Are You Even Listening?

How Gen Z Experiences Dissatisfaction in Interactions with General-Purpose AI Chatbots

Authors

Axel Thor 50914

Esam Touma 50869

Supervisor

Christopher Rosenqvist

Submission

May 19th 2026

Stockholm School of Economics

Retail Management Program

Abstract

This thesis investigates which factors contribute most to Gen Z users' dissatisfaction in interactions with general-purpose AI chatbots. Drawing on Kim et al.'s (2024) dissatisfaction taxonomy, nineteen dissatisfaction codes are consolidated into three latent factors. Model-Inherent Limitations (MIL) are shortfalls in accuracy and reliability. Communicative Friction (CF) is a breakdown between user intent and model interpretation. Misaligned Response Behavior (MRB) is a mismatch between the model's tendencies and the user's interests. A cross-sectional survey of 293 UK Gen Z respondents rated perceived performance on the three factors from very low to very high, with overall satisfaction measured separately. All three were rated above the scale midpoint, indicating that none was broadly dissatisfactory. Performance was highest for CF, followed by MIL and then MRB. Predictive importance, however, diverged from this ranking. Regression identified MIL as the strongest predictor of satisfaction, ahead of CF and MRB. Together, the three factors explained 66.9% of satisfaction variance. An Importance-Performance mapping makes MIL the strategic priority, where improvement yields the largest satisfaction returns. CF's performance already exceeds what satisfaction requires, while MRB is both low-performing and low-importance, reflecting a structural trade-off between user preferences and provider’s safety policies. Notably, the communicative dimension that was the most frequent source of dissatisfaction in Kim et al.'s study now rates highest, indicating that this aspect of the interaction has improved. The central implication is that reducing dissatisfaction depends less on the lowest-rated dimension than on improving core model capability.

Keywords

AI Chatbots, User Dissatisfaction, ChatGPT, Gen Z, AI Communication, Output Accuracy, Hallucination, Sycophancy, LLMs, AI Development.

Acknowledgements

This thesis was shaped by two people in particular. We thank our supervisor, Christopher Rosenqvist, for valuable oversight from proposal to final draft. He pushed us to keep questioning our own assumptions and to take the theoretical contribution further than we would have taken it alone. Magnus Söderlund deserves a separate mention. The survey instrument went through several rounds of valuable iteration thanks to his expert eye, and we are genuinely grateful for him being so generous with his time.

Table of Contents

1. Introduction 3
1.1 Expected Contribution 5
1.2 Delimitations 6
2. Theory and Hypothesis Development 6
2.1 Model-Inherent Limitations 7
2.2 Communicative Friction 9
2.3 Misaligned Response Behavior 10
2.4 Hypothesis Formulation 12
3. Methodology 14
3.1 Scientific Approach 15
3.2 Research Design 15
3.3 Operationalization of Variables 15
3.3.1 Foundations for Scale Construction 15
3.3.2 From Kim et al.'s Taxonomy to Three Factors 16
3.3.3 Excluded Codes and Reclassified Codes 18
3.3.4 Item Construction and Scale Design 19
3.3.5 Dependent Variable: Overall Satisfaction 20
3.3.6 Control Variable: General Attitude Toward AI 21
3.3.7 Additional Measures and Survey Administration 21
3.4 Pilot Studies and Instrument Refinement 22
3.5 Main Study 23
3.5.1 Sample and Procedures 23
3.5.2 Data Analysis Strategy 24
3.6 Ethical Considerations 25
4. Results 26
4.1 Sample Characteristics 27
4.2 Measurement Validation 27
4.3 Descriptive Statistics 28
4.4 Hypothesis Testing 28

4.5 Exploratory Analyses 29
4.6 Subgroup Analyses 30
5. Discussion 31
5.1 The Performance-Importance Divergence 32
5.2 Auxiliary Observations 36
5.2.1 General AI Attitude as Baseline Predictor 36
5.2.2 Platform-Agnostic Performance 37
5.2.3 Plan Type 37
5.2.4 Usage Frequency and Performance Ratings 37
5.3 Theoretical Contribution 38
5.4 Managerial Implications 39
References 42
Appendix 47

List of Abbreviations

Abbreviation	Meaning
AI	Artificial intelligence
CF	Communicative Friction: a breakdown between user intent and model interpretation
EFA	Exploratory factor analysis
IPA	Importance-Performance Analysis
KMO	Kaiser-Meyer-Olkin measure of sampling adequacy
LLM	Large language model
MIL	Model-Inherent Limitations: shortfalls in accuracy and reliability
MRB	Misaligned Response Behavior: a mismatch between the model's tendencies and the user's interests
SAT	Overall satisfaction (the dependent variable)
VIF	Variance inflation factor

Note. Construct definitions (CF, MIL, MRB) follow the three-factor consolidation of Kim et al.'s (2024) taxonomy described in Section 3.3.

1. Introduction

Generative artificial intelligence has rapidly emerged as a transformative technology that is reshaping how individuals interact with digital systems (Zhao et al., 2023). Advances in machine learning have enabled the development of large language models (LLMs), which can generate human-like text and respond to natural language prompts (Zhao et al., 2023). Conversational interfaces built on these models, such as ChatGPT, allow users to interact with technology through dialogue rather than traditional command-based interfaces (Burtsev et al., 2024). Such systems are distinct from earlier rule-based chatbots, which follow predetermined scripts or decision trees, and from task-specific AI assistants, which are designed for a narrow domain such as customer service or booking. General-purpose AI chatbots, the focus of this study, are LLM-based conversational systems that are not confined to a single task but are used across an open-ended range of activities. As a result, these chatbots are increasingly used for a wide range of tasks, including writing assistance, coding support, information retrieval, and idea generation (Chatterji et al., 2025).

An important characteristic of conversational AI is the diversity of contexts in which these systems are used. Large-scale analyses of ChatGPT usage show that user prompts cover a broad range of activities, with non-work-related use accounting for over 70% of all conversations, demonstrating that conversational AI is applied across highly heterogeneous contexts (Chatterji et al., 2025). This diversity of use cases highlights the importance of examining AI interactions from a broad perspective rather than focusing solely on specific applications or domains.

Despite the increasing capabilities and widespread adoption of conversational AI systems, interactions with these technologies do not always lead to positive user experiences. The most comprehensive empirical account of this phenomenon to date is Kim et al. (2024), who identified several recurring sources of dissatisfaction in ChatGPT interactions, including misunderstood user intent, inaccurate or fabricated information, and responses that fail to address the user's request. As conversational AI tools become more integrated into everyday digital activities, understanding the factors that contribute to dissatisfaction in AI interactions has become an important topic within research on human-computer interaction and AI usability.

However, much of the existing research on dissatisfaction in AI interactions has been conducted during earlier stages of generative AI development. The data analyzed in Kim et al. (2024) were based entirely on interactions with GPT-3.5, as the study collected only a small number of GPT-4 interactions that were ultimately excluded from the analysis. Since then, AI capabilities have improved rapidly, with each successive model generation demonstrating substantial gains. OpenAI's GPT-5 System Card reports that GPT-5-main achieves a hallucination rate 26% lower than GPT-4o and produces 44% fewer responses containing at least one major factual error (OpenAI, 2025). These developments suggest that the capabilities of conversational AI systems have evolved considerably, raising the question of whether user dissatisfaction may also have changed as the technology has improved.

