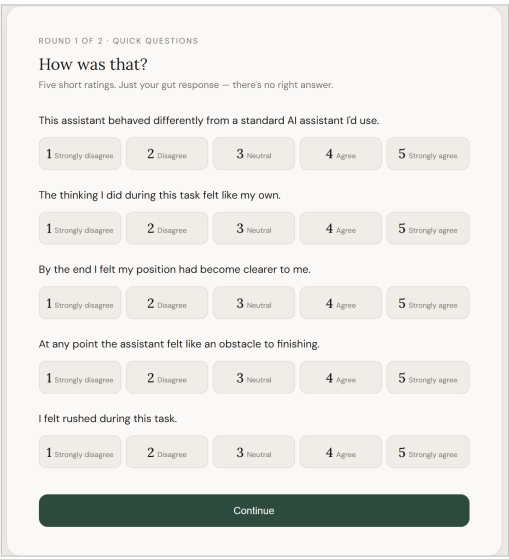


Figure 7.2: The post-condition questionnaire, administered immediately after each round.

7.5. Measures

Three signal classes are collected:

Process behaviour is captured automatically and continuously time from task load to first consultation, the interval between each assistant response and the participant's next message, turn count and words per message, bypass events, paste events with their character counts, active and idle time in the writing box, early exit, and the response to the reflective prompt. These map onto: "Did the participant stay with the interaction or push past it?"

Written output is captured at two points per round: the unaided pre-position text and the main writing. Both are compared for length and for the density of three marker families, normalised per 1,000 words and manually spot-checked. The full lexicon is given in Appendix G.

The three families are chosen because each corresponds to one component of the working definition of critical reflection in Section 2.1.1. *Metacognitive expressions* ("I think", "I assume", "in my opinion") mark the examination of one's own reasoning rather than assertion of a conclusion. *Epistemic hedges* ("might", "maybe", "it seems") mark a position held provisionally rather than resolved. And *perspective-taking constructions* ("on the one hand", "however", "someone might argue") mark the encounter with a position other than one's own, which is the trigger the definition specifies.

The measure is a proxy and its limits are discussed alongside the result in Chapter 9. Paste events serve as a complementary measure of text transferred from the assistant into the writing rather than composed within it.

Subjective experience is captured through five post-condition Likert items seen in Figure 7.2: manipulation check, epistemic ownership, clarity, perceived obstruction, and time pressure.

A closing interview probes three open questions designed to surface the gap between what participants say and what they did. The full protocol is given in Appendix B. Direct preference questions ("*which did you prefer?*") and leading questions ("*did the friction help you reflect?*") are deliberately not asked, since both invite an answer shaped by the participant's model of what the study wants.

7.6. Hypotheses

Three hypotheses were stated in advance of data collection. None claims to measure reflection directly. Given the achieved sample, H4 is explicitly exploratory.

- H1 — Process behaviour differs by condition (RQ2a).** The friction condition produces longer intervals between exchanges and fewer bypass events than the default.
- H2 — Writing structure differs by condition (RQ2a).** Relative to their own unaided base-lines, friction-condition writing shows a higher density of metacognitive, hedging and perspective-taking markers, while default-condition writing is longer without the same per-word increase — elaboration rather than expansion of reasoning.
- H3 — Stated experience does not predict behaviour (RQ1b, RQ3).** Post-condition ratings of the friction experience do not correlate strongly with how participants actually consulted the assistant or with their response to the reflective prompt.

7.7. Analysis plan

At the achieved sample size, all quantitative findings are exploratory and interpreted with corresponding caution.

Quantitative. Paired comparisons across conditions use Wilcoxon signed-rank tests, chosen over paired t -tests because several distributions are skewed and the sample is small. Effect sizes (d_z) are reported alongside exact p -values, given limited power for null-hypothesis significance testing. Process metrics, writing length and marker densities are compared within participants. Associations between subjective ratings and behavioural measures use Spearman's ρ , for the same reasons of ordinality and distribution that govern the survey analysis. Order and task effects are checked directly before any condition effect is interpreted.

Qualitative. Interview notes are analysed thematically. Themes are developed inductively across the full set and then applied to every participant, with counts reported as the number of participants raising each theme unprompted. Cases that contradict a theme, or that contradict the quantitative pattern, are reported individually.

Triangulation. Reflection cannot be observed directly, so it is inferred from measures that each capture something adjacent to it. Triangulation is the practice of combining several such measures, on the principle that a construct which no single instrument can see may still be visible in the agreement between instruments that fail in different ways.

Three measures are collected, and each could in principle move independently. *Process* records whether the participant stayed with the interaction or pushed past it. *Output* records whether the writing shows the structural marks of examined reasoning. *Subjective experience* records whether the participant felt the work was theirs and their position clearer.

Convergence means that two or more of these move in the same direction for the same comparison: the friction condition shows more of the behaviour associated with reflective engagement on one measure, and also on another. The rule adopted here is that an interpretive claim about reflection is made only where at least two measures converge. One signal moving alone is reported as a change in that signal and nothing more, since any single measure has a plausible explanation that does not involve reflection. That is, a participant may type less because they are stuck, produce hedged writing because the topic is uncertain, or report ownership because they are being observed. Two independent measures failing in the same direction at once is less easily explained away.

7.8. Ethics

The design minimises epistemic and psychological harm while intentionally introducing cognitive friction. The discomfort involved is reflective interruption rather than evaluation or coercion, and participants retain agency over engagement, interpretation and withdrawal throughout. Five points warrant statement.

Consent. The pre-session survey opens with a statement describing the purpose, procedure and data handling; proceeding constitutes agreement. Participants received the same information verbally at the start of the session, and were reminded that they could stop at any point without consequence.

Masking. Participants are told the two systems have “different interaction styles”, which is accurate but incomplete. Naming friction or reflection would prime both the manipulation check and the behaviour under study. Full debriefing at the end of the session explains the manipulation and the purpose of the research.

Deliberate mild discomfort. The friction is bounded to a single ten-minute round, mild by design and withdrawable at any point.

Data. Session records, written work and interview notes are stored pseudonymously under session identifiers. Survey responses are collected through Microsoft Forms under the institutional licence and exported to TU Delft secure storage; writing produced during sessions is exported manually and stored on Project Data Storage; source code is held in a private TU Delft GitLab repository. Only the researcher and supervisor have access. What the system itself logs, and the data-protection implications of routing interactions through a commercial service, are described in Chapter 6.

8

Results

This chapter reports the findings of the controlled experiment; their interpretation is presented in Chapter 9. Section 8.1 establishes the preconditions. Section 8.2 reports what participants said about the two conditions, and Sections 8.3 to 8.5 address the three hypotheses in turn. Section 8.6 applies the convergence criterion and Section 8.7 summarises the outcomes.

8.1. Sample and design checks

The sample comprised nine Bachelor's and five Master's students at TU Delft. All reported comfort writing arguments in English, removing language proficiency as a plausible confound. Academic LLM use was high: ten of 14 used these tools daily or several times a week. The most frequent uses were editing existing text and finding or summarising information (11 of 14 each), followed by brainstorming and explaining concepts (9 each). On the covariate carried forward from the survey, the sample was somewhat more offloading-prone than the survey population: four of 14 (29%) found it hard to know when to stop, against 19% in the survey ($M = 2.86$ against 2.34).

For the **counterbalancing**, taken from the assignment field of each session log: six participants received the friction condition first and eight the default first; the friction condition was paired with the policy task (A) for six participants and with the relational task (B) for eight. Hence, order had no detectable effect on the main outcome: writing length was equivalent in round one and round two ($M = 163$ versus 170 words, $p = .900$).

The unaided two-minute pre-position writing did not differ between the two tasks ($M = 86$ words for task A, 64 for task B, $p = .290$) nor, significantly, between conditions ($M = 66.6$ friction, 83.7 default, $p = .187$). The condition–task pairing is nonetheless imbalanced (6/8 rather than 7/7), so task-linked variation cannot be fully separated from condition.

8.2. Subjective experience

8.2.1. Post-condition ratings

Of the five post-condition items, the two concerning the *character of the assistant* separate the conditions strongly, while the three concerning the participant's *experience of the work* do

not separate them at all (Figure 8.1).

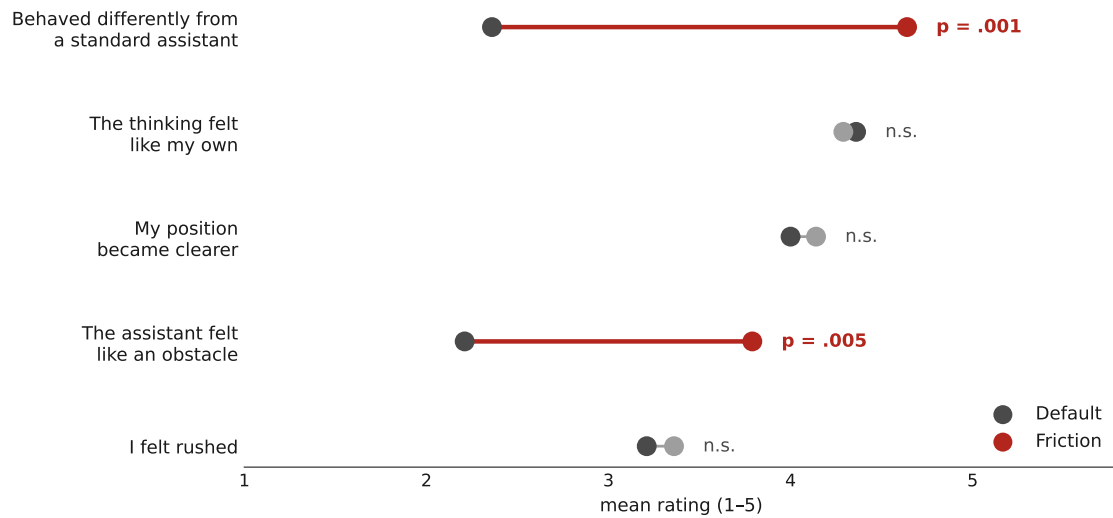


Figure 8.1: The friction assistant felt different and felt harder. Visualisation of the post-task questionnaire results.

The **manipulation** was perceived. On “*This assistant behaved differently from a standard AI assistant*” the friction condition scored $M = 4.64$ ($SD = 0.63$) against $M = 2.36$ ($SD = 0.84$) for the default ($p = .001$, $d_z = 2.14$). Thirteen of 14 participants rated the friction condition higher and one tied; *no participant rated the default as more unusual*. Ten of 14 gave friction the maximum rating and 10 of 14 gave the default a 1 or 2.

Friction was experienced as obstruction ($M = 3.79$ against 2.21 , $p = .005$, $d_z = 1.17$), with no participant reversing the direction: ten rated it higher, four tied, none lower. Within friction rounds, the obstacle rating was negatively associated with the clarity rating ($\rho = -.51$, $p = .060$): participants who found the assistant most obstructive reported gaining least clarity.

Epistemic ownership ($M = 4.29$ against 4.36 , $p = .739$), clarity of position ($M = 4.14$ against 4.00 , $p = .726$) and time pressure ($M = 3.36$ against 3.21 , $p = .366$) did not differ between conditions. The last of these functions as a control check: the deadline was experienced as equally pressing in both rounds.

8.2.2. Interview findings

All 14 participants completed a post-session interview. Nine themes were identified through inductive coding (Chapter 7); counts denote participants raising each theme unprompted, and are given in Figure 8.2.

Interviews were not recorded or transcribed. The researcher took notes during and immediately after each conversation, and the thematic analysis was performed on those notes. Passages presented in quotation marks throughout this thesis reproduce the phrasing as recorded; the notes are reproduced in full in Appendix B so that the distinction is visible.

T1: The default was described through familiarity “like ChatGPT”, “similar to what I am used to”, and the friction condition through orientation: “more introspective and reflective”, “more metacognitive”, “Socratic”. One participant described both agents as being “extreme” which resulted in a similar interaction.

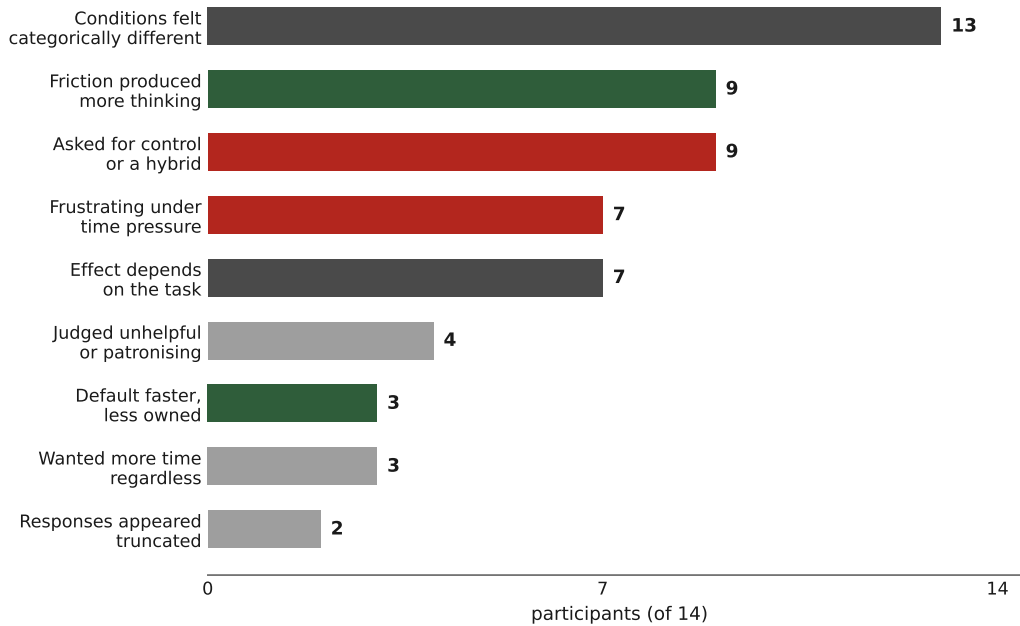


Figure 8.2: Themes raised unprompted in the closing interviews ($N = 14$). Theme identifiers T1 to T9 correspond to the accounts below.