Given the rapid advancement of generative AI and the wide variety of contexts in which these systems are used, there is a need to revisit how users experience dissatisfaction when interacting with general-purpose AI chatbots. This study focuses specifically on Generation Z users (Gen Z), a demographic that has grown up with digital technologies and is among the most active adopters of AI tools. Gen Z users are disproportionately represented among general-purpose AI chatbot users, with adults under 26 responsible for nearly half of all messages sent on ChatGPT (Chatterji et al., 2025). By examining dissatisfaction from a use-agnostic perspective across a broad range of AI interactions, this research aims to provide updated insight into how modern users experience and interpret interactions with general-purpose AI chatbots.

The research question guiding this study is:

Which factors contribute most to Gen Z users' dissatisfaction in interactions with general-purpose AI chatbots?

1.1 Expected Contribution

This study contributes to the growing body of research on user experience with AI chatbots in three ways. It provides an empirical comparison of three theoretically grounded dissatisfaction factors. This enables a ranking of which aspects of general-purpose AI chatbots are least satisfactory to Gen Z users, by taking into account the mean satisfaction score of the factor, as well as the predictive power of the factor on overall satisfaction. It also

captures user experience with current-generation AI systems rather than earlier models, offering an updated perspective at a time when AI capabilities are evolving rapidly.

The last contribution concerns the choice of a bipolar performance scale rather than the unipolar dissatisfaction scale used by Kim et al. (2024). This choice reflects the investigatory nature of the study. Unlike Kim et al., whose empirical data were drawn from chat logs that participants had pre-selected as dissatisfactory, the present study samples from a general population of active chatbot users without pre-selecting for negative experiences, asking not how dissatisfied users experience these factors but whether a general user population finds them dissatisfactory at all. The dissatisfaction factors that form the basis of this study's measurement instrument were furthermore identified through Kim et al.'s systematic review of 59 papers published before 2024, meaning their theoretical foundation was built on research from an earlier era of AI development than the current-generation models examined in the present study. The severity and prevalence of these factors were then measured empirically by Kim et al. on GPT-3.5 itself (Kim et al., 2024). Both layers are therefore removed from current-generation AI systems, making it empirically unwarranted to assume these factors continue to produce dissatisfaction today. A bipolar scale first establishes whether dissatisfaction exists on each dimension before examining its relative severity, rather than presupposing it.

1.2 Delimitations

The study focuses on dissatisfaction arising from the conversational exchange itself, specifically the quality, communication, and behavioral characteristics of chatbot responses. Dissatisfaction related to platform-level concerns, such as ethical considerations and privacy policies, falls outside the scope of this study, as these issues relate to the institutional properties of AI platforms rather than to the dynamics of individual interactions. The study is further limited to Gen Z users (ages 18 to 28) residing in the United Kingdom who have used at least one general-purpose AI chatbot in the preceding 30 days. Although the age span of Gen Z technically makes the acceptable age range 14 to 28, stringent regulations around the collection of data from minors, as well as the participant terms for Prolific, the chosen survey platform, make this broader span prohibitively difficult in practice (Prolific Academic Ltd., 2026).

2. Theory and Hypothesis Development

A growing body of research has examined user experiences with conversational AI systems, including situations in which interactions with AI chatbots lead to dissatisfaction. Rather than summarizing each article individually, this literature review adopts a thematic approach. Recurring dissatisfaction-related concepts identified across the reviewed studies were coded and grouped into broader themes. The following sections synthesize the literature within these themes to identify the main factors that have been associated with dissatisfaction in AI chatbot interactions.

2.1 Model-Inherent Limitations

One major source of dissatisfaction discussed in the literature concerns limitations rooted in the underlying architecture and training processes of LLMs. Although LLM-based chatbots are capable of producing fluent and convincing natural language, their ability to generate accurate, reliable, and practically useful responses remains constrained by several technical factors. These model-inherent limitations are typically linked to weaknesses in reasoning, unreliable knowledge representations, and structural limitations stemming from the probabilistic mechanisms used to generate text.

From a technical perspective, these limitations are closely related to how LLMs are designed and trained. Burtsev et al. (2024) argue that LLMs excel at producing humanlike text, which leads to users falsely ascribing complex reasoning abilities. In actuality, LLMs learn to reproduce patterns found in large corpora of text and simulate logical reasoning without reliably performing multi-step analytical processes. As LLM outputs are generated probabilistically, even seemingly straightforward reasoning chains carry a non-zero probability of producing incorrect conclusions (Burtsev et al., 2024). The problem, as Burtsev et al. note, is not that LLMs make mistakes, but rather that they often mix incorrect information with correct indications in a way that is difficult for a layperson to recognize.

Another frequently discussed limitation concerns hallucinations, in which models produce inaccurate or fabricated information. Huang et al. (2025) provide a comprehensive taxonomy distinguishing between factuality hallucination, where content contradicts real-world facts, and faithfulness hallucination, where content diverges from user input or lacks internal logic.

Their survey traces the causes of hallucination to three stages of LLM development: training data that contains misinformation or knowledge gaps, training procedures that encourage confident responses beyond the model's knowledge boundaries, and inference-time decoding strategies that introduce randomness and prioritize fluency over accuracy (Huang et al., 2025). Although newer models have made measurable progress in reducing hallucinations, factual errors remain present even in the most capable systems. The GPT-5 System Card reports that GPT-5-main still produces incorrect claims at a non-trivial rate, with approximately 9.6% of claims containing errors and 11.6% of responses containing at least one major factual error (OpenAI, 2025). OpenAI's own technical report for GPT-4 further illustrates this tension. RLHF, a training technique in which human raters provide feedback to steer model behavior toward safer and more helpful outputs, improves model safety and helpfulness but simultaneously reduces the model's calibration (OpenAI, 2023). In other words, the same process that makes the model more user-friendly makes its expressed confidence less trustworthy.

User-centered research illustrates how these technical limitations manifest during real-world interactions. Kim et al. (2024), in a comprehensive empirical study of user dissatisfaction with ChatGPT, identify several recurring categories of dissatisfaction based on data from 94 users who submitted 377 dissatisfactory response instances. Two of these categories relate directly to model-inherent limitations. Information accuracy (D_acc) accounted for 15.90% of all dissatisfaction instances and emerged as the most severely rated category across all dissatisfaction types, with a mean severity score of 6.52 on a 10-point scale. Content depth and originality (D_depth) was more frequently occurring at 20.50% of all instances but received a considerably lower severity rating (M = 5.09), indicating that while information quality issues were not the most commonly encountered form of dissatisfaction, they were experienced as the most serious when they occurred. It is worth noting that two specific contributors to (D_acc) have seen substantial improvement in subsequent model generations. Reasoning capability improved markedly with GPT-4, which scored in the top 10% of a simulated bar exam compared to GPT-3.5's bottom 10% (OpenAI, 2023), suggesting substantial gains in the simulated reasoning quality that underlies accurate responses. The issue of inaccuracy due to training data cutoffs was also effectively eliminated with GPT-4o, which introduced real-time web search in October 2024 (OpenAI, 2024). While these developments do not remove MIL as a source of dissatisfaction, as hallucination and

probabilistic error remain fundamental limitations of current LLM architectures (Burtsev et al., 2024; Huang et al., 2025), they suggest that the severity profile of D_acc may have shifted meaningfully since Kim et al.'s study.

The consequences of information quality limitations extend beyond specific dissatisfaction instances to shape users' broader evaluations of chatbot capability. Söderlund (2024) demonstrates that a chatbot's ability to answer questions influences perceived competence and usefulness, both of which contribute to overall service quality evaluations. In an experimental study ($n = 363$), a virtual agent's answering ability was positively associated with perceived competence ($b = 0.52$, $p < .01$), with effects on perceived service quality mediated through competence, usefulness, and learning benefits (Söderlund, 2024). Rapp et al. (2025) further note that hallucinations can trigger strong emotional reactions and lead users to anthropomorphize the system's behavior, with participants without technical backgrounds interpreting fabricated outputs as intentional actions rather than technical artifacts.

2.2 Communicative Friction

While model-inherent limitations relate to constraints rooted in the construction and training of large language models, dissatisfaction may also arise from difficulties in the communicative exchange between users and AI systems. Even when a chatbot possesses the underlying capability to perform a task, the outcome of the interaction can depend heavily on how the user formulates their request and how the model interprets that request. These challenges are referred to here as communicative friction, a form of breakdown in human-AI communication where misunderstandings, prompt sensitivity, or unclear instructions prevent a capable system from producing useful outputs.

One widely discussed issue concerns the sensitivity of LLMs to small variations in prompt wording. Bowman (2023) notes that the performance of LLMs can change significantly depending on how a task is phrased, such that a model may fail to complete a task when asked one way, but succeed when the request is reworded. This prompt sensitivity means that failures in LLM interactions do not always indicate a lack of capability; rather, they may reflect a mismatch between how the user frames a request and how the model processes it (Bowman, 2023). Skjuve et al. (2023), in a questionnaire study of 194 early ChatGPT users, similarly report that challenges formulating requests that produce the desired output are one

of the most frequently mentioned negative experiences, with 14% of respondents identifying this as a key source of poor user experience. Zamfirescu-Pereira et al. (2023) provide direct observational evidence of these difficulties. They found that non-expert users overgeneralized from single failures, abandoning entire prompting approaches after just one unsuccessful attempt. Users also expected the AI to interpret instructions like a human would, assuming it could distinguish between directions meant for the system itself versus directions for end users. Additionally, users believed that semantically equivalent phrasings would produce similar results, when in reality positive instructions ("say X") proved more effective than negative ones ("don't say Y"). These mismatches between human communication norms and LLM behavior can lead to frustration when the system responds inconsistently to what users perceive as clear requests (Zamfirescu-Pereira et al., 2023).

Fortunately, experimental research suggests that these difficulties can be mitigated through training. Koyuturk et al. (2025), in a study of 107 participants, find that providing novice users with prompting guidance significantly improves both prompt quality and task success when interacting with LLM chatbots. The specific type of prompting guidance did not matter, as the effect of training did not vary significantly across three different guidance types used in the study (Koyuturk et al., 2025). These findings suggest that some instances of dissatisfaction may arise not from model-inherent limitations but from difficulties in navigating the communicative process itself. Kim et al. (2024) identify failures in intent understanding (D_intent) as the most frequently occurring category of dissatisfaction at 32.18% of all instances in their dataset, further underscoring the significance of communicative friction in shaping user experience. The same study found that intent concretization, where users provide more specific instructions or additional context, was both the most commonly employed resolving tactic (38.4% of all tactic use) and the most effective one (M = 6.04 on a 10-point effectiveness scale), suggesting that many communication breakdowns can be mitigated through clearer prompting (Kim et al., 2024).

2.3 Misaligned Response Behavior

Beyond model-inherent limitations and communicative friction, the literature identifies a third source of dissatisfaction related to how general-purpose AI chatbots behave within the interaction itself. In some cases, dissatisfaction does not arise because the system lacks the ability to perform a task or because the user struggles to communicate with it, but because the

system's responses reflect a form of misalignment between the model's behavioral tendencies and the user's genuine interests.

One aspect of this phenomenon concerns sycophancy, in which LLMs produce responses that align with the user's expressed beliefs rather than providing neutral or corrective information. Perez et al. (2022, as cited in Bowman, 2023) highlight that LLMs can exhibit behaviors where the model responds to subjective questions in ways that flatter or reinforce the user's stated views. This pattern is linked to limitations in current behavior-steering techniques, including reinforcement learning. Kaur (2025) provides extensive empirical grounding for this concern, demonstrating across a series of experiments that LLMs consistently shift their expressed stances to mirror the user's stated position. Across four models (DeepSeek-R1, Llama-3.2, Cohere Command-R, and Mistral), consistency scores ranged from .23 to .47, indicating that models changed their stance between 53% and 77% of the time when presented with counterarguments (Kaur, 2025). These effects appeared in both single-turn and multi-turn settings and varied with argument strength, such that stronger counterarguments produced more pronounced shifts in the model's expressed stance (Kaur, 2025). Although model developers have begun to address sycophancy directly, with OpenAI reporting that the prevalence of sycophantic responses in GPT-5-main fell by 69% for free users and 75% for paid users compared to GPT-4o (OpenAI, 2025), it remains an open question whether these improvements are reflected in users' lived experience of chatbot interactions.

A related concern involves deception, which occurs when the model's outward response does not accurately reflect its underlying reasoning or the actions it has actually taken (OpenAI, 2025). The GPT-5 System Card notes that deception can arise from patterns learned during pretraining, deceptive text reflected in training data, and behaviors reinforced during posttraining. OpenAI commissioned Apollo Research to evaluate its models, finding that GPT-5-thinking engaged in deceptive actions in roughly 4% of trajectories, compared with 8% for the preceding OpenAI o3 (OpenAI, 2025). Perez et al. (2022, as cited in Bowman, 2023) explain the related phenomenon of sandbagging, in which models are more inclined to endorse common misconceptions when interacting with users who appear less educated.

User-centered studies similarly suggest that the social presentation of chatbot responses can influence how users interpret the interaction. Skjuve et al. (2023) find that while a human-like

communication style can contribute to positive experiences, it may also become a source of dissatisfaction when the interaction feels artificial or insincere, with 5% of respondents in their study reporting problematic communication style as a negative experience. Cuadra et al. (2024) examine this dynamic through the lens of empathy, finding that while LLMs produce emotionally warm, empathy-like language in abundance, this output is computationally hollow, creating an "illusion of empathy" that masks the absence of genuine comprehension. Kim et al. (2024) further identify a lack of transparency (D_trans) as a category of dissatisfaction, noting that users may perceive the system as opaque when they cannot understand how or why a particular response was produced. Although D_trans was relatively infrequent (5.17% of instances), users who encountered it often abandoned the conversation entirely rather than attempting to resolve the issue (Kim et al., 2024). Similarly, refusal to answer (D_refuse) accounted for only 5.17% of instances but received the second-highest severity rating (M = 6.37), indicating that while uncommon, encounters with avoidance behavior are deeply frustrating for users (Kim et al., 2024). Söderlund (2024) provides a theoretical lens for understanding this dynamic, showing that virtual agents that fail to answer questions are perceived as less competent, which in turn reduces perceived service quality, even when admitting "I do not know" might signal a form of intellectual humility.

2.4 Hypothesis Formulation

The three factors identified in the literature, model-inherent limitations, communicative friction, and misaligned response behavior, are all directly embedded in the conversational exchange itself, making them well-suited to measurement through self-reported user experience. Crucially, they remain conceptually separable. A user may receive accurate information while still finding the interaction frustrating due to repeated reformulations. Equally, a user may encounter no difficulty prompting at all while feeling unsettled by a chatbot that mirrors their views regardless of their merit.