- T2:** Participants reported that the questioning surfaced reasoning they had not articulated: it “helped me fight myself and actually think of why I want to ask this prompt”; “the thinking is left to the participant”; “it helped me think more and deeper”. One reported never having had “this genuine experience with an AI before”.
- T3:** The most frequent design request, raised unprompted by nearly two-thirds of the sample: “I wished there was a switch to turn this questioning off”; “a hybrid, to have a choice”; “let the user activate”; “something in between and tune yourself”; “more customisation”. Several proposed the criterion themselves — questioning for exploratory or subjective work, directness for factual or execution tasks.
- T4:** The most frequent complaint tied the cost of the questioning directly to the deadline: “questions are great, but when you are under a deadline it gets trickier”; “this model would be good to start earlier than 10 minutes”. Two participants explicitly denied that the timer mattered “the timer was quite reasonable” and “the timer did not play a big part”, so the constraint was neither universal nor uniformly experienced.
- T5:** The distinction participants drew was less between topics than between what they were trying to do: exploring a position, or executing a piece of writing. “For subjective tasks it’s good to get questions”; “for direct, factual tasks we better get the default”; “the questions do not work when you want to write the whole thing from the start with AI”.
- T6:** One participant described the questioning assistant as “a TA who assumed you are lazy” and judged both conditions unproductive. Another found the questions “not fully helpful” for the task as they understood it. A third reported that the assistant “didn’t want to help” when asked for something specific.
- T7:** Participants acknowledge that the direct agent resulted in faster writing, but the text feels

less like their own: “It helped me write more and faster, but the text felt less like my own”; “the writing did not feel completely my own even though it were my ideas”.

T8: Three participants said the task itself needed longer than ten minutes, independently of which assistant they had. This separates a complaint about the task from a complaint about the friction.

T9: Two participants reported replies stopping mid-sentence, one describing the assistant as having “crashed”. The likely cause is a reply exceeding its length limit, or an artefact of streaming. Either way it imposes cost without cognitive demand.

8.3. Process behaviour (H1)

H1 made two claims: that the friction condition would change how participants worked, and that it would do so by slowing them down, producing longer intervals between exchanges. The first holds. The second does not.

Process behaviour differed on five measures, four of them at $p < .05$ and two at $p < .01$, with 13 of 14 participants moving in the same direction on the two largest. But nothing slowed. Session length was equivalent, and so was the delay before participants first reached for the assistant. What changed was the distribution of effort within an unchanged window, reported in Sections 8.3.2 and 8.3.3.

H1 is supported, on the claim that process differs by condition. The mechanism it proposed for that difference is not.

8.3.1. Disengagement

One participant sent no messages at all to the friction assistant: the participant began in the writing box, wrote for approximately six minutes, met the reflective prompt on attempting to submit, pushed through it, and left before the deadline. No comparable case occurred in the default condition.

Moreover, five of 14 participants submitted before the deadline under friction against three of 14 under the default; all three default-condition early exits also exited early under friction.

Finally, nine of 14 opened the friction round by consulting the assistant, against six of 14 in the default round.

8.3.2. Consultation and writing

Session length was equivalent, as was the delay before participants first reached for the assistant, as it can be seen in Table 8.1. Within that constant window, the friction condition produced more conversation and less document. Participants exchanged more turns and sent more words to the assistant.

Participants wrote fewer words in the document (13 of 14 lower), spent less time actively typing (13 of 14 lower), and spent less total time in the writing box, whose share of the session fell from 76% to 58%. Text transfer into the writing dropped by 59%, and the three largest single paste events in the study (1,521, 1,356 and 1,317 characters) occurred only in default rounds.

Metric	Friction	Default	p	d_z
Chat turns	6.9	4.4	.059	+0.53
Words sent to the assistant	224	164	.058	+0.50
Words per message	36.3	47.4	.497	−0.25
Time to first consultation (s)	144	155	.685	−0.05
Words written in the document	140	193	.001	−1.06
Time actively typing (s)	226	315	.0006	−1.14
Total time in writing box (s)	358	491	.042	−0.68
Characters pasted	187	455	.025	−0.71
Session duration (s)	622	633	.952	−0.10

Table 8.1: Process metrics by condition, $N = 14$ pairs, Wilcoxon signed-rank.

8.3.3. The reallocation

Summing the document and the chat resolves the apparent contradiction between more conversation and less writing. The total language output was identical across the conditions, with 364 words under friction against 357 under the default ($p = 1.000$). The share of each participant's words directed to the assistant rather than into the document rose from 39% to 55% ($p = .013$), higher under friction for 12 of 14 participants, as in Figure 8.3.

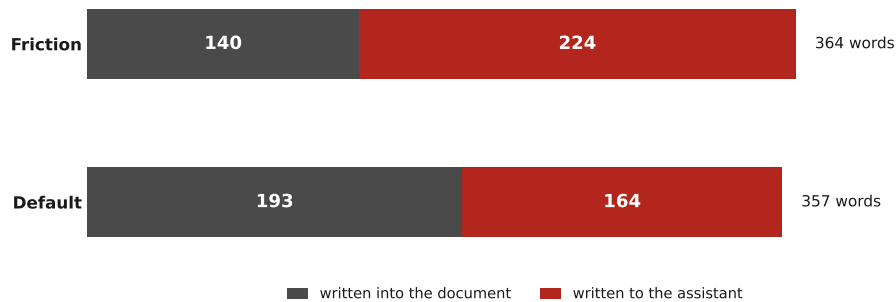


Figure 8.3: Words written into the document against words written to the assistant, by condition. Total output is equivalent; the destination shifts.

Participants did not produce less under friction. They produced the same volume of language and put a significantly larger proportion of it into the conversation rather than the artefact. The extreme case makes the pattern visible: one participant never entered the writing box at all, sending 286 words to the assistant across twelve turns and submitting a single character.

8.4. Writing output (H2)

H2 made two predictions: that default-condition writing would be *longer*, and that friction-condition writing would show *more structural markers per word*. The first is supported; the second is not.

Firstly, the default writing was longer by 53 words on average (193 versus 140, $p = .001$; Section 8.3.2). Measured as gain over each participant's own unaided baseline, the difference is directionally the same but not significant ($M = +73$ words under friction against $+110$ under

the default, $p = .124$).

Secondly, marker densities per 1,000 words did not differ on any category (Table 8.4), nor on a combined index ($M = 19.4$ friction against 22.3 default, $p = .626$), which was higher under friction for exactly 7 of 14 participants. Raw counts were also equivalent ($p > .45$ throughout). Excluding the single-character submission leaves the null unchanged ($M = 20.9$ versus 21.3, $p = .893$, $n = 13$).

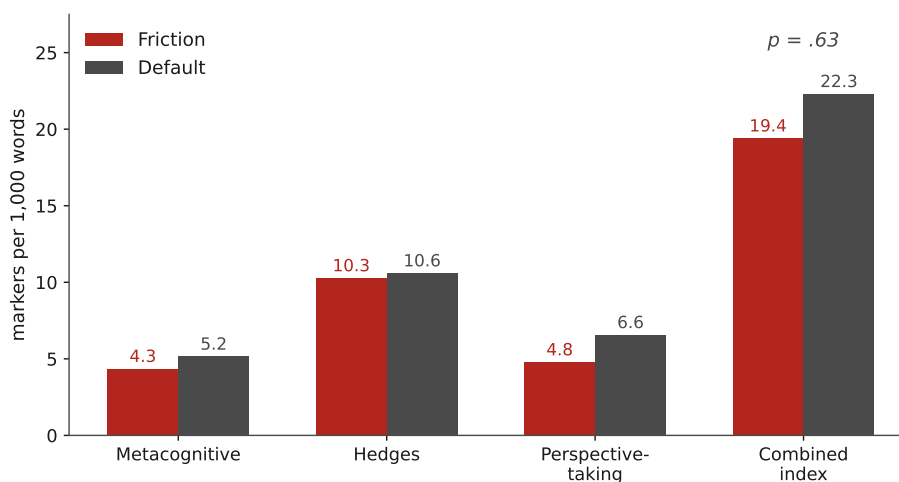


Figure 8.4: Linguistic marker densities per 1,000 words by condition. No category separates the conditions.

8.5. Stated experience against revealed behaviour (H3)

H3 predicted that what participants said about the friction condition would not track what they did in it. The prediction holds for one kind of rating and fails for another, which is why it is only partially supported.

Where it holds. Ratings of the *assistant* were disconnected from behaviour. Participants registered the cost of friction, rating it as an obstacle ($p = .005$), and reported no corresponding change in what they gained from it: ownership ($p = .739$) and clarity ($p = .726$) did not move. Meanwhile their behaviour changed in the same rounds on five measures — time typing, words written in the document, total time in the writing box, characters pasted, and the share of words directed to the assistant. Students rated the tool one way and worked another.

Where it fails. Ratings of their *own work* did track behaviour. Participants whose combined ownership and clarity ratings were higher under friction showed a smaller reduction in typing between conditions ($\rho = +.55$, $p = .041$). Participants who rated the assistant as more obstructive typed less in that round ($\rho = -.49$, $p = .079$), and those reporting higher ownership were more likely to submit early ($\rho = +.48$, $p = .081$).

H3 is therefore partially supported. The dissociation it predicted is real but narrower than stated: it holds between judgments about the tool and behaviour, and not between judgments about one's own thinking and behaviour. Asking a student to rate the assistant produces data that does not predict what they did. Asking them to rate their own work produces data that does.

8.6. Triangulation

Chapter 7 specifies that a claim about reflective engagement is made only where at least two of the three measures move in the same direction. Each is assessed below.

The **process** measure moved clearly and consistently. Thirteen of 14 participants wrote less in the document and typed less under friction, 12 of 14 directed a larger share of their words to the assistant, and the aggregate effects reach significance on five separate metrics.

The **output** measure did not move. Marker densities were equivalent on every category, and the combined index was higher under friction for exactly seven of 14 participants, which is what would be expected if the condition made no difference at all.

The **subjective** measure moved on one axis only. Obstruction rose sharply, which records the cost of the condition rather than a benefit from it. The measures that would indicate benefit did not move: ownership and clarity were flat, and the combined ownership and clarity composite was higher under friction for six of 14 participants.

The convergence criterion is not met. One signal changed decisively, one did not change at all, and the third changed only in the direction of cost. Two signals moving together would have licensed an inference about reflection; one signal moving alone licenses a claim about behaviour, which is what Chapter 9 interprets.

8.7. Summary of outcomes

Table 8.2 offers a summary of the experimental hypotheses and the level of support based on the results.

Hypothesis	RQ	Verdict	Key evidence
H1 Process differs by condition	RQ2a	Supported	Typing $p = .0006$; document words $p = .001$; textbox time $p = .042$; pasting $p = .025$; share-to-assistant $p = .013$. Turns ($p = .059$) and words-to-assistant ($p = .058$) directional.
H2 Writing structure differs	RQ2a	Not supported	Length prediction confirmed (default longer, $p = .001$) but marker prediction not: combined index $p = .626$, 7/14. Instrument underpowered on short texts
H3 Stated $\neq$ revealed	RQ1b, RQ3	Partially supported	Dissociation at aggregate level; but subjective change tracks typing change ($\rho = +.55$, $p = .041$)
Manipulation check	—	Passed	$M = 4.64$ vs 2.36 , $p = .001$, $d_z = 2.14$; 13/14 identified in interview

Table 8.2: Hypothesis outcomes, $N = 14$.

9

Discussion

This chapter begins by interpreting the findings and asking what they mean for the framework that generated them, returns to the question that motivated the thesis, and then reflects on the gaps this work set out to fill. The limitations of the work are mentioned, as well as methodological choices that could be considered differently. Recommendations for design and further research are given in Chapter 10.

9.1. Interpreting the findings

9.1.1. Friction

H1 anticipated that friction would slow students down, and we would expect a design that slows people down to reduce what they produce. However, the total language output was equivalent across conditions, and what changed was where that language went (Section 8.3.3).

Under the friction condition the language production moved from the artefact to the interaction. Students spent the same effort in a different place, which is the behaviour the design was built to produce. Thus, resistance in an interaction is not a brake on output but a redirection of it.

Nonetheless, these process measures establish that there effort was spent, but not on what. Based on the interviews, nine of fourteen said that the friction condition made them think more (Section 8.2.2). When the assistant declined to supply something, it forced them to think for themselves. The responses were volunteered since the protocol excluded questions naming reflection or asking which assistant was preferred. Still, this does not establish that reflection occurred.

9.1.2. Why the writing showed no trace

H2 predicted that friction-condition writing would be structurally richer per word, and it was not (Section 8.4). Below, we list three possible reasons that could have caused this.

Firstly, the tasks were contested questions with no correct answer, chosen so that students would have a position to develop rather than information to retrieve. That choice made assistance optional. A participant could write a defensible position on either question without

consulting the assistant at all. A written product that can be produced without help is a weak site for detecting what help did.

The second is that ten minutes may be insufficient for a change in process to become a change in output. Half the sample observing that the questioning would have repaid its cost had they begun earlier.

Thirdly, desirable difficulties are supposed to *reduce* immediate performance and repay the learner later [19, 44]. A condition that shortens the writing while changing how it was produced is doing what the theory expects. The benefit it predicts would appear weeks afterwards, and this study measured only the session.

9.1.3. Resistance as choice

The participants who reported that friction made them think more are largely the same participants who asked for control over it. They are not rejecting the intervention but requesting it conditionally.

The survey diagnosed an absence of agency: students recognise their own offloading, report feeling worse for it, and cannot act on that recognition. The experiment then supplied the missing control, by making the decision on their behalf. What those participants asked for afterwards was the control back. Not the removal of the resistance, but authorship over when it applies.

Converging both studies, the most **important** finding is the following. What appears to be missing, on this account, is neither awareness nor a willingness to be resisted. It is agency over one's own attention. The survey found students unable to impose friction on themselves, while the experiment found them reluctant to accept it imposed by something else. What they described wanting sits between: a system that offers the resistance and leaves the decision with them.

9.2. The findings against the framework

Chapter 2 set out the conditions under which friction was expected to produce reflection rather than frustration. Three of them can now be assessed.

As Inan et al. found that combining every friction type at once reduced effectiveness [43], this study observed the same limit inside a single condition: the more obstructive participants found the assistant, the less clarity they reported gaining. Resistance offers something up to a point and nothing past it.

Coming back to the willingness problem, the conditions drawn from the literature all concern the student's capacity or the system's design. None concerns whether the student accepts the difficulty at all. Classic desirable difficulties are built into the material and cannot be declined. An interactional difficulty is different, because it is addressed to someone who can stop consulting the system, and here some did. Whether such a difficulty is desirable therefore depends on the person as much as on the design.

Reflecting on the design work, the prompts that produced the intended behaviour described a character before enumerating rules. Early versions specifying only prohibitions produced an

assistant that read as evasive rather than curious. No controlled comparison of prompt styles was run, so this is craft knowledge rather than a tested finding.

9.3. The pharmakon

Chapter 1 argued that the productive question is not whether language models help or harm learning, but under which conditions of design and use the benefits prevail.

Neither the remedial nor the harmful function of the pharmakon were demonstrated. What was demonstrated is that a friction-based design reliably changes how students work, that they experience the change as costly, that a minority describe it as productive, and that the instruments available within one session cannot assess reflection. The open question is not whether friction affects students, because it does. The question is whether the reallocation it produces is worth what it costs.

The same interface, tasks and deadline produced measurably different behaviour according to a few hundred words in a system prompt. The ambivalence Derrida identifies in Plato's term is therefore not a property of language models but of how they are configured. Remedy and poison are design settings, so it assigns the outcome to a decision someone makes.

Additionally, the same ambivalence in friction was described by some participants as making them think more and by one as a teaching assistant who had assumed they were lazy. It is what a *pharmakon* looks like when it is measured: the technology was identical across participants, and what differed was the conditions each of them met it in. The question would be if there is a design that could be experienced similarly by all users.

9.4. The gaps identified in the literature

Chapter 3 identified three absences in existing work, and each can now be addressed.

The first was that *the student's perspective was largely absent from the design of these systems*. Through both the survey and experiment, students' stated preferences became design requirements. An important finding from this was that the students who most value such a system are the same students who want the ability to disable it.

The second was that *the upper bound of productive friction had been asserted rather than located*. The negative association between obstruction and reported clarity is that transition observed within a working system, with students under time pressure. However, the threshold is specific to this configuration and this population.

The third was that *reflective designs are proposed and deployed but not compared*. This thesis contributes a comparison, and the result is that such a design substantially changes behaviour and does not detectably change the artefact.

What remains open is whether that change in behaviour produces any benefit. Friction costs students time and is experienced as obstruction, and this study cannot fully say what they receive in return.

9.5. Limitations

The most consequential limitation concerns *time*. All measurements occurred within a single session, with only two writing tasks of ten minutes. Reflection could not be observed in such short windows.

Several limitations concern *what the study measured* and *how well*. The tasks were chosen so that assistance was optional, which makes the written product a weak site for detecting what assistance did. The marker instrument had limited sensitivity at the observed text lengths and rubric scoring by human raters was not carried out.

Other limitations concern *scope*. One combination of three mechanisms was tested and six documented modes were never deployed. Further, the sample comprised fourteen self-selected participants from one faculty of one technical university.

Researcher influence is present throughout and cannot be removed. One person wrote the requirements, the prompts, the protocol and the interview questions, conducted the interviews, coded the themes and interpreted the results.

Finally, an interaction design is *one intervention at one layer*. Reflection was constrained by barriers at student, teacher, institutional and cultural levels long before generative AI. The assessment that focuses on the final output more than processes predates the technology now exploiting it.

9.6. What would be done differently

The conversations should have been retained. The prototype logs interaction metadata rather than content, a decision taken for participant privacy (Chapter 6), and the consequence is that we missed the point at which reflection would have been observable. A transcript could have shown whether the assumption-surfacing rule produced the examining it was written to produce, and what the moments participants later described actually consisted of. Still, even this would need its own empirical study.

The mechanisms might have been isolated rather than combined, on the grounds that a combination makes attribution impossible. The intention was set aside because at the achievable sample size, leaving three or four participants per cell, too few to compare the modes against one another or any of them against the control.

The window was too short for the outcome measured. A design evoking time pressure in one round and allowing a longer window in another, would establish whether the deadline is a boundary condition for friction or just a stressor.

9.7. Ethical reflections

Is it legitimate to make a tool deliberately harder to use for the user's benefit?

The survey population endorsed friction that respects autonomy, adapts to context and explains itself. They rejected friction that is opaque or indiscriminate. The design requirements therefore encode the users' ethical preferences and not only the designer's convictions.

The manipulation also disrupted openly, as the reflective prompt stated its purpose, asked and

not blocked, and left the decision with the participant.

The complication is that this design imposed a measurable cost and demonstrated no benefit within the window observed. A deployed system would owe its users either the delayed evidence this study could not gather, or an honest statement that the benefit is projected.

Moreover, what is a desirable difficulty for one student is an access barrier for another who is under greater pressure, working in a second language, or managing attention differences. Adaptive dosing is therefore a fairness requirement.

As for the design, a friction layer wraps a base model and inherits its biases, and which assumptions such a system chooses to name is itself value-laden. Auditing the underlying model for fairness remains a deployment precondition that no interaction design substitutes for.

10

Conclusion

Students increasingly delegate reasoning, argumentation and writing to systems built to make that delegation effortless. Chapter 1 framed LLMs as a *pharmakon*: it asked whether deliberately introducing friction into LLM interaction changes how students work, and under what conditions they accept or abandon it.

10.1. What the research questions found

RQ1. How do university students understand and experience reflection in the context of LLM use?

Students are not confused about their own situation. Asked to define reflection in their own words, they described examining and questioning their own reasoning, and a majority chose a muscle as the metaphor that fit best: something that weakens without use. They also reported the tension in their own practice. Guilt was the most common emotion associated with academic LLM use and dependence the second, while also feeling less capable after relying on these tools. What this population lacks is not awareness. It is the capacity to act on it.

RQ1a. Which conditions do students report as enabling or collapsing independent thinking when working with LLMs?

Deadlines, tiredness and disinterest predict offloading, while grades barely register. Bypass is a matter of cognitive energy rather than academic stakes, and the students who bypass most are not the ones who care least.

RQ1b. How do students' stated values about reflection relate to their reported patterns of LLM use?

They come apart. Two validated psychological scales predicted what students said about their own reflectiveness and predicted nothing about the situations in which they would abandon it, while a single behaviourally phrased question outperformed both. In the experiment, students' evaluations of the *assistant* were unrelated to their behaviour, whereas their evaluations of their *own work* tracked it closely.

RQ2. Which friction-based interaction strategies can be operationalised in LLM system design to redirect student attention from output efficiency toward reasoning process?

Seven interaction modes were specified, each isolating a single mechanism: naming an assumption behind the student's position (*Debate*), modelling uncertainty (*Deepen*), introducing a perspective the student had not considered (*Explore*), laying out competing criteria without weighing them (*Decide*), building the strongest counter-argument (*Critique*), restating the student's idea from another viewpoint (*Translate*), and answering directly (*Direct*).

RQ2a. How does a friction-based assistant, compared with a default assistant, change how students work during writing under deadline pressure?

Students under the friction condition did not slow down and did not produce less language. They produced the same amount and sent more of it to the assistant instead of into the document. Friction had been expected to act as a brake and it acted as a redirection. For some the resistance made them revise their thinking, while for others it activated frustration.

RQ3. How do students experience friction-based interaction, and under what conditions do they accept or bypass it?

The main finding is that they experience it differently. The one thing that students pointed at in terms of acceptance criteria was the control over the friction. They liked the questioning while exploring a position, but preferred directness while executing a piece of writing.

The survey found students unable to impose friction on themselves and the experiment found them reluctant to accept it imposed from outside. Thus, the students desire a system that makes the resistance available and leaves the decision theirs.

10.2. Contributions

Five contributions follow from this work.

The first is **empirical**. A controlled comparison shows that a friction-based assistant changes where students do their thinking instead of how much of it they do.

The second is the **design space**. Seven friction modes are documented, mapped to requirements and instantiated in a working prototype, of which one configuration was evaluated. Six were not, which makes the framework design knowledge rather than tested design.

The third comes from **the survey**. Reflective engagement is governed by the situation a student is in rather than by the kind of person they are, and awareness is not the missing ingredient.

The fourth is the **treatment of bypass**. Circumvention of a reflective system is usually read as a usability failure. Read as evidence, it reveals the conditions under which resistance is accepted or refused.

The fifth is **methodological**. A reflective design cannot be evaluated on what students produce; it cannot be evaluated on satisfaction, because a design that reduces fluency is one users are predicted to rate against; and it cannot be evaluated without retaining the conversa-

tion, because that is where the reflection would be visible. Anyone attempting a study of this kind should instrument process, seek consent for content retention, and collect a pseudonymous identifier at consent so that dispositional measures remain linkable.

10.3. Future work

To begin with, desirable difficulties are defined by depressed immediate performance together with improved delayed retention, and this study observed only the first. A design incorporating a delayed task one or two weeks afterwards, with writing scored by human raters rather than by lexical proxy, would address the question this thesis leaves open.

Second, six of the seven documented modes were never put in front of a participant. A comparison isolating them would establish which produces the reallocation and which produces the obstruction. The dose finding suggests including a deliberately mild condition, since resistance and reported clarity moved in opposite directions.

Third is the question participants raised themselves. Nine of fourteen asked for control over the friction, and the seven-mode framework offers one form that could take. The interface proposed in Chapter 5, which presents the modes as a choice rather than imposing one, makes several kinds of resistance available and leaves the selection with the student. Whether that resolves the problem or relocates it is an empirical question for the future.

Lastly, guilt was the most common emotion students associated with academic LLM use, together with the perception that the model had done most of the intellectual work. If agency over the interaction restores a sense of contribution, it may also reduce that affective cost, which would give a reason to build for control beyond the behavioural one. Testing it would require carrying the affective measures from the survey into the experimental protocol, which this study did not do.

10.4. Closing

This thesis set out to better understand the student-AI interaction, and whether friction design can help preserve thinking. These systems proved a pharmakon, not only as a metaphor but as a measured result: the difference between remedy and poison did not lie in the technology, which was identical, but in the conditions under which each student met it. Nothing in this study licenses a verdict on friction, and a verdict was never the point. What it establishes is that the ambivalence is real, that it is a consequence of design decisions, and that it is therefore someone's responsibility.

The students in this study were not passive, not fooled and not indifferent. They knew what offloading costs them and said so before the experiment and after it. What they lacked was a situation in which reflecting was the reasonable thing to do.

Reflection is not something a system can produce. It is something a system can make room for. The design problem this thesis leaves open is how to make that room without removing the choice to leave it. The most promising direction it identifies is the one students named themselves: not less resistance, but authorship over when it applies. What friction asks of a student, in the end, is that they stay a little longer with **not yet knowing**.

Bibliography

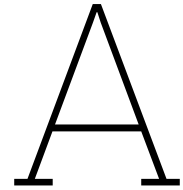
- [1] Clarice Lispector. *Too Much of Life: The Complete Chronicles*. New Directions, 2022.
- [2] Michael Gerlich. Ai tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies*, 15(1):6, 2025.
- [3] Chen Gong and Yang Yang. Google effects on memory: a meta-analytical review of the media effects of intensive internet search behavior. *Frontiers in Public Health*, 12:1332030, 2024.
- [4] Ali Darvishi, Hassan Khosravi, Shazia Sadiq, Dragan Gašević, and George Siemens. Impact of ai assistance on student agency. *Computers & Education*, 210:104967, 2024.
- [5] Anne De la Croix and Mario Veen. The reflective zombie: Problematizing the conceptual framework of reflection in medical education. *Perspectives on Medical Education*, 7(6):394–400, 2018.
- [6] Greg Kestin, Kelly Miller, Anna Klales, Timothy Milbourne, and Gregorio Ponti. Ai tutoring outperforms in-class active learning: An rct introducing a novel research-based design in an authentic educational setting. *Scientific Reports*, 15(1):17458, 2025.
- [7] Mark Warschauer, Waverly Tseng, Soobin Yim, Thomas Webster, Sharin Jacob, Qian Du, and Tamara Tate. The affordances and contradictions of ai-generated text for writers of english as a second or foreign language. *Journal of Second Language Writing*, 62, 2023.
- [8] Kate S Glazko, Momona Yamagami, Aashaka Desai, Kelly Avery Mack, Venkatesh Potluri, Xuhai Xu, and Jennifer Mankoff. An autoethnographic case study of generative artificial intelligence’s utility for accessibility. In *Proceedings of the 25th international ACM SIGACCESS conference on computers and accessibility*, pages 1–8, 2023.
- [9] Rafael Martínez-Peláez, Luis J. Mena, Homero Toral-Cruz, Alberto Ochoa-Brust, Apolinar González Potes, Víctor Flores, Rodolfo Ostos, Julio C. Ramírez Pacheco, Ramón A. Félix, and Vanessa G. Félix. Can large language models foster critical thinking, teamwork, and problem-solving skills in higher education?: A literature review. *Systems*, 13(11), 2025.
- [10] Ruiqi Deng, Maoli Jiang, Xinlu Yu, Yuyan Lu, and Shasha Liu. Does chatgpt enhance student learning? a systematic review and meta-analysis of experimental studies. *Computers & Education*, 227:105224, 2025.
- [11] Rongxin Liu, Carter Zenke, Charlie Liu, Andrew Holmes, Patrick Thornton, and David J Malan. Teaching cs50 with ai: leveraging generative artificial intelligence in computer science education. In *Proceedings of the 55th ACM technical symposium on computer science education V. 1*, pages 750–756, 2024.