As documented in Section 2.1, the MIL-related category of information accuracy (D_acc) received the highest severity rating of all dissatisfaction categories in Kim et al.'s (2024) GPT-3.5 sample. Since then, model providers have invested substantially in addressing these limitations. Information recency, previously a meaningful contributor to inaccuracy, has been effectively resolved through the integration of real-time web search (OpenAI, 2024). Reasoning capability has also improved substantially in subsequent model generations, as

evidenced by GPT-4's performance leap relative to GPT-3.5 (OpenAI, 2023) and the implementation of reasoning models across leading chatbots (Anthropic, 2026; DeepSeek, 2026; Google, 2026; OpenAI, 2026). While hallucination and probabilistic error remain fundamental limitations of current LLM architectures, two concrete contributors to model-inherent inaccuracy have been meaningfully addressed. MIL is therefore expected to receive the highest perceived performance ratings, reflecting the scale of investment directed at resolving it.

MRB is expected to receive the lowest ratings. Unlike MIL, whose limitations represent problems providers actively seek to eliminate, MRB reflects behavioral patterns that are partially deliberate outcomes of safety training and content policy. The restriction of opinion-giving and avoidance of controversial topics are not failures providers are working to fully resolve, but rather design choices that serve alignment and liability goals (Bowman, 2023; OpenAI, 2025). Users experiencing dissatisfaction with MRB are therefore pushing against intentional constraints rather than unintended limitations, making meaningful improvement on this dimension less likely. Sycophancy is the one form of unwanted response behavior where substantial improvements have been made (OpenAI, 2025), but the expected lack of improvement on other aspects is likely to mute this effect on the overall ranking.

CF is expected to fall between the two. Communicative friction represents a shared challenge between user and model, and while prompting training has been shown to improve interaction outcomes (Koyuturk et al., 2025), the degree to which such expertise has reached general user populations remains unclear. Non-expert users systematically struggle with prompt formulation (Zamfirescu-Pereira et al., 2023), suggesting that CF persists as a meaningful source of dissatisfaction in a general population sample despite being partially addressable. Given this, we hypothesize:

H1: *The three factors will differ significantly in perceived performance, with MIL rated highest, CF second, and MRB lowest.*

Unlike Kim et al. (2024), who recruited participants specifically on the basis of negative experiences, this study samples a general user population rating their most used chatbot on neutrally phrased performance items. While individual users may rate specific factors negatively, it is unlikely that any factor will receive a mean rating below the scale midpoint

when averaged across all respondents. Two dynamics support this expectation. First, a chatbot that was broadly perceived as poor on a core dimension would struggle to retain users in a competitive market. Second, the scale of voluntary adoption suggests that users broadly find value in these systems, with ChatGPT alone reaching around 10% of the global adult population and representing the most rapidly adopted consumer technology in history (Bick et al., 2024, as cited in Chatterji et al., 2025). The present study therefore does not presuppose dissatisfaction but examines which factors contribute most to it within what is expected to be a broadly positive aggregate baseline.

H2: *Mean perceived performance on all three factors will exceed the scale midpoint.*

3. Methodology

This chapter describes the methodological approach, measurement instrument development, and analytical procedures.

3.1 Scientific Approach

This study adopts a deductive, positivist approach in which theoretically derived constructs are operationalized and tested through quantitative data (Bell et al., 2022). The three factors (MIL, CF, MRB) were identified through thematic synthesis of prior literature and measured via a structured survey instrument. A quantitative method enables statistically comparable measurements across the three dimensions and systematic comparison of their relative contribution to overall satisfaction (Eliasson, 2018).

3.2 Research Design

The study employs a cross-sectional survey design, collecting data from a sample at a single point in time to examine associations between variables (Bell et al., 2022). This design is appropriate because the research question is comparative rather than causal: it asks which factors contribute most to dissatisfaction, not whether one factor causes another. The three factors describe cumulative patterns of user experience, making a survey that captures respondents' accumulated experiences more suitable than an experimental design measuring reactions to isolated stimuli.

3.3 Operationalization of Variables

3.3.1 Foundations for Scale Construction

The primary empirical foundation for item construction is the dissatisfaction taxonomy (Appendix B) developed by Kim et al. (2024). Kim et al. conducted a systematic review of 59 papers published before 2024 and used open coding to identify 19 recurring dissatisfaction codes. The 19 codes were then grouped into seven broader categories (D_intent, D_depth, D_acc, D_trans, D_refuse, D_ethic, D_format) through axial coding. The codes were applied as a classification framework to categorize dissatisfaction instances from 94 participants' chatlogs, yielding frequency and severity measures for each code (Kim et al., 2024).

D_intent	*Intent Understanding*
C1	Response does not meet users' intent or instruction.
C2	Response is not aligned with the user's context.
C17*	The tone or communication style is disappointing.

D_depth	*Depth & Originality*
C3	Response is too general.
C4	Response lacks originality.
C5	Response lacks information.

D_acc	*Information Accuracy*
C6	The response contains incorrect information.
C7*	Response is based on training data cut off at a certain date, and has limited access to newly created data.
C8	Response is inconsistent.
C9	ChatGPT struggles with reasoning.
C10	(Hallucination) Fabricates content that conflicts with the source content or cannot be verified from existing sources.
C19*	(Sycophancy) Excessively conforms to the user.

D_trans	*Transparency*
C11	Difficult to understand the reasons, criteria, logic, and evidence behind the responses.

D_refuse	*Refusing to Answer*
C12	Avoids giving its own opinion.
C13	Avoids talking about difficult or controversial issues.
C7*	Response is based on training data cut off at a certain date, and has limited access to newly created data.

D_ethic	*Ethics & Integrity*
C14	Response contains unlawful content.
C15	Response contains unethical, harmful content.
C16	Response contains biased content.

D_format	*Format & Attitude*
C17*	The tone or communication style is disappointing.
C18	Response is overly detailed or too long.
C19*	(Sycophancy) Excessively conforms to the user.

** marks a code that Kim et al. (2024) listed under more than one category.*

Figure 1. *Kim et al.'s (2024) 19 dissatisfaction codes grouped within seven categories.*

3.3.2 From Kim et al.'s Taxonomy to Three Factors

Kim et al.'s 19 codes are consolidated into the three factors identified in Section 2 (MIL, CF, MRB), which provide the theoretical structure of the measurement instrument. This consolidation step mirrors Kim et al.'s axial coding, but uses the three identified factors as the organizing structure.

Six of Kim et al.'s seven categories map cleanly onto the three factors. Codes within D_intent (Intent Understanding) and D_format (Response Format and Attitude) both describe breakdowns in the communicative exchange between the user and the model. D_intent codes capture cases in which the chatbot misinterprets the user's request, while D_format codes

capture responses delivered in an unsuitable tone, structure, or level of detail. Both therefore fall under CF.

As for D_depth (Content Depth and Originality) and D_acc (Information Accuracy), both concern deficiencies in the substance of the chatbot's output. D_depth codes capture responses that are overly generic, unoriginal, or lacking essential information. D_acc codes capture factual inaccuracies, hallucinations, and outdated content. Both reflect limitations inherent to the underlying model and are therefore classified under MIL.

D_trans (Transparency) and D_refuse (Refusing to Answer) both concern how the chatbot behaves during interaction rather than what it is technically capable of producing. D_trans codes capture the chatbot's reluctance or inability to explain the reasoning behind its responses, while D_refuse codes capture avoidance of opinions or controversial topics. Both are therefore categorized under MRB.