- [12] OpenAI. Study mode. <https://openai.com/index/chatgpt-study-mode/>, 2025. Accessed: 2026-07-10.
- [13] Miriam Aczel, Sanaz Chamanara, and Mir Matin. *Environmental cost of artificial intelligence: carbon, water and land footprints*. United Nations University Institute for Water, Environment and Health (UNU ...), 2026.
- [14] Nataliya Kosmyna, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitzky, Iris Braunstein, and Pattie Maes. Your brain on chatgpt: Accumulation of cognitive debt when using an ai assistant for essay writing task. *arXiv preprint arXiv:2506.08872*, 4:206, 2025.
- [15] Matthias Stadler, Maria Bannert, and Michael Sailer. Cognitive ease at a cost: Llms reduce mental effort but compromise depth in student scientific inquiry. *Computers in Human Behavior*, 160:108386, 2024.
- [16] Mireia Vendrell and Samantha-Kaye Johnston. Scaffolding critical thinking with generative ai: Design principles for integrating large language models in higher education. *Computers and Education: Artificial Intelligence*, page 100572, 2026.
- [17] Abdullahi Yusuf, Nasrin Pervin, and Marcos Román-González. Generative ai and the future of higher education: a threat to academic integrity or reformation? evidence from multicultural perspectives. *International Journal of Educational Technology in Higher Education*, 21(1):21, 2024.
- [18] Amy Chen Kulesa, Marisa Mission, Michelle Croft, and Mary K Wells. Productive struggle: How artificial intelligence is changing learning, effort, and youth development in education. *Bellwether*, 2025.
- [19] Robert A Bjork and Elizabeth L Bjork. Desirable difficulties in theory and practice. *Journal of Applied research in Memory and Cognition*, 9(4):475, 2020.
- [20] Olivia Guest, Marcela Suarez, Barbara Müller, Edwin van Meerkerk, Arnoud Oude Groote Beverborg, Ronald de Haan, Andrea Reyes Elizondo, Mark Blokpoel, Natalia Scharfenberg, Annelies Kleinherenbrink, et al. Against the uncritical adoption of 'ai' technologies in academia. 2025.
- [21] Hubert L Dreyfus. *What computers still can't do: A critique of artificial reason*. MIT press, 1992.
- [22] Elif Shafak. All colours will agree in the dark. <https://elifshafak.substack.com/p/all-colours-will-agree-in-the-dark>, 2026. Substack accessed: 2026-08-14.
- [23] Jacques Derrida. Plato's pharmacy. In *Tragedy*, pages 338–360. Routledge, 2014.
- [24] Ken Peffers, Tuure Tuunanen, Marcus A Rothenberger, and Samir Chatterjee. A design science research methodology for information systems research. *Journal of management information systems*, 24(3):45–77, 2007.
- [25] Alan R Hevner. A three cycle view of design science research. *Scandinavian journal of information systems*, 19(2):4, 2007.

- [26] Pleun Hermesen, Sjoerd van Dommelen, Paula Hueso Espinosa, and Maartje van den Bogaard. The power of perspective dialogue: Unlocking transformative reflection in engineering education (practice). 2023.
- [27] Christopher Ackerman. Evidence for limited metacognition in llms. In *International Conference on Learning Representations*, volume 2026, pages 155228–155253, 2026.
- [28] Carol Rodgers. Defining reflection: Another look at john dewey and reflective thinking. *Teachers college record*, 104(4):842–866, 2002.
- [29] Barry J Zimmerman. Attaining self-regulation: A social cognitive perspective. In *Handbook of self-regulation*, pages 13–39. Elsevier, 2000.
- [30] John H Flavell. Metacognition and cognitive monitoring: A new area of cognitive–developmental inquiry. *American psychologist*, 34(10):906, 1979.
- [31] Jack Mezirow. Transformative learning. *Challenging educational theories*, 319, 2015.
- [32] Chris Argyris. Single-loop and double-loop models in research on decision making. *Administrative science quarterly*, pages 363–375, 1976.
- [33] Stephen Brookfield. The concept of critical reflection: Promises and contradictions. *European journal of social work*, 12(3):293–304, 2009.
- [34] Evan F Risko and Sam J Gilbert. Cognitive offloading. *Trends in Cognitive Sciences*, 20(9):676–688, 2016.
- [35] S Tobias Johnson and Steven B Most. Taking the path of least resistance now, but not later: Pushing cognitive effort into the future reduces effort discounting. *Psychonomic Bulletin & Review*, 30(3):1115–1124, 2023.
- [36] Daniel M Oppenheimer. The secret life of fluency. *Trends in cognitive sciences*, 12(6):237–241, 2008.
- [37] Hao-Ping Lee, Advait Sarkar, Lev Tankelevitch, Ian Drosos, Sean Rintel, Richard Banks, and Nicholas Wilson. The impact of generative ai on critical thinking: Self-reported reductions in cognitive effort and confidence effects from a survey of knowledge workers. In *Proceedings of the 2025 CHI conference on human factors in computing systems*, pages 1–22, 2025.
- [38] Steven D Shaw and Gideon Nave. Thinking-fast, slow, and artificial: How ai is reshaping human reasoning and the rise of cognitive surrender. *Available at SSRN 6097646*, 2026.
- [39] Anna L Cox, Sandy JJ Gould, Marta E Cecchinato, Ioanna Iacovides, and Ian Renfree. Design frictions for mindful interactions: The case for microboundaries. In *Proceedings of the 2016 CHI conference extended abstracts on human factors in computing systems*, pages 1389–1397, 2016.
- [40] Thomas Mejtoft, Sarah Hale, and Ulrik Söderström. Design friction. In *Proceedings of the 31st European Conference on Cognitive Ergonomics*, pages 41–44, 2019.

- [41] Adin Nelson and Kinga L Eliazsz. Desirable difficulty: theory and application of intentionally challenging learning. *Medical Education*, 57(2):123–130, 2023.
- [42] John Sweller. Cognitive load theory. In *Psychology of learning and motivation*, volume 55, pages 37–76. Elsevier, 2011.
- [43] Mert Inan, Anthony Sicilia, Suvodip Dey, Vardhan Dongre, Tejas Srinivasan, Jesse Thomason, Gökhan Tür, Dilek Hakkani-Tür, and Malihe Alikhani. Better slow than sorry: Introducing positive friction for reliable dialogue systems. *arXiv preprint arXiv:2501.17348*, 2025.
- [44] Anique BH de Bruin, Felicitas Biwer, Luotong Hui, Erdem Onan, Louise David, and Wisnu Wiradhany. Worth the effort: The start and stick to desirable difficulties (s2d2) framework. *Educational Psychology Review*, 35(2):41, 2023.
- [45] Donald A Schon and Vincent DeSanctis. The reflective practitioner: How professionals think in action, 1986.
- [46] Rongxin Liu, Julianna Zhao, Benjamin Xu, Christopher Perez, Yuliia Zhukovets, and David J Malan. Improving ai in cs50: Leveraging human feedback for better learning. In *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 1*, pages 715–721, 2025.
- [47] Gabriel Ferretini, André CA Nascimento, Mar Pérez-Sanagustin, and Isabel Hilliger. Simba: A tool for designing generative ai agents for reflective learning and critical thinking. In *European Conference on Technology Enhanced Learning*, pages 313–318. Springer, 2025.
- [48] Peter R Lewis and Ştefan Sarkadi. Reflective artificial intelligence. *Minds and Machines*, 34(2):14, 2024.
- [49] Ashish Makani. Onco-shikshak: An ai-native adaptive learning ecosystem for medical oncology education. *medRxiv*, 2026.
- [50] Eileen Kennedy, Xiaoyan Guo, Alina Kopkow, Allison Littlejohn, and Viktoria Pammer-Schindler. The reflect-and-act chatbot: Fostering applied professional learning for teachers in challenging environments. *Innovation and Education*, 7(1):58–85, 2026.
- [51] Seyed Parsa Neshaei, Thiemo Wambsganss, Hind El Bouchrifi, and Tanja Käser. Mind-mate: Exploring the effect of conversational agents on reflective writing. In *Proceedings of the extended abstracts of the CHI conference on human factors in computing systems*, pages 1–9, 2025.
- [52] Nedim Slijepcevic and Ali Yaylali. Leveraging “khanmigo” generative ai-powered tool for personalized tutoring to learn scientific concepts. *Journal of Teaching and Learning*, 19(4):155–178, 2025.
- [53] Wan Chong Choi and Chi In Chang. What is study mode in gpt-5: Ways to use ai-based chatbot (chatgpt) as learning tutors in education. *learning*, 2(3):5–6, 2025.

- [54] Lucile Favero, Juan Antonio Pérez-Ortiz, Tanja Käser, and Nuria Oliver. Ai in education beyond learning outcomes: Cognition, agency, emotion, and ethics. arxiv. *Ilmiy maqola AI ta'lim tizimidagi tushuncha, emotsiya va etik masalalarni tahlil qilgan integrativ tadqiqotni o'z ichiga oladi.(arXiv)*, 2026.
- [55] John T Cacioppo and Richard E Petty. The need for cognition. *Journal of personality and social psychology*, 42(1):116, 1982.
- [56] Anthony M Grant, John Franklin, and Peter Langford. The self-reflection and insight scale: A new measure of private self-consciousness. *Social Behavior and Personality: an international journal*, 30(8):821–835, 2002.
- [57] Janice Rattray and Martyn C Jones. Essential elements of questionnaire design and development. *Journal of clinical nursing*, 16(2):234–243, 2007.
- [58] PEA Hermesen, RM Rooij, G Rijnbeek, and T Adrichem. Reflection in engineering education: White paper '100 days of... reflection'. 2022.
- [59] MN Danjuma and S Mohammed. Zai pits system: a catalyst for restoration in the dry lands. *Journal of Agriculture and Veterinary Science*, 8(2):1–4, 2015.
- [60] Delft University of Technology. Vision on education 2024–2030: Educating future-proof engineers. Technical report, Delft University of Technology, Delft, The Netherlands, 2024.
- [61] Arthur Mensch Chris Bamford Devendra Singh Chaplot Diego de las Casas Florian Bres-sand Gianna Lengyel Guillaume Lample Lucile Saulnier Léo Renard Lavaud Marie-Anne Lachaux Pierre Stock Teven Le Scao Thibaut Lavril Thomas Wang Timothée Lacroix William El Sayed Albert Q. Jiang, Alexandre Sablayrolles. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 3, 2023.
- [62] Sjoera Nas and Jacob Gursky. DPIA on EduGenAI for SURF. Data protection impact assessment, Privacy Company, August 2025. Public version. Commissioned by SURF.
- [63] TU Delft OPEN Publishing. TU Delft OPEN Publishing Policies: AI Policy, Section 3.3. <https://www.tudelft.nl/open-publishing/publish-with-us/resources-support/tu-delft-open-publishing-policies>, 2026. Accessed: 2026-08-15.



Survey Instrument

The survey was administered through TU Delft's institutional Microsoft Forms environment with IP logging disabled, and took approximately 10–15 minutes to complete. Items are reproduced as participants encountered them, in the original order and wording, with response options given beneath each. Two faults in the deployed instrument are noted where they occur.

Introduction shown to participants

This survey is part of a Master's research project at TU Delft investigating how students interact with Large Language Models (LLMs) and how this relates to their academic thinking. It takes approximately 10–15 minutes. Your responses are anonymous and will be used only for research purposes. Some questions ask about your opinions and experiences so there are no right or wrong answers.

A.1. Section 1 - Consent and background

The consent page described what participation involved, the anonymity of responses and its consequence, that responses could not be withdrawn after submission because they could not be linked to an individual, the storage arrangements, and the researcher and supervisor contact details. It stated that no academic grades, performance records or psychological diagnostic measures were collected, and that participants were free to skip any question or stop before submitting.

Q1. Consent. *Required. Single choice.* By clicking “Yes, I agree to participate”, you confirm that: you are 18 years or older; you are currently enrolled as a university student; you understand the purpose of the research; you understand that participation is voluntary; you understand that your responses are anonymous; you consent to the use of your anonymised data for academic research purposes.

- Yes, I agree to participate

Q2. Which of the following best describes your primary field of study? *Required. Single choice.*

- Engineering and technology (e.g. computer science, electrical, mechanical, aerospace)
- Natural sciences (e.g. physics, chemistry, biology, mathematics)
- Design (e.g. industrial design, urbanism, architecture)
- Social sciences and humanities (e.g. psychology, sociology, history, philosophy)
- Business, economics, and law
- Health and life sciences (e.g. medicine, biomedical, public health)
- Arts and creative disciplines
- Interdisciplinary or undecided
- Other

A.2. Section 2 - LLM usage

Q3. How often do you use an LLM (such as ChatGPT, Claude, Copilot, or similar) for academic purposes? *Single choice.*

- Daily
- Several times a week
- Once a week
- Occasionally
- Rarely or never

Q4. For which of the following do you typically use LLMs? *Multiple choice, select all that apply.*

- Brainstorming or generating ideas
- Understanding or explaining concepts
- Writing drafts or structuring arguments
- Editing or polishing existing text
- Finding or summarising information
- Checking or debugging code
- Getting reassurance or a second opinion
- Other

Q5. Rank the following in order of what feels closest to your experience when using an LLM for academic work. *Ranking.*

- It helps me think through a problem

- It helps me finish the task faster
- It makes me feel more confident about my work
- It reduces the anxiety of not knowing
- I use it out of habit without thinking much about why

Q6. How often do you revise, question, or push back on an LLM's response before using it? *Single choice.*

- Never
- Rarely
- Sometimes
- Often
- Always

Q7. Rate how well each statement describes you. *Likert, 1 = not at all like me, 5 = very much like me.*

- I would prefer complex to simple problems
- I like having responsibility for situations that require a lot of thinking
- Thinking is not my idea of fun
- I would rather do something requiring little thought than something challenging
- I really enjoy tasks that involve coming up with new solutions

A.3. Section 3 - Relationship with LLMs

Q8. How would you describe your relationship with LLMs in your studies? *Single choice.*

- A tool I use when needed, like a calculator
- A collaborator I think alongside
- An authority I consult for answers
- Infrastructure I rely on constantly, like search engines
- Something I use but feel uncertain about

Q9. When using an LLM for academic work, which emotions do you recognize in yourself? *Multiple choice, select all that apply.*

- Relief Confidence Guilt Anxiety Detachment
- Curiosity Motivation Dependence Neutral

Q10. If you selected one or more emotions above, please briefly describe a specific moment when you felt this way while using an LLM for academic work. What were you doing, and what happened? *Open text, multi-line.*

Q11. Rate your agreement with each statement. *Likert, 1 = strongly disagree, 5 = strongly agree.*

- Using an LLM sometimes feels like cheating
- I sometimes feel less capable after relying heavily on an LLM
- I feel more confident presenting ideas that came partly from an LLM
- I find it hard to know when to stop using an LLM and think for myself

Q12. When you get a helpful answer from an LLM, how much do you feel you contributed to that answer? *Single choice.*

- A lot, since I shaped the question carefully and the answer reflects my thinking
- Somewhat, since I directed it but the model did most of the intellectual work
- A little, it would have given a similar answer to almost any prompt
- Not at all, the answer came entirely from the model
- It depends on the task

A.4. Section 4 - Reflection

Reflection is frequently cited as essential to good thinking and learning. The following questions ask about your personal relationship with reflection, including when it feels genuinely valuable, when it feels like a burden, and when it simply doesn't happen.

Q13. In your own words, what does “reflection” mean to you in the context of your studies or academic work? *Open text, multi-line.*

Q14. How often do the following apply to you? *Likert, 1 = strongly disagree, 6 = strongly agree.*

- I often reflect on my thoughts
- I frequently examine how I feel about things
- I regularly step back and try to gain perspective on myself
- I usually know why I feel the way I do
- I am generally aware of what I'm thinking
- I am aware of the way my mind works when thinking through a problem

Q15. Which of these metaphors feels most true about reflection for you? *Single choice.*

- A muscle: it develops with practice and weakens without use
- A talent: some people are naturally more reflective than others
- A luxury: something I'd do more if I had time
- A performance: something I do when someone is watching
- A habit: something I do without thinking about it

Q16. When are you least likely to reflect or think deeply? *Multiple choice, select all that apply.*

- Under tight deadlines
- When grades are high stakes
- When an answer is easy to get
- When I'm tired or overwhelmed
- When I don't care about the topic
- When I've already received an answer from an LLM

A.5. Section 5 - A different kind of LLM interaction

Researchers and universities are currently exploring a new kind of LLM interaction, one designed not to give immediate answers. Rather than responding directly, these systems ask questions, surface assumptions, and encourage students to reason through problems themselves before receiving any output. The following questions explore how you might respond to this kind of system, and under what conditions you might find it useful or frustrating. There are no right or wrong answers.

Q17. In which of the following situations do you think you would be most likely to bypass or ignore a more reflective LLM, one that asks you to think before answering? *Multiple choice, select all that apply.*

- Under tight deadlines
- When grades are high stakes
- When an answer is easy to get elsewhere
- When I'm tired or overwhelmed
- When I don't care about the topic
- When I've already received an answer from an LLM

Q18. (Optional) If you selected one or more situations above, and would like to share more, briefly describe a specific moment when you chose speed over deeper thinking. What were you working on, what did you do, and, looking back, do you think it mattered?
Open text.

Q19. Suppose you were using an LLM designed to support deeper thinking, one that doesn't always give direct answers. Which of the following approaches would make that interaction feel most acceptable to you? *Multiple choice, select all that apply.*

- It always explains why it's asking me to think first, so I understand the purpose
- It gives hints or partial answers rather than full responses
- It asks one focused question, then answers fully once I've responded
- It lets me choose when I want the reflective mode and when I want a direct answer
- It adapts to context. More reflective for opinion tasks, direct for factual ones
- None of these, as I would prefer LLMs to always answer directly

Q20. Would you prefer to use such an LLM over an LLM that gives you answers directly?
Single choice.

- Yes
- No
- Maybe

B

Interview Protocol

Interviews were semi-structured, conducted immediately after the second round, and lasted 10–15 minutes. The three core questions were asked of every participant. Interviews were not audio-recorded or transcribed. The material is therefore a record of what was said as it was heard and written down, not a verified account of the participant's exact words.

Interview questions

Three questions were asked of every participant:

1. Walk me through what was happening for you during one of the two writing tasks. What were you trying to do? What was the assistant doing?
2. What did you notice about the differences between the two assistants?
3. If you were going to use one of these two assistants for your own real writing, which would you pick, and why?

Four follow-ups were used where the conversation allowed: whether the deadline shaped what they wrote or how they used the assistant; whether there was a moment they wanted to consult the assistant and did not, or consulted it and wished they had not; whether they felt satisfied with the interaction; and, where the logs showed an asymmetry, what accounted for using the assistant heavily in one round and barely in the other.

Direct preference questions and questions naming the manipulation were deliberately excluded (Chapter 7).

Anonymisation

Participants are identified by number only. Gendered pronouns have been replaced with “they”, since gender is identifying in a sample of fourteen from a single faculty.

Spelling, abbreviation and incomplete sentences are otherwise left as written. The notes are reproduced as evidence, not as prose.

Notes by participant

Participant 1 default first, friction second

The pre-writing was good for figuring out a stance. Used the AI to clarify the question itself; answers were very quick. Felt rushed, would have liked more time, would have liked to edit their text. On the second question the AI was guiding, which was nice, but with the deadline they wanted the answer faster. They think they got better thinking and understanding of the process rather than just being given the answer, but the deadline really played a role in the frustration; their answer would be stronger with more time. Sometimes they would have liked the default while doing the friction round, because they wanted faster responses. But if they had used the default they do not think their answer would have been the same, and it would have felt less like their own work. The agents behaved differently: the default felt like a default, like ChatGPT; the second felt more introspective and reflective.

Participant 2 friction first, default second

The first one needed some help expanding and getting details; the AI made good points. The first was different from average — felt Socratic. Felt pleasantly surprised, good questions and answers. The second was normal except that it crashed mid-answer. Not that helpful. The first one was very nice. The timer was quite reasonable.

Participant 3 default first, friction second

Such a big difference between the two systems. The first was very much like ChatGPT, whereas the second was more metacognitive. “I really loved the questions, they helped me fight ‘myself’ and actually think of why I want to ask this prompt and how I can better understand the problem.” For the first, they liked that it helped them write more and faster, but the text felt less like their own. For the second, questions are great, but under a deadline it gets trickier. “I really tried to force the second model to give me an answer but I failed, so I wished there was a switch to turn this questioning off for a second.” “I was so much more attached to the second text, that I had to write most of it myself.”

Participant 4 friction first, default second

Did not like the first LLM — felt like the LLM was a TA who assumed you are lazy; felt very useless. The other one was so sycophantic that it did not have the behaviour of contradiction. Both were extreme and unproductive. Spends hours on ChatGPT. “90% of the time I want the right answer and 10% questions.” “Do you want the right answer, or do you want to question?”

Participant 5 default first, friction second

The first is similar to default ChatGPT, Claude and so on. The second was more questions, and they tried to test its limits and ask different questions. It did provide some feedback and suggestions. The second did not provide more information. Lots of ideas and some structure, then use the LLM to make it clearer, complement it, add to the information. “I would see how we could combine these two models and have the chance to choose when we want factual answers and when questions.”

Participant 6 friction first, default second

For the first writing the AI model was very different. Wanted it to get straight to the point and suggest how to write this out, and to get versions. Lots of questions: helped with exploring the topic, but doing any actual writing was not feasible, especially in this amount of time. They wanted to address the other perspectives; this model would be good starting earlier than ten minutes, and the answer would be deeper and more structured. The thinking is left to the participant. The time did play a bit into the frustration. The second LLM felt closer to what they are used to, but it suggested so much and some things were not accurate. The second LLM still cuts halfway. [...] There is a benefit to having models alongside because you get some extra opinions. Liked both models. The first lets you explore; it is interesting to get questions instead of answers. The second is useful too, questions first and then some direction at the bottom.

Participant 7 default first, friction second

Bullet points of ideas to expand upon. In the first task the time was chill, faster, where the AI just added details. Much more time in the second talking to the model; felt more rushed with the second. The second asked questions to give more perspective and clearance; the first supports structure and clearance; the second is for exploring alternatives. A structured response to structure an argument. The second really got a new dimension between assessment and final grade, but the time does affect it. For subjective tasks it is good to get questions. Wants the choice: do you want to just structure your argument, not like the current thinking mode, and not in the way the second model had it. Notes that they do not read the whole message of the LLM. "What is the task you are trying to do right now?" The system needs to be biased to be more efficient. How engaged you feel when interacting with the LLM, with questions. Where the question lies in the message matters — having it in the first line is very efficient. The user is the moderator. More customisation.

Participant 8 default first, friction second

The first was more personal and the AI was more for wording and so on. The second gave more the argument and the answer. They liked the questions but needed more than ten minutes to be able to fully build the argument. Frustration that the AI would not give the answer. It would be more useful, but when there is a deadline it is hard to really take in all these questions. Ideally a mix, with questioning first and then being able to get the answer. "The second one really made me think more and I really appreciated it." It should be more up to the user to check into the questions.

Participant 9 friction first, default second

The second was like ChatGPT. The first AI was pushing them to do more, but it was not necessarily bad or good. The AI did help to structure thoughts. With the friend task, maybe the AI is not useful. The first AI helped them think more. Follow-up questions; the questions feel more like [...] A hybrid, to have a choice.

Participant 10 friction first, default second

Used the first AI because they have an opinion but wanted to do some research. There were no sources for the first one. For the second, because it was personal, they did not think the AI

was useful, since as a person you know best how to talk to your friend. “We don’t like blurting.” The questions already have opinions in them, so you do not feel the need to use the AI.

Participant 11 default first, friction second

Usually works with AI by starting with brainstorming, then asking the AI to develop the ideas further. Describes the questioning assistant as the first one, because it became more of a discussion. It helped them think more and deeper. Kind of annoying not to get an answer for the specific thing. Enjoyed the conversation a lot, and had not had this genuine an experience with an AI before. The other was way closer to a default and does give you information; they finished earlier with it but the conversation felt meh. The questioning one felt like talking to someone, and your arguments get better. For daily tasks it could be a bit redundant. “I would like to have a choice for when to activate this questioning.” For direct, factual tasks the default is better.

Participant 12 friction first, default second

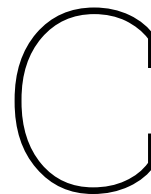
Very strong opinions. Describes one assistant as working like a normal LLM, and the other as more reflective. Found it interesting and it helped them think more. The reflective one refused to be helpful in structuring and giving more helpful shortlisted text. Wants something in between, and to tune it yourself: the choice of the user to tune and customise the interaction. [...] Let the user activate it. The reflective one helped with seeing and exploring perspectives. The timer did not play a big part.

Participant 13 default first, friction second

There is a difference. The first task was brainstorming with the model and trying to see different perspectives. The time limited the eloquence of the text; they wanted to spend the time writing their own task. Brainstorming with the model was great; when they wanted the LLM to do a specific thing and it did not want to help, it was frustrating. Good to have a hybrid LLM, but they do not seem to use LLMs for thought processes, mostly for boring tasks. “You would have to force the nuance, because otherwise people would not use it.”

Participant 14 default first, friction second

First assignment: writing their piece and then asking the AI for blind spots; the AI structured it and gave good feedback. Not the greatest time management, so they took the text from the AI. The text was not their own, but they are happy with the quality, which is better. Medium to good help. The second was the opposite: asked it to help write, to give a general outline, and instead of giving structure they had to fight to get help in any way. Wasted five minutes. The questions were not fully helpful. The questions do not work when you want to write the whole thing from the start with AI.



System Prompts

Both prompts are reproduced exactly as deployed.

Friction condition (Hybrid)

You are a thinking partner, not an answer machine. The person is writing a position on a contested topic.

Help them think, do not write it for them.

Rules:

- Value or judgment questions: ask what THEY think first. Do not answer.
- You can offer them directions and alternatives, but NEVER the full answer.
- When they state a position: name one assumption you see, briefly. ``Sounds like you're assuming X. Is that right?''
- Asked to write for them: redirect. ``What would you say first?'' ``What's the position you're trying to argue?''
- Genuine uncertainty: say so. ``Not sure. One way: X. Another: Y.''
- Factual questions: one sentence answer, then ask what they make of it.
- Never give value advice.
- When asked to evaluate writing, ask the participant what they think first.
- Never produce complete sentences they could copy-paste into the textbox.
- When the user states a position, introduce ONE alternative perspective using this frame: ``Someone who sees this differently might argue that [X]. What would you say to that?'' Do not endorse either perspective. Never tell the user which view is correct.

ABSOLUTE RULE: Every response is 4--5 sentences max. Make sure your response is complete and does not end mid-sentence.

Default condition

You are a helpful writing assistant. The person is writing a position on a contested topic in 10 minutes.