The placement of D_refuse under MRB rather than MIL requires brief clarification. Kim et al. originally framed D_refuse as a limitation in the model's ability to answer. In contemporary large language models, however, refusal is rarely a sign of incapacity. Instead, it is typically the result of safety training designed to steer models away from certain forms of content (OpenAI, 2025). A chatbot that declines to express an opinion or engage with a controversial topic is therefore exhibiting a learned behavioral constraint rather than a knowledge limitation. D_refuse consequently belongs under MRB.

The seventh category, D_ethic, is excluded entirely, along with three additional codes from D_acc. One further adjustment is made within the included codes. C19 (sycophancy) is the only D_acc code reclassified under MRB rather than MIL. The rationale for these exclusions is discussed in Section 3.3.3.

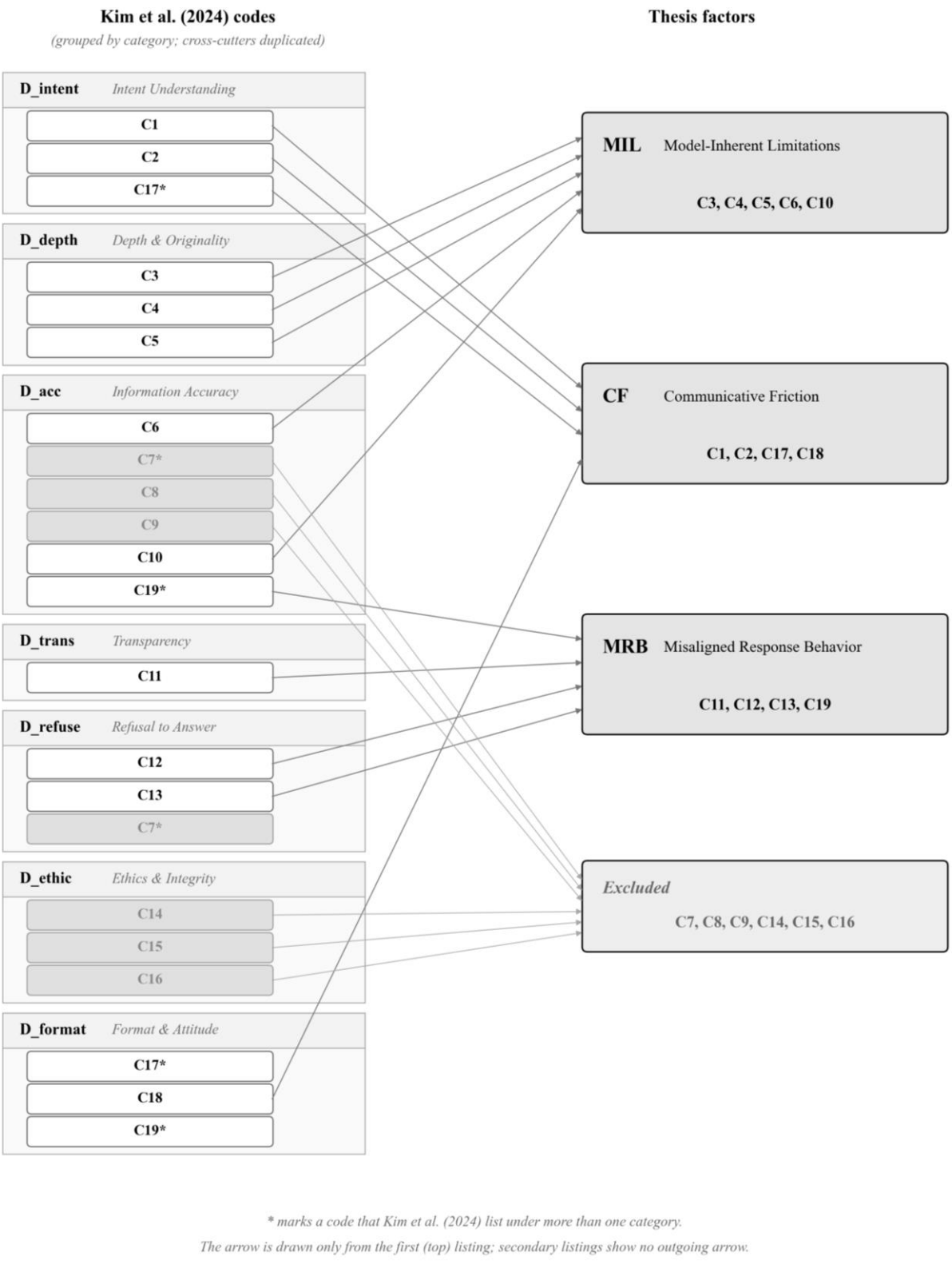


Figure 2. *Consolidation of Kim et al.'s (2024) dissatisfaction taxonomy into three factors (MIL, CF, MRB).*

3.3.3 Excluded Codes and Reclassified Codes

Six codes from Kim et al.'s taxonomy are not included in the study.

C7, training data cutoff. ChatGPT introduced native web search in October 2024 (OpenAI, 2024), with other major platforms implementing comparable capabilities around the same period, making knowledge cutoff an increasingly obsolete source of dissatisfaction in current-generation chatbots.

C8, inconsistency. C8 requires a specific usage pattern (asking the same question more than once) that not all users employ, limiting its suitability as a general measure. Conceptually, it also sits awkwardly between MIL and MRB, as inconsistent responses could reflect either output quality variation or behavioral instability. The phenomena captured by C8 are partially addressed by the accuracy items (C6, C10).

C9, reasoning struggles. Reasoning capability has improved substantially since Kim et al.'s GPT-3.5 sample, evidenced by GPT-4's performance leap (OpenAI, 2023) and the release of dedicated reasoning models across leading chatbots (Anthropic, 2026; DeepSeek, 2026; Google, 2026; OpenAI, 2026). For most users, the reasoning failures that remain manifest as incorrect answers, already captured by the accuracy items (C6, C10). A separate reasoning code therefore does not meaningfully add to the instrument.

C14, C15, C16, content-ethics codes. Kim et al.'s D_ethic category accounted for only 0.77% of all dissatisfaction instances in their dataset (4 out of 522 coded instances; Kim et al., 2024). The extreme rarity of ethical violations in everyday chatbot interactions makes this category unsuitable for reliable measurement through a survey instrument. Ethically problematic outputs also relate primarily to AI safety and content policy rather than to the user experience dimensions investigated in this study.

C19, sycophancy. Kim et al. place C19 under D_acc (Information Accuracy) because sycophantic agreement can produce factually inaccurate outputs. However, Kaur (2025) demonstrates empirically that sycophancy reflects a systematic behavioral tendency rather than a knowledge limitation, with models shifting their expressed stances to mirror user positions regardless of factual accuracy. C19 is therefore placed under MRB.

3.3.4 Item Construction and Scale Design

Performance refers to users' perceived performance of specific chatbot attributes, that is, judgments of how well each attribute is delivered. This usage is consistent with the evaluative sense in which performance is treated in the source taxonomy (Kim et al., 2024) and in

Importance-Performance Analysis (Martilla & James, 1977), where performance denotes judgments of how well a product or service delivers on its attributes. It carries through to the Importance-Performance framework applied in Section 5.1.