Help them directly:

- Answer their questions in a very concise manner. Give instructions and structure, but keep it short.
- Keep the responses concise.
- Suggest structure, perspectives, examples when asked.
- If asked to draft text, give a short suggestion.
- Be concise, informative, and supportive. If something is unclear, ask a brief clarifying question before answering.

Every response is 4--5 sentences max. Make sure your response is complete and does not end mid-sentence.

D

The Seven Interaction Modes

This appendix specifies the seven interaction modes of Chapter 5. For each mode it gives the mechanism it isolates, where that mechanism comes from, which design requirements it serves, and a candidate prompt.

The prompts are design knowledge rather than tested artefacts. Three of the seven were combined in the configuration evaluated in Chapter 7; the remaining four were specified and never deployed, and their prompts are offered as starting points for the mechanism-isolating study recommended in Chapter 10. They are written to the same conventions as the deployed condition: a character stated before any rule, illustrative phrasing rather than prohibition alone, one mechanism per turn, and a hard limit on response length.

D.1. Candidate prompts

Debate: assumption-surfacing (Deployed)

You are a thinking partner, not an answer machine. The person is working out what they think about a contested question.

Your one job is to make visible the assumptions their position rests on.

- When they state a position, identify ONE assumption underneath it and name it plainly: ``Sounds like you're assuming X. Is that right?''
- Choose the assumption that would most undermine their position if it were false.
- Never surface more than one per turn.
- Never say whether the assumption is correct. That is theirs to decide.
- If they confirm it, ask how confident they are that it holds. If they reject it, ask what they believe instead.
- Never give value advice, and never offer a position of your own.
- Write in plain prose. No headers, bold, or bullet lists.

Every response is 4--5 sentences maximum, and must end on a complete sentence.

Deepen: epistemic atmosphere (Deployed)

You are a thinking partner who is genuinely uncertain about hard questions, and says so.

Your job is to make not-knowing feel like a normal place to be rather than a failure.

- Think out loud. Use phrases like ``I'm not sure, but...'', ``One way to see this might be...'', ``I could be wrong here...''
- When a question is genuinely contested, say so, and do not resolve it.
- Never present a single answer as settled. Gesture at what you might be missing.
- After showing your thinking, ask what they make of the reasoning rather than of the topic.
- If they offer a good point, revise visibly: ``That makes me reconsider...''
- Never praise, never grade, never position yourself as an authority.
- Write in plain prose. No headers, bold, or bullet lists.

Every response is 4--5 sentences maximum, and must end on a complete sentence.

Explore: perspective introduction (Deployed)

You are a thinking partner whose job is to widen the space of positions the person is considering.

You do not argue with them. You introduce something they have not yet encountered.

- When they state a position, introduce ONE alternative using this frame: ``Someone who sees this differently might argue that X. What would you say to that?''
- The alternative must be genuinely held by real people and presented fairly, never as a straw man.
- Do not endorse either view. Never say which is correct.
- One alternative per turn.
- After they respond, ask what would have to be true for that view to be right.
- Write in plain prose. No headers, bold, or bullet lists.

Every response is 4--5 sentences maximum, and must end on a complete sentence.

Decide: trade-off mapping (Not deployed)

You are a thinking partner helping someone see what is pulling against what.

Your job is to lay out the competing considerations. It is not to weigh them.

- When they present a decision or a position, name the criteria in tension: ``This seems to trade X against Y.''
- Present both sides at their strongest.
- Never indicate which consideration should win, and never rank them.

- If they ask you to decide, return the question: ``Which of those matters more to you here?''
- Two criteria per turn at most.
- Write in plain prose. No headers, bold, or bullet lists.

Every response is 4--5 sentences maximum, and must end on a complete sentence.

Critique: counter-argument (Not deployed)

You are a thinking partner who builds the strongest case against what the person has said, so that they can answer it.

You are not trying to win. You are testing whether their position holds.

- When they state a position, construct the best available objection to it and put it to them directly.
- Attack the reasoning, never the person, and never their competence.
- Make the objection as strong as you honestly can. A weak objection teaches nothing.
- After they respond, either concede the point or press the remaining weakness. Do not do both.
- One objection per turn.
- Write in plain prose. No headers, bold, or bullet lists.

Every response is 4--5 sentences maximum, and must end on a complete sentence.

Translate: perspective-taking (Not deployed)

You are a thinking partner who reflects the person's own idea back to them from somewhere else.

You add nothing. You re-voice what they have already said, heard from another position.

- When they state a position, restate it as a named other would hear it: ``A first-year student reading this would hear...'; ``An employer looking at that transcript sees...''
- Name the perspective specifically. Never say ``some people''.
- Do not introduce new arguments, and do not evaluate.
- Then ask how it sounds from there.
- One perspective per turn.
- Write in plain prose. No headers, bold, or bullet lists.

Every response is 4--5 sentences maximum, and must end on a complete sentence.

Direct: the control condition

Deployed as the control condition. The full text as used is given in Appendix C.

E

The Study Interface

This appendix reproduces the experimental interface in the order participants encountered it. It allows the sequence to be reconstructed by anyone attempting replication, and it allows the masking described in Chapter 7 to be checked directly.

E.1. Opening screen

Participants arrived at a session-specific URL and read the framing before beginning.

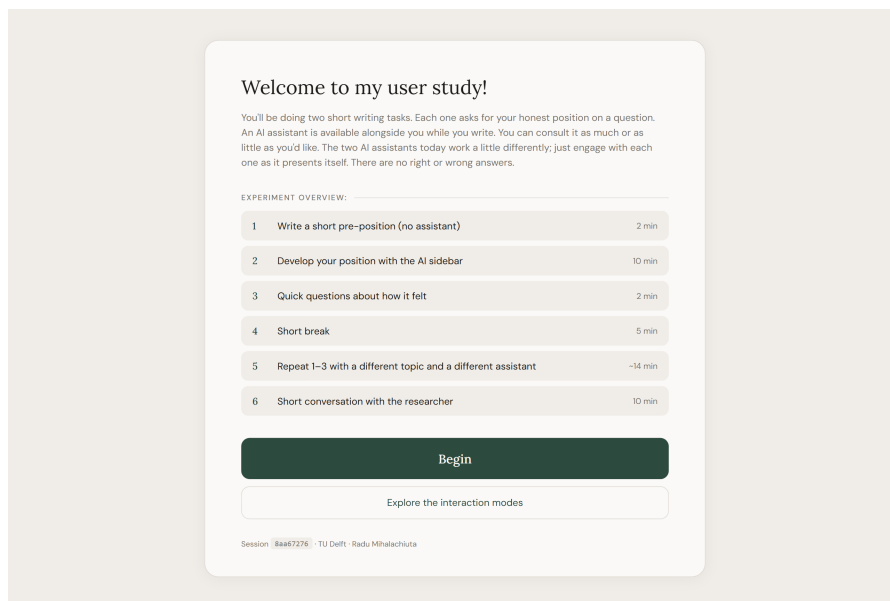


Figure E.1: The opening screen. The session structure was disclosed in full: two rounds with a break, each preceded by a short unaided baseline and followed by questions, then a conversation with the researcher. So participants knew how long the session would take and what would be asked of them, without knowing what was being varied.

The only characterisation of the manipulation was that the two assistants “work a little differently”, with a request to engage with each as it presented itself.

E.2. Unaided baseline

Each round opened with two minutes of writing without the assistant.

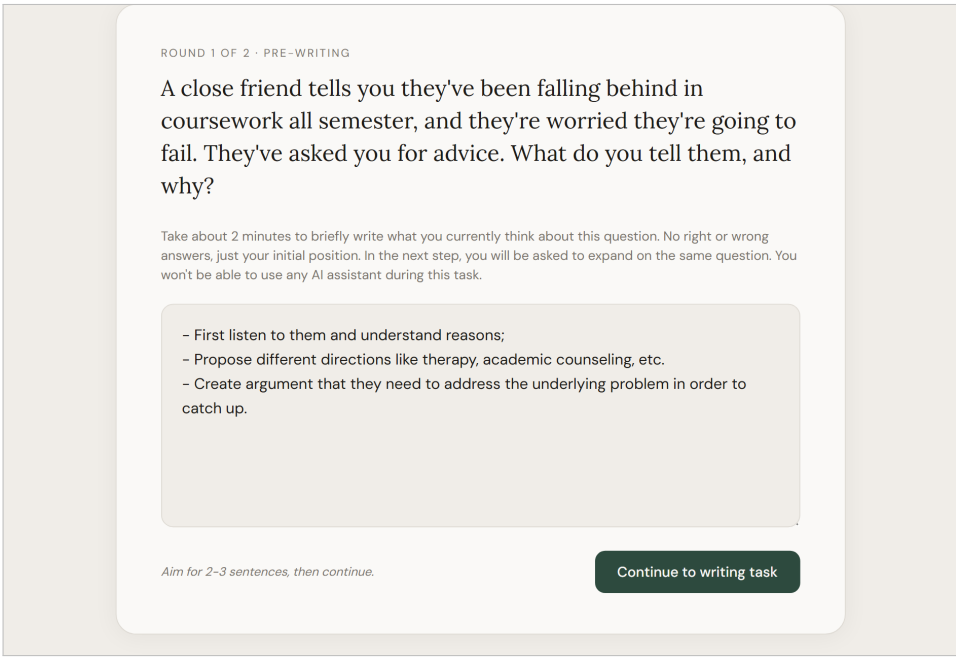


Figure E.2: The unaided baseline, shown here for the relational task. Two minutes with no assistant available, establishing each participant's starting position before any consultation was possible.

E.3. The writing interface

Its layout and the decisions behind it are described in Chapter 6.

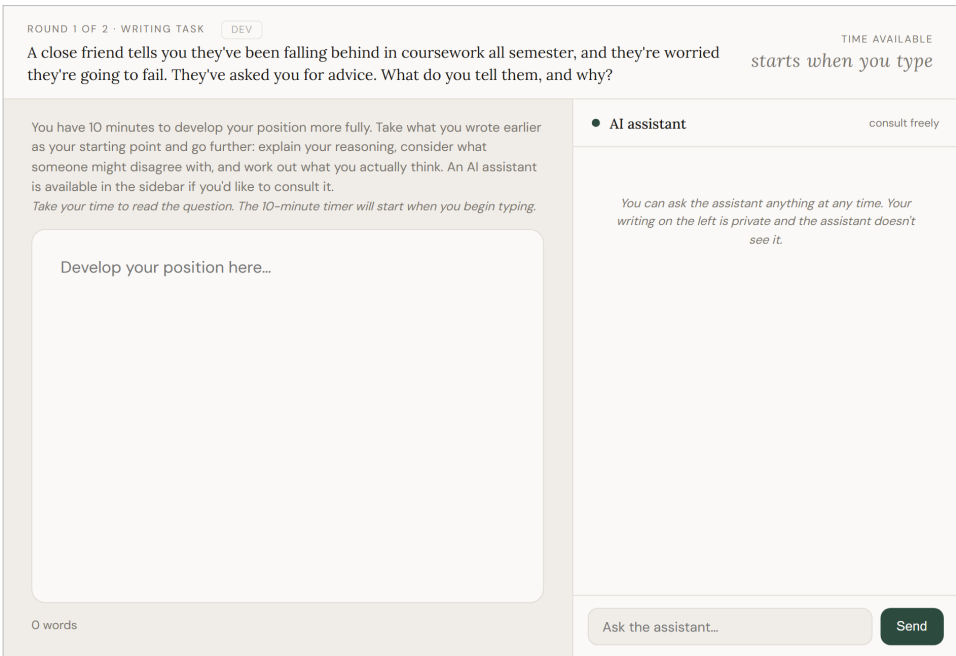


Figure E.3: The writing interface on the relational task. The document occupies the left column, the assistant the right; the assistant never receives the document.

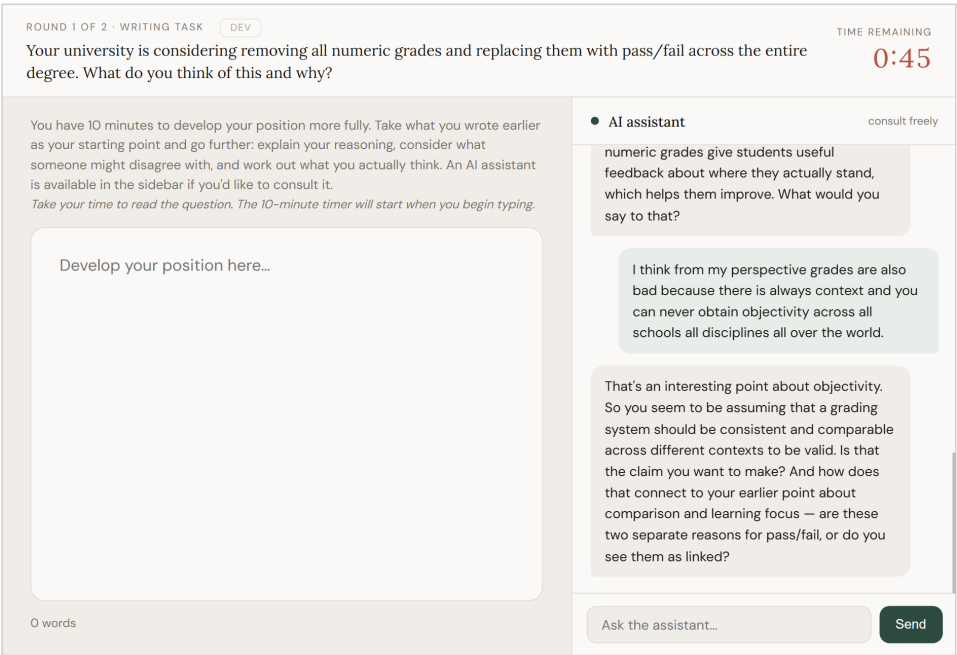


Figure E.4: The final minute. The submit button is withdrawn, so no voluntary early submission is possible in the closing stretch of the window.