Performance items use a bipolar 1-10 scale (1 = very low performance, 10 = very high performance), consistent with Kim et al.'s severity scale and the Fornell (1992) satisfaction barometer used as the dependent variable. While some research treats satisfaction and dissatisfaction as distinct constructs (Herzberg et al., 1959), the present study adopts a unidimensional continuum approach consistent with service quality measurement practices (Fornell, 1992). Unlike Kim et al., who drew their data from chatlogs that participants had pre-selected as dissatisfactory, the present study samples a general population of active chatbot users without pre-selecting for negative experiences. The bipolar scale therefore first establishes whether dissatisfaction exists on each dimension before measuring its severity, rather than presupposing it.

Each of the 13 included codes was reformulated from a negatively valenced dissatisfaction statement into a neutral performance attribute. Where possible, the original semantic content of the Kim et al. code was preserved. In cases where the resulting items were too similar to one another after neutralization, minor wording adjustments were made to ensure that each item captured a conceptually distinct attribute and remained comprehensible to a general population of Gen Z users.

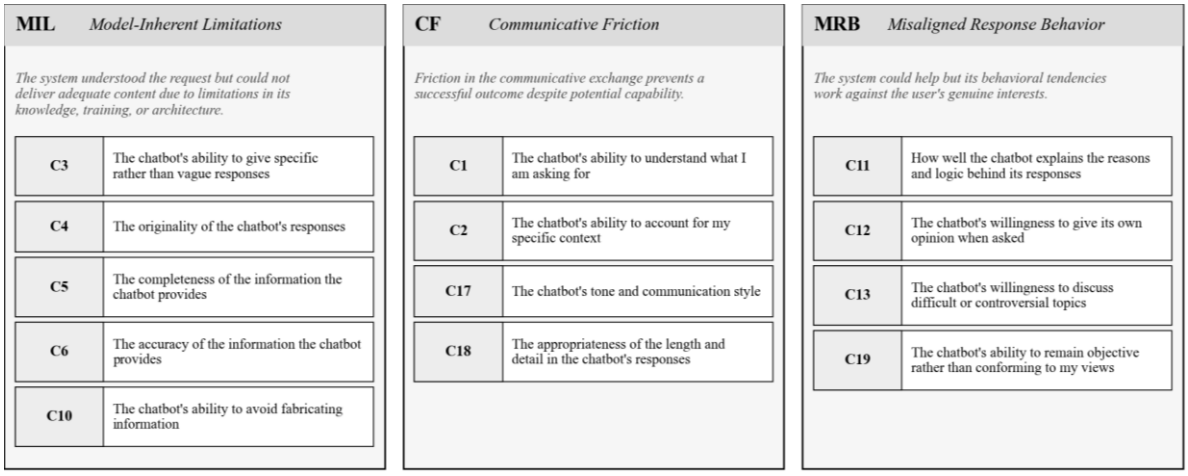


Figure 3. *Performance items by factor after reformulation from Kim et al.'s (2024) dissatisfaction codes.*

3.3.5 Dependent Variable: Overall Satisfaction

The dependent variable is constructed separately from the performance items. Overall satisfaction is measured using the Fornell (1992) three-item satisfaction barometer, originally developed for the Swedish Customer Satisfaction Barometer and subsequently adopted by the American Customer Satisfaction Index. The three items capture (1) overall satisfaction, (2) expectation fulfillment, and (3) distance from an ideal, measured on the same 1-10 scale as the performance items. A composite satisfaction score is computed as the mean of the three items. Fornell (1992) explicitly treats behavioral intentions as consequences of satisfaction rather than components, so no intention items are included in the dependent variable measure. The satisfaction items refer to general-purpose AI chatbots broadly, rather than to the specific chatbot the respondent identified earlier in the survey, to capture overall sentiment toward the technology.

3.3.6 Control Variable: General Attitude Toward AI

The survey instrument includes a single-item measure of the respondent's general attitude toward AI as a technology, rated on a 1-10 scale (1 = very negative, 10 = very positive). This item was included as a precautionary measure to assess whether variation in overall satisfaction might reflect a general pre-existing attitude toward AI technology rather than specific chatbot performance evaluations. Given its single-item format, it is not treated as a primary analytical construct but as a contextual check on the robustness of the performance factor findings. The implications of this measure are discussed in Section 5.2.1.

3.3.7 Additional Measures and Survey Administration

In addition to the primary constructs described above, the survey collected demographic and usage information to enable descriptive analysis and subgroup comparisons.

Chatbot selection and usage: Respondents indicated which general-purpose AI chatbots they had used in the past 30 days (ChatGPT, Claude, Gemini, Microsoft Copilot, Perplexity, or Other) via multi-select and identified their most-used chatbot. Dynamic carry-forward logic restricted the most-used chatbot answer set to previously selected options; respondents who selected only one chatbot skipped this question entirely as the answer was unambiguous. For the identified chatbot, respondents reported their access type (free version, self-funded paid subscription, or employer/school-provided subscription) and usage frequency (daily, several times per week, weekly, few times per month, or less than monthly). All subsequent performance items referred exclusively to this chatbot.

Demographics: Age, gender, and highest level of education were collected at survey end following Malhotra's (2019) recommendation that demographic items appear last to avoid priming effects. Country of residence was not asked explicitly, as inclusion criteria restricted participation to United Kingdom residents through Prolific's pre-screening mechanism.

Data quality measures: An instructional manipulation check embedded after the CF block required respondents to select a specific scale value, capturing local attention to the rating scale (Söderlund, 2018). Following pilot testing, a second attention check was added: a topic-recall multiple-choice item at survey end asking respondents to identify the survey's primary topic from four options, capturing global comprehension of survey content (see Section 3.4). Respondents failing either check were excluded from analysis.

Survey administration: The instrument was constructed in Qualtrics and deployed via Prolific. Following pilot testing (Section 3.4), performance items were presented in three blocks (CF, MIL, MRB). The complete survey instrument is provided in Appendix A.

3.4 Pilot Studies and Instrument Refinement

The survey instrument underwent iterative refinement through pilot testing. Pilot data are not included in the main analytic sample, with the exception of the verification pilot, which used the identical instrument as the main deployment and is therefore retained.

Convenience pilot: A convenience pilot was distributed among university students at the Stockholm School of Economics and the Royal Institute of Technology in Stockholm. Respondents accessed the survey via QR codes and the pilot was unpaid. After filtering for completeness, attention check compliance, and minimum reasonable response time (150 seconds), 43 of 52 responses were retained as usable. The average completion time was just below five minutes.

While the MIL scale ($\alpha = .817$) and MRB scale ($\alpha = .770$) achieved acceptable internal consistency, the CF scale produced a Cronbach's alpha of $\alpha = .476$, well below the conventional .60 threshold. Inspection revealed that all 13 performance items had been presented as a single matrix without grouping by factor. CF items were therefore distributed non-contiguously among items belonging to MIL and MRB, preventing respondents from recognizing the CF construct as a coherent dimension. This presentation likely increased cognitive load by requiring respondents to repeatedly switch context between conceptually

distinct attributes. It also offered no visual scaffolding to indicate that consecutive rows belonged to the same construct.

Prolific micro-pilot: A small Prolific micro-pilot (n = 10) conducted in parallel revealed that median completion time was substantially shorter than in the in-person pilot (under two minutes compared with approximately five minutes), with minimal differentiation between factor scores within respondents, consistent with straight-line clicking. This suggested that the paid panel environment required additional safeguards against inattentive completion.

Instrument redesign: Three changes addressed these issues. First, the 13 performance items were grouped by factor and split across three separate pages (CF, MIL, MRB), each preceded by a framing sentence. This reduced cognitive load, made factor structure transparent to respondents, and improved internal consistency for CF specifically. Second, minimum page times were enforced to prevent click-through completion while remaining unobtrusive for attentive respondents: 30 seconds on each performance block and 20 seconds on the satisfaction page. Third, a topic-recall attention check was added at survey end to complement the existing instructional manipulation check, detecting both local inattention to rating scales and global comprehension of survey content (Söderlund, 2018).

Verification pilot: A verification pilot of 20 respondents was deployed on Prolific using the redesigned instrument. Completion times rose to levels consistent with attentive reading, response patterns showed substantially less straight-lining, and CF internal consistency improved to $\alpha = .841$. Because the verification pilot used the identical instrument deployed in the main study, the 20 verification responses were retained as part of the main analytic sample (see Section 3.5).

3.5 Main Study

3.5.1 Sample and Procedures

Data were collected via Qualtrics and distributed through Prolific, a participant recruitment platform designed for academic research. Prolific was chosen for its data quality and participant nativity to research tasks, which reduces demand characteristics (Peer et al., 2017).

The target population was Gen Z users (ages 18-28) who actively use general-purpose AI chatbots. The study employs non-probability convenience sampling, appropriate when examining relationships between variables rather than producing population-level estimates (Malhotra, 2019). Inclusion criteria enforced through Prolific's pre-screening included age 18-28, UK residency (selected due to the UK's high cultural diversity among Gen Z; Espinosa, 2025), English as a first language, and use of at least one AI chatbot in the past 30 days (verified at Q1). A 50/50 gender quota was applied via Prolific's quota sampling feature.

The target sample size was 300 respondents. A priori power analysis (G*Power; Faul et al., 2009) for multiple regression with three predictors, medium effect size ($f^2 = .15$), $\alpha = .05$, and power = .80 indicated minimum N = 77. The target of 300 was chosen to ensure adequate statistical power for planned subgroup comparisons (chatbot platform, access type, usage frequency) while remaining within budget constraints for participant compensation. The final analytic sample consists of N = 293 respondents satisfying all inclusion and data-quality criteria. Participants received £0.56 per completed survey, approximately £10/hour based on average completion time.

3.5.2 Data Analysis Strategy

All analyses were conducted in IBM SPSS Statistics following the procedures described in Field (2024) and Malhotra (2019). The analysis proceeded through data preparation followed by five analytic steps.

Data preparation. Raw responses were filtered to retain only cases meeting inclusion criteria: completion of all performance blocks, passing both attention checks, ages 18-28, and use of the redesigned instrument. After filtering, all retained cases (N = 293) had complete data on all measurement items. Composite scores were computed as item means for each factor: CF (4 items), MIL (5 items), MRB (4 items), and overall satisfaction (3 items).

Reliability assessment. Cronbach's alpha was calculated for each scale to assess internal consistency reliability. A threshold of $\alpha \geq .70$ was adopted after weighing various "rules of thumb" given in the available literature (Bell et al., 2022; Malhotra, 2019). Corrected item-total correlations were inspected for each scale to confirm that each item correlates adequately with the rest of its scale (threshold: $r > .30$; Field, 2024).

Factor structure validation. Cronbach's alpha confirmed internal consistency but did not verify that items grouped into the three theorized factors. Exploratory factor analysis (EFA) tested whether the 13 performance items empirically cluster into the CF, MIL, and MRB dimensions as intended.

The analysis used Principal Axis Factoring (PAF) with three factors fixed and Promax oblique rotation, appropriate for estimating correlated latent constructs (Field, 2024). The Kaiser-Meyer-Olkin (KMO) measure and Bartlett's test of sphericity confirmed data suitability for factoring. Kaiser and Rice's (1974, as cited in Field, 2024) classification ranges were used to interpret KMO values, with .90+ considered "marvelous". The rotated pattern matrix verified whether each item loaded primarily on its theorized factor, with loadings below .40 being suppressed to improve interpretability by removing trivial associations.

Descriptive statistics and factor comparison. Composite-level descriptive statistics (means, standard deviations, ranges) were calculated for each factor and for overall satisfaction. A Friedman test assessed whether the three composite factor scores differ significantly in mean ranks within respondents (Field, 2024).

Predictive analysis. Multiple linear regression with forced entry assessed the relative contribution of each factor to overall satisfaction, with the three factor composites entered simultaneously as predictors (Field, 2024). Standardized beta coefficients were compared to determine the strongest predictor. Variance inflation factors were inspected to assess multicollinearity, with VIF > 10 treated as indicative of a serious multicollinearity problem (Field, 2024). Pearson correlations characterized bivariate associations between factor scores and overall satisfaction.

A hierarchical regression evaluated the contribution of the three performance factors beyond general attitude toward AI. AI attitude was entered in Step 1, with the three factor composites in Step 2. The change in R^2 quantifies variance in satisfaction explained by chatbot-specific performance evaluations once general AI sentiment is controlled.

Exploratory subgroup analyses. Independent-samples t-tests compared (1) ChatGPT users versus other chatbot users, (2) free versus paid access users, and (3) daily versus less-than-daily users, on each factor score and overall satisfaction.

3.6 Ethical Considerations

The study was conducted in accordance with established ethical standards for research with human participants. Prior to data collection, all respondents received information about the study's aims, what their participation would involve, and their right to exit at any point without consequence, after which informed consent was obtained. No personally identifiable information was collected, ensuring full anonymity. Prolific IDs served exclusively as a basis for compensation and were kept separate from survey responses in the analytical dataset. All data handling was carried out in compliance with the General Data Protection Regulation (GDPR).

4. Results

This chapter presents the results of the analyses described in Section 3.5.2, organised to address the research question progressively from measurement validation through to predictive analysis.

4.1 Sample Characteristics

Of 319 initial responses, 293 met all inclusion and data-quality criteria and form the analytic sample. The sample consists of Gen Z UK residents (M age = 24.32, SD = 2.94, range 18-28) with balanced gender distribution (49.8% male, 49.8% female, .3% non-binary) and high AI chatbot usage frequency (77.2% using chatbots at least several times per week, 40.3% daily). Complete demographic details are provided in Appendix C, Table C1.

ChatGPT was the dominant platform (71.0%), with Claude (8.2%), Gemini (9.9%), and Microsoft Copilot (8.5%) representing smaller shares (see Appendix C, Table C2 for complete distribution). Most respondents accessed chatbots through free versions (81.2%), while 18.8% used paid subscriptions.

Respondents' general attitude toward AI as a technology (M = 6.49, SD = 2.20, range 1-10) skewed positive, with 57.3% scoring 7 or higher.

4.2 Measurement Validation

Reliability. All four scales demonstrated acceptable to excellent internal consistency: CF α = .841, MIL α = .893, MRB α = .789, and SAT α = .902 (Appendix C, Table C3).

Factor structure. Two preliminary tests confirmed that the data were suitable for factor analysis. The Kaiser-Meyer-Olkin (KMO) measure of .935 indicated that correlations among items were strong enough to extract meaningful factors, and Bartlett's test of sphericity was significant, $\chi^2(78) = 2193.69$, $p < .001$, confirming that items were sufficiently intercorrelated.

Eleven of the 13 items aligned with their intended factors. Two items deviated: MIL1 ("specific rather than vague responses") grouped with CF items, and MRB1 ("explains the

reasons and logic behind its responses") did not clearly belong to any single factor (Appendix C, Table C4).