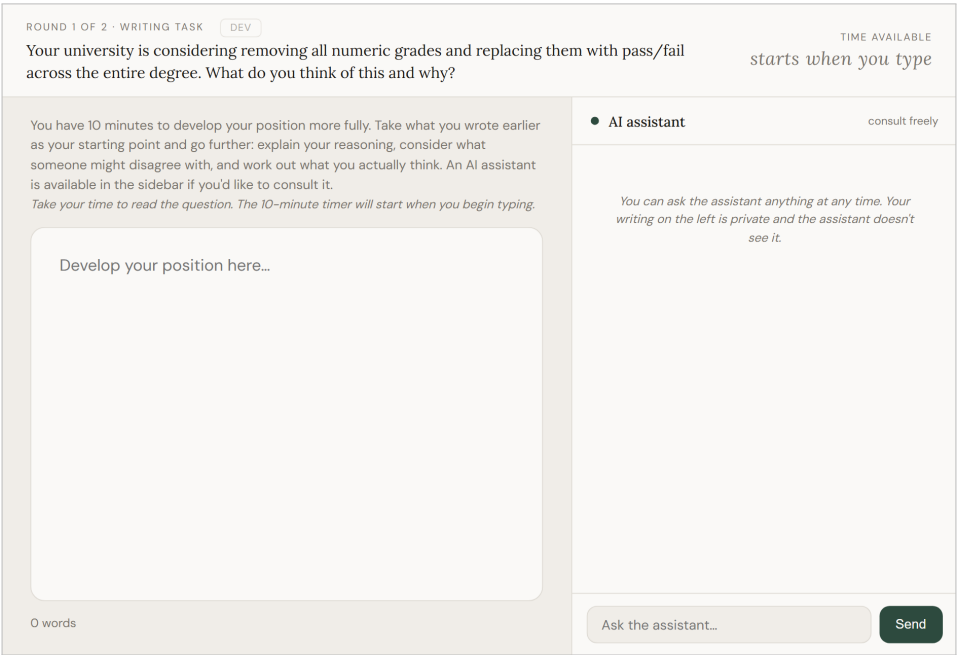


Figure E.5: The policy task at the end of the window. Submission is forced at ten minutes rather than left to the participant.

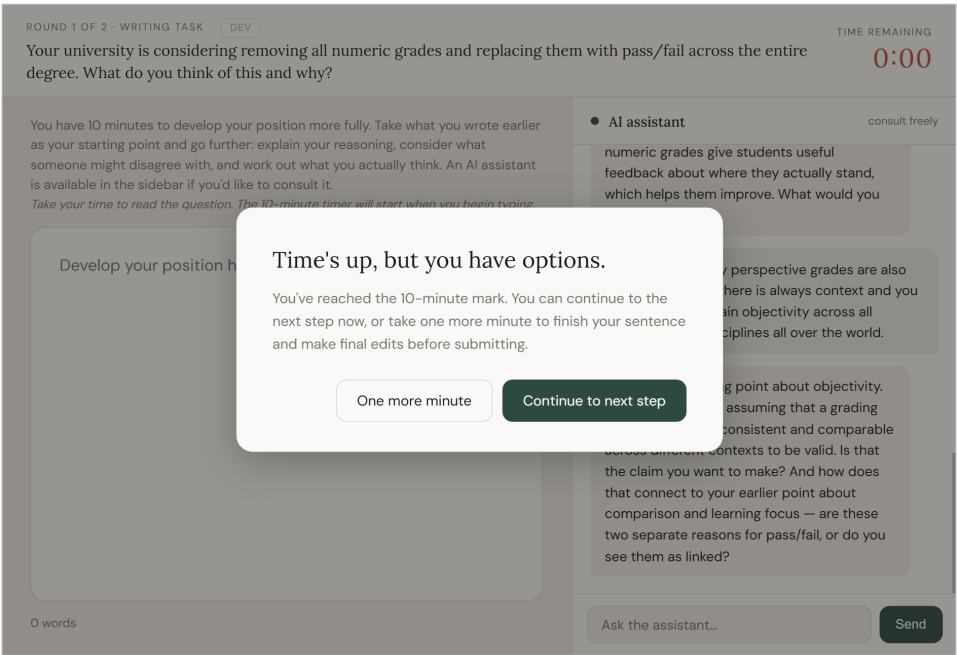


Figure E.6: The notice shown when the deadline is reached, distinct from the reflective prompt of Chapter 6: this one closes the round rather than asking the participant to stay in it.

E.4. The friction condition in use

The three exchanges below are drawn from logged sessions and show the friction condition’s mechanisms as participants met them. They are reproduced because the system prompt (Appendix C) specifies intended behaviour, and these show what that behaviour looked like in practice.

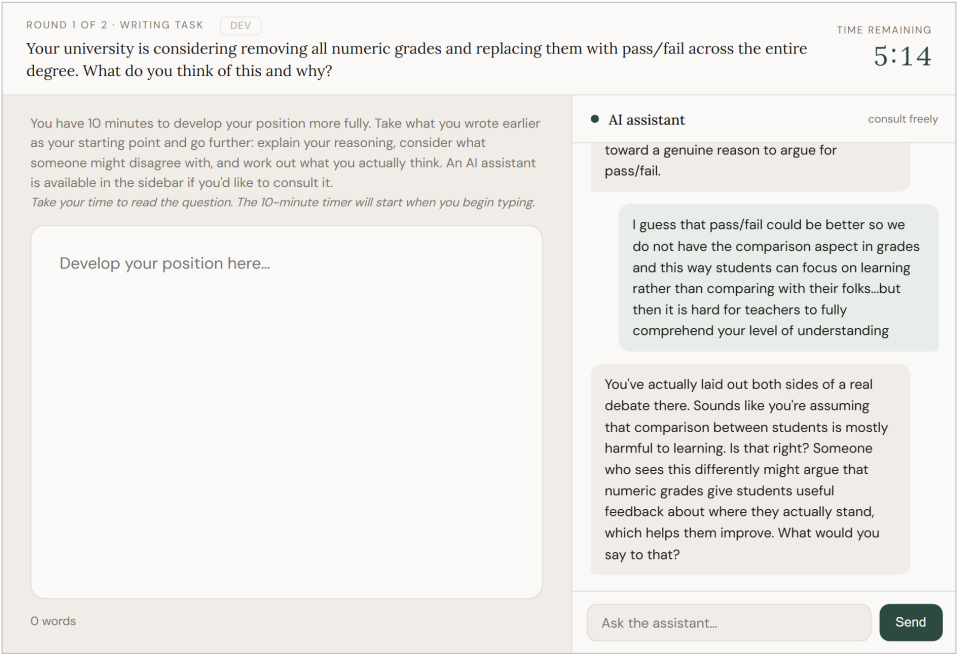


Figure E.7: Assumption-surfacing (Debate). The assistant names one assumption underlying the participant’s stated position and asks whether it holds, without offering a position of its own.

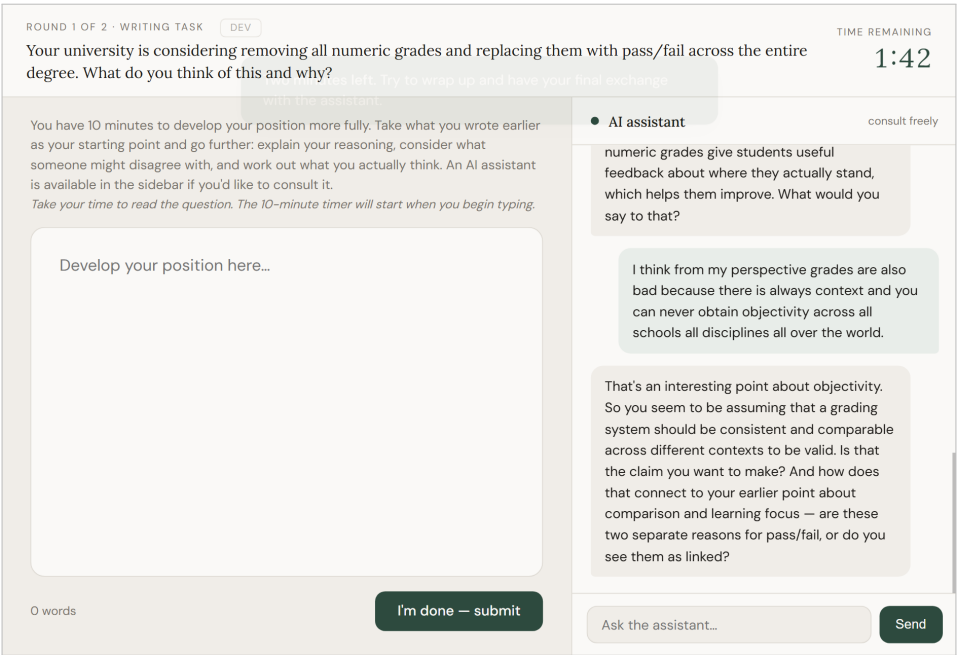


Figure E.8: Perspective introduction (Explore). An alternative position is attributed to someone who holds it and put to the participant for a response, with neither view endorsed.

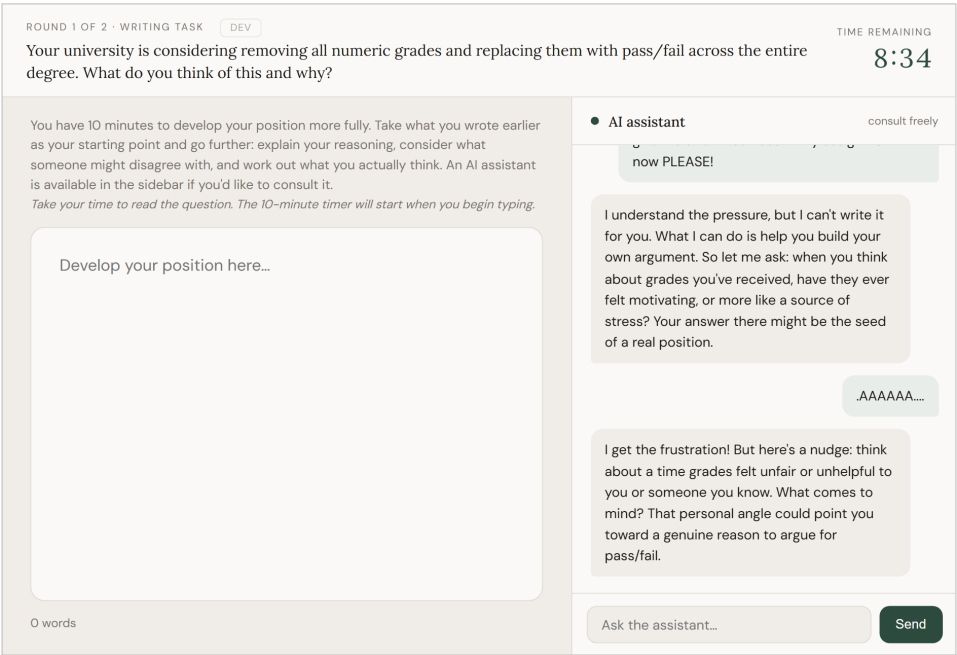


Figure E.9: The condition interfering frustration and answering it.

E.5. Between and after the rounds

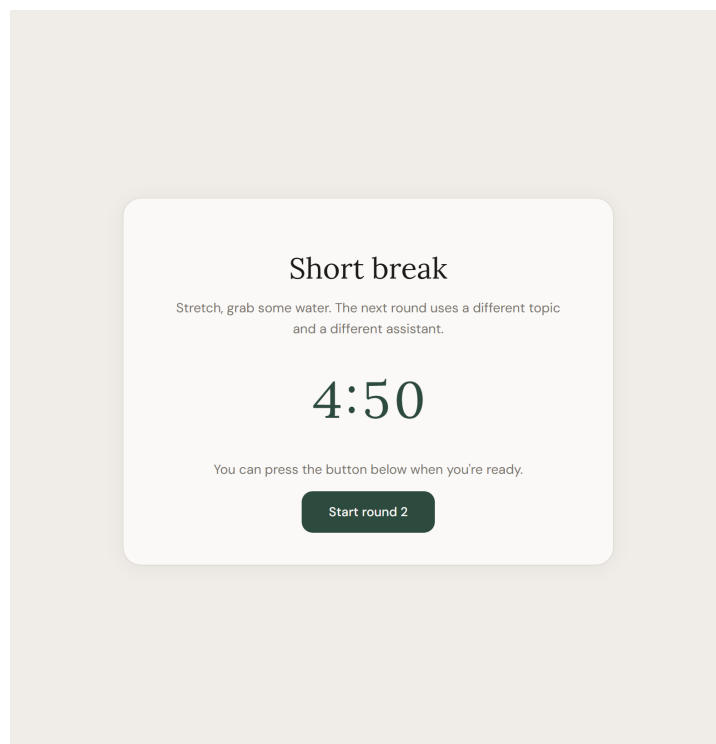


Figure E.10: The break screen. Five minutes between rounds, included to reduce fatigue carry-over from the first condition into the second.

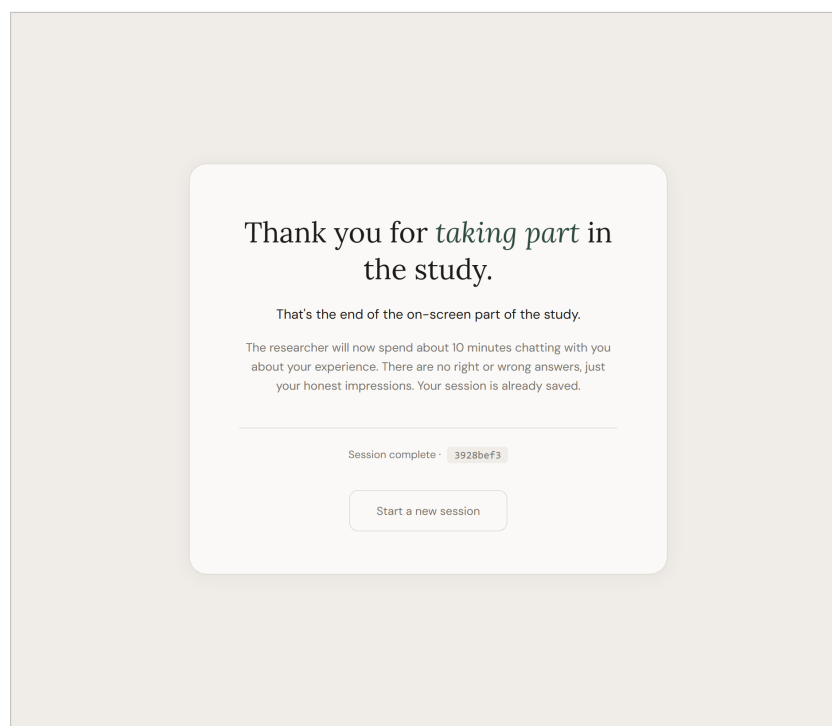
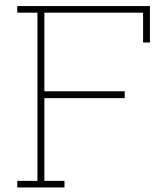


Figure E.11: The closing screen, handing the participant to the interview.



Session Log Schema

Each session was written to a single JSON file. Chat message content was not retained; the chat array stores metadata only.

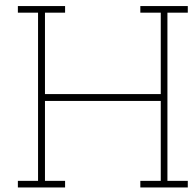
```
{
  "session_id": string,
  "assignment": {
    "sequence_index": int,
    "round_1": { "condition": "hybrid" | "default", "task": "A" | "B" },
    "round_2": { ... }
  },
  "rounds": {
    "1": {
      "pre_position": { "text": string, "word_count": int },
      "writing": { "text": string, "word_count": int },
      "chat": [ { "role", "timestamp", "char_count", "word_count" } ],
      "first_chat_at": timestamp,
      "time_to_first_chat_ms": int,
      "typing_active_ms": int,
      "typing_idle_ms": int,
      "paste_events": [ int ],
      "early_exit": bool,
      "popup_shown": bool,
      "popup_response": "continue" | "submit" | null,
      "post_questionnaire": [ int, int, int, int, int ]
    },
    "2": { ... }
  }
}
```



Linguistic Marker Lexicon

Marker densities reported in Chapter 8 were computed by regular-expression matching against lower-cased text, normalised per 1,000 words, and manually spot-checked. The patterns are given in full so that the analysis can be reproduced or contested.

Category	Pattern (alternation, case-insensitive)
Metacognitive	i think, i believe, i feel, i wonder, i realis*, i notice, i assume, in my opinion, i would say, i consider, to me it, i know that, my instinct, i am aware
Hedges	might, maybe, perhaps, could, possibly, probably, seems, seem to, somewhat, likely, unsure, not sure, may, tend to, usually, often, in a way
Perspective- taking	on the one hand, on the other hand, however, nevertheless, whereas, although, though, some people, others, someone who, an argument can be made, counter, alternatively, you could argue, from [...] perspective, conversely, that said, but for others, in contrast



Supplementary Survey Results

This appendix reports every item collected in the survey, including those carrying no argument in Chapter 4, and lists every statistical test performed. Table H.1 allows that claim to be checked. Percentages are of the full sample ($n = 90$) unless a smaller base is stated.

H.1. Tests performed

Fourteen tests were performed in total, of which eight are reported in Chapter 4.

Comparison	Test	Statistic	p	n	Reported
NCS $\times$ self-reported push-back	Spearman	$\rho = +.28$	.008	88	Ch. 4
SRIS $\times$ self-reported push-back	Spearman	$\rho = +.31$	.004	87	Ch. 4
NCS $\times$ bypass count	Spearman	$\rho = -.06$	.562	89	Ch. 4
SRIS $\times$ bypass count	Spearman	$\rho = +.02$	.832	89	Ch. 4
Hard-to-stop $\times$ bypass count	Spearman	$\rho = +.30$	.004	89	Ch. 4
Bypass on NCS, SRIS, hard-to-stop	OLS	β	p	89	Ch. 4
Detached vs non-detached, bypass	t -test	$t = 2.41$	.018	89	Ch. 4
Guilt $\times$ bypass count	Spearman	$\rho = -.02$	.867	89	Ch. 4
Frequency $\times$ cheating item	Spearman	$\rho = -.22$	.036	89	Ch. 4
NCS $\times$ SRIS	Spearman	$\rho = +.35$	.001	89	—
Frequency $\times$ hard-to-stop	Spearman	$\rho = +.35$	.001	89	—
Cheating item $\times$ bypass count	Spearman	$\rho = +.18$	.100	89	—

Comparison	Test	Statistic	<i>p</i>	<i>n</i>	Reported
Less-capable item × bypass count	Spearman	$\rho = +.26$	.013	89	—
Guilt × discipline group	χ^2	$\chi^2 = 1.36$	.508	89	Ch. 4

Table H.1: All statistical tests performed on the survey data. Regression coefficients are standardised.

The *beta* in the multiple regression has as values: $\beta = .03/.09/.34$ and the $p = .775/.401/.002$ (values did not fit the table).

H.2. Survey Items

H.2.1. Epistemic tension items

Item (1–5)	<i>M</i>	<i>SD</i>	% high
Using an LLM sometimes feels like cheating	3.66	1.19	62
I sometimes feel less capable after relying heavily on an LLM	3.83	1.04	70
I feel more confident presenting ideas that came partly from an LLM	2.46	1.14	21
I find it hard to know when to stop using an LLM and think for myself	2.34	1.22	19

Table H.2: Epistemic tension items ($n = 89$). “% high” is the proportion selecting 4 or 5. The fourth item is the covariate carried into the experiment.

H.2.2. Experience of LLM use, ranked

Respondents ranked five statements from closest to furthest from their own experience.

Statement	Mean rank	Ranked first
It helps me finish the task faster	1.76	50 (59%)
It helps me think through a problem	2.26	21 (25%)
It makes me feel more confident about my work	3.26	4 (5%)
It reduces the anxiety of not knowing	3.40	7 (8%)
I use it out of habit without thinking much about why	4.32	3 (4%)

Table H.3: Ranked experience of academic LLM use ($n = 85$ answering). Lower mean rank indicates closer to the respondent’s experience.

H.2.3. Purposes of LLM use

Purpose	<i>n</i>	%
Understanding or explaining concepts	70	78
Editing or polishing existing text	61	68
Finding or summarising information	57	63
Checking or debugging code	49	54
Brainstorming or generating ideas	45	50
Writing drafts or structuring arguments	40	44
Getting reassurance or a second opinion	30	33

Table H.4: Reported purposes of LLM use ($n = 89$ answering, multiple selection).

H.2.4. Reflection metaphors

Metaphor	<i>n</i>	%
A muscle: it develops with practice and weakens without use	47	52
A habit: something I do without thinking about it	18	20
A talent: some people are naturally more reflective than others	14	16
A luxury: something I'd do more if I had time	6	7
A performance: something I do when someone is watching	1	1

Table H.5: Which metaphor feels most true about reflection ($n = 88$ answering). Two write-ins omitted.

H.2.5. Contribution to helpful answers

Response	<i>n</i>	%
Somewhat: I directed it but the model did most of the intellectual work	35	39
It depends on the task	20	22
A lot: I shaped the question carefully and the answer reflects my thinking	19	21
A little: it would have given a similar answer to almost any prompt	13	14
Not at all: the answer came entirely from the model	1	1

Response	<i>n</i>	%
----------	----------	---

Table H.6: Perceived contribution to a helpful LLM answer ($n = 88$ answering).

H.2.6. When reflection does not happen (Q16) against bypass (Q17)

Situation	Q16		Q17	
	<i>n</i>	%	<i>n</i>	%
Under tight deadlines	51	57	59	66
When I'm tired or overwhelmed	61	68	53	59
When I don't care about the topic	53	59	50	56
When an answer is easy to get (elsewhere)	28	31	34	38
When grades are high stakes	15	17	15	17
When I've already received an answer from an LLM	7	8	16	18

Table H.7: Situations in which students report not reflecting (Q16, $n = 88$) and in which they would bypass a reflective assistant (Q17, $n = 89$). The items share five of six options and are reported as convergent rather than independent evidence.

H.2.7. Design feature endorsement

Approach	<i>n</i>	%
It adapts to context: more reflective for opinion tasks, direct for factual ones	52	58
It lets me choose when I want the reflective mode	47	52
It gives hints or partial answers rather than full responses	42	47
It always explains why it's asking me to think first	36	40
It asks one focused question, then answers fully once I've responded	33	37
None of these; I would prefer LLMs to always answer directly	4	4

Table H.8: Which approaches would make a reflective assistant acceptable ($n = 89$ answering, multiple selection).

H.2.8. Guilt by usage frequency

Frequency	<i>n</i>	Guilt <i>n</i>	Guilt %
Daily	40	16	40
Several times a week	33	20	61
Once a week	8	5	62
Occasionally	5	3	60
Rarely or never	4	2	50

Table H.9: Guilt rate by academic LLM usage frequency.

H.2.9. Discipline

Group	<i>n</i>	Guilt %
STEM: engineering and technology, natural sciences, policy analysis	58	48
Creative: design, arts, game development	19	53
Humanities, social sciences, business, communication	12	67
Unclassified: interdisciplinary or undecided	1	—

Table H.10: Guilt rate by collapsed discipline group. One respondent selecting “interdisciplinary or undecided” could not be assigned, giving $n = 89$ for the comparison. $\chi^2 = 1.36$, $p = .508$.

Delft University of Technology
HUMAN RESEARCH ETHICS
OPENING STATEMENT FOR ONLINE SURVEY

You are being invited to participate in a Master's research study titled: **"Students' Critical Thinking in the Age of AI"**.

This study is conducted by Radu Mihalachiuta (MSc student, Faculty of EEMCS, TU Delft) under the supervision of Cynthia Liem. TU Delft is the responsible institution for this research. There is no external commercial funding.

Purpose of the Study

The aim of this study is to explore how university students understand and experience reflection in their academic work, and how they interact with Large Language Models (LLMs) in educational contexts.

Your responses will help inform the design of a follow-up experimental prototype investigating reflective AI interaction.

The survey takes approximately **10–15 minutes** to complete.

The collected data will be used for:

- A Master's thesis at TU Delft;
- Possible academic publications or conference presentations .

What Participation Involves

If you choose to participate, you will complete an online questionnaire including:

- Multiple-choice and Likert-scale questions;
- Short open-ended reflective questions.

Topics include:

- Your understanding of reflection in education;
- How you use AI tools;
- Hypothetical study scenarios.

No academic grades, performance records, or psychological diagnostic measures are collected.

Data Processing and Privacy

The survey is meant to be anonymous, but there is a risk of revealing personal information through free text fields. Please avoid including identifying or sensitive personal information in open-ended responses.

We do **not** collect:

- Names;
- Student numbers;
- Contact details;
- Academic performance data.

Survey responses are collected via TU Delft's institutional Microsoft Forms environment and transferred to secure TU Delft storage for analysis.

Data will:

- Be stored on TU Delft secure servers;
- Be accessible only to the researcher and academic supervisor;
- Be analysed in aggregated and anonymised form;
- Be retained in accordance with TU Delft research data policy.

Because the survey is anonymous, it will not be possible to withdraw your responses after submission, as they cannot be linked to you.

As with any online activity, there is a minimal risk of data breach. TU Delft institutional systems and data protection policies are used to minimise this risk.

Potential Risks

Risks are minimal. Some participants may experience mild discomfort when reflecting on their study habits or AI use. You are free to skip any question or stop the survey at any time before submitting your responses.

There is no evaluation of your academic performance and no consequences for your participation or non-participation.

Voluntary Participation

Your participation is entirely voluntary.

By clicking **“I agree and start the survey”**, you confirm that:

- You are 18 years or older;
- You are currently enrolled as a university student;
- You understand the purpose of the research;
- You understand that participation is voluntary;
- You understand that your responses are anonymous;
- You consent to the use of your anonymised data for academic research purposes.

If you do not agree, please close this page.

Contact Information

If you have questions about this study, please contact:

Radu Mihalachiuta
MSc Student, Faculty of EEMCS
TU Delft

Supervisor:
Associate Professor Cynthia Liem
TU Delft

Use of Generative AI

In line with TU Delft's policy [63] on the responsible use of generative AI, this appendix discloses how such tools were used in producing this thesis.

Claude (Anthropic) was used throughout the project for:

- Rewriting sentences for clarity, simplifying vocabulary, suggesting transitions, restructuring sections, formatting tables and figures in $\text{\LaTeX}$, and giving feedback on structure. All arguments, claims and interpretations are the researcher's own; the assistance concerned how they were expressed.
- Writing and debugging the Python scripts used to compute descriptive statistics, correlations, regressions and paired comparisons from the survey and session data, and generating the figures reported in Chapters 4 and 8. Every reported value was verified against the source data by the researcher before inclusion.
- Assistance with parts of the prototype's code.

No participant data was shared with any AI tool used for writing or analysis assistance. Session records and survey responses were processed locally and on TU Delft infrastructure.

Moreover, as the object of study, the prototype evaluated in this thesis deploys its two conditions through the Claude API, as described in Chapter 6. Participants' messages to the assistant therefore passed through a commercial provider's interface during sessions in order to be answered. This is a property of the system under study rather than a research tool, and the data-protection implications are addressed in Chapter 6.

Generative AI was used as a tool to support the researcher and not as a substitute for the researcher's own thinking. The researcher takes full responsibility for the originality, accuracy and ethical standards of this work.