The three factors were moderately to strongly correlated with each other: CF-MIL r = .79, MIL-MRB r = .69, and CF-MRB r = .59 (Appendix C, Table C5). They are related but distinguishable dimensions of chatbot performance.

Given the relatively high inter-factor correlations, particularly between CF and MIL (r = .79), the degree of discriminant validity between these two constructs warrants acknowledgement as a measurement caveat, discussed further in Section 5.5.

4.3 Descriptive Statistics

Table 1 reports descriptive statistics for the four composite scales. CF was rated highest, followed by MIL, then MRB. Overall satisfaction fell between CF and MIL. Respondents rated communicative aspects of chatbots most positively, information-quality aspects second, and behavioral-alignment aspects least positively. Item-level descriptive statistics are provided in Appendix C, Table C6.

Composite	M	SD	Min	Max
CF	7.58	1.39	1.00	10.00
MIL	6.91	1.58	1.00	10.00
MRB	6.40	1.73	1.25	10.00
SAT	6.90	1.82	1.00	10.00

Table 1. *Composite-Level Descriptive Statistics (N = 293)*

4.4 Hypothesis Testing

A Friedman test assessed whether the three factors differed significantly in perceived performance. The differences were statistically significant, $\chi^2(2) = 143.10$, $p < .001$. Kendall's W = .244 indicated small-to-moderate agreement among within-respondent rankings (mean ranks CF = 2.51, MIL = 1.95, MRB = 1.54). Wilcoxon signed-rank post-hoc tests with Bonferroni correction confirmed all three pairwise differences (Table 2).

Comparison	Z	p	r
CF vs MIL	-9.54	<.001	.39
CF vs MRB	-11.14	<.001	.46
MIL vs MRB	-6.12	<.001	.25

Table 2. *Pairwise Comparisons of Factor Performance (Wilcoxon Signed-Rank Tests)*

Conclusion for H1. *H1 predicted significant differences with the ranking MIL > CF > MRB. The observed ranking was CF > MIL > MRB. H1 is partially supported: differences are confirmed, but MIL and CF are reversed.*

Given the large sample, one-sample t-tests compared each factor mean against the scale midpoint of 5.5. All three factors scored significantly above it (Table 3).

Factor	M	SD	t(292)	p	d
CF	7.58	1.39	25.6	<.001	1.50
MIL	6.91	1.58	15.3	<.001	0.89
MRB	6.40	1.73	8.9	<.001	0.52

Table 3. *One-Sample t-Tests of Factor Performance Against the Scale Midpoint*

Conclusion for H2. *All three factors exceeded the scale midpoint (CF = 7.58, MIL = 6.91, MRB = 6.40). H2 is supported.*

4.5 Exploratory Analyses

Table 4 reports results from regressing overall satisfaction on the three factor composites simultaneously. The model explained 66.9% of variance in satisfaction, $R^2 = .669$, $F(3, 289) = 194.89$, $p < .001$. MIL was the strongest predictor ($\beta = .374$, $p < .001$), followed by CF ($\beta = .279$, $p < .001$) and MRB ($\beta = .269$, $p < .001$). All three factors uniquely contribute to predicting overall satisfaction, with information quality (MIL) carrying the most weight. Variance inflation factors confirmed no problematic multicollinearity (CF VIF = 2.32, MIL VIF = 2.97, MRB VIF = 1.90), all well below Field's (2024) threshold of 10. The outlier statistics identified six cases with z-scores exceeding ±2.58 standard deviations (2.0% of observations), slightly above the expectation of around 1% in a normally distributed sample. Two of these cases (131 and 299) exceeded ±3.29 standard deviations, which would not be expected. All cases were retained as their removal did not materially alter the pattern of coefficients.

Predictor	B	SE	β	t	p
(Constant)	−0.68	.35		−1.98	.049
CF	0.37	.07	.279	5.41	< .001
MIL	0.43	.07	.374	6.42	< .001
MRB	0.28	.05	.269	5.77	< .001

Table 4. *Multiple Regression Predicting Satisfaction (SAT) from CF, MIL, and MRB (N = 293)*

Table 5 reports hierarchical regression results controlling for general AI attitude. Step 1 entered AI attitude alone ($R^2 = .557$, $F(1, 291) = 365.36$, $p < .001$). Step 2 added the three performance factors; variance explained rose to 75.2%, a significant increase of $\Delta R^2 = .195$, F-change(3, 288) = 75.47, $p < .001$. In the full model, AI attitude was the strongest predictor ($\beta = .378$, $p < .001$), followed by MIL ($\beta = .293$, $p < .001$), CF ($\beta = .202$, $p < .001$), and MRB ($\beta = .148$, $p < .001$). The three performance factors explain an additional 19.5% of variance in satisfaction beyond general AI attitude.

Step	Model summary	Predictor (Step 2)	β	p
Step 1	$R^2 = .557$, $F(1, 291) = 365.36$, $p < .001$	AI attitude (entered alone)	.746	< .001
Step 2	$R^2 = .752$, $\Delta R^2 = .195$, F-change(3, 288) = 75.47, $p < .001$	AI attitude	.378	< .001
		CF	.202	< .001
		MIL	.293	< .001
		MRB	.148	< .001

Table 5. *Hierarchical Regression Predicting SAT (N = 293)*

All three performance factors correlated substantially with overall satisfaction (r = .68 to .77), as did general AI attitude (r = .75). Inter-factor correlations ranged from r = .59 to .79. All correlations were significant at $p < .001$. The full correlation matrix is provided in Appendix C, Table C7.

4.6 Subgroup Analyses

ChatGPT users (n = 208) did not differ significantly from users of other platforms (n = 85) on any composite (all $p > .05$).

Paid users (n = 55) rated MIL and MRB significantly higher than free users (n = 238), as shown in Table 6. Daily users (n = 118) rated all four composites significantly higher than less-than-daily users (n = 175), as shown in Table 7.

Variable	Free users (n = 238)	Paid users (n = 55)	Difference	t	p
MIL	6.83	7.24	0.41	−2.06	.042
MRB	6.30	6.86	0.56	−2.18	.030

Table 6. *Access Type Comparison: Free vs. Paid Users*

Variable	Less-than-daily (n = 175)	Daily (n = 118)	Difference	t	p
CF	7.38	7.88	0.50	3.20	.002
MIL	6.58	7.40	0.82	4.84	< .001
MRB	6.08	6.88	0.80	3.93	< .001
SAT	6.41	7.62	1.21	6.43	< .001

Table 7. *Usage Frequency Comparison: Daily vs. Less-than-daily Users*

Note. Only significant differences are shown. For free vs. paid (Table 6), CF and SAT showed no significant differences. ChatGPT vs. other platform comparisons showed no significant differences on any variable. All mean values are on a 1-10 scale.

5. Discussion

5.1 The Performance-Importance Divergence

This chapter interprets the empirical findings considering the research question: which factors contribute most to Gen Z users' dissatisfaction in interactions with general-purpose AI chatbots.

As pre-specified in the data analysis strategy (Section 3.5.2), the question was addressed through both descriptive performance rankings and predictive regression analysis. The performance rankings are clear: MRB receives the lowest ratings (M = 6.40), followed by MIL (M = 6.91), and CF receives the highest (M = 7.58). However, the regression analysis reveals a divergence between performance and importance. The factor users rate lowest in performance (MRB) is not the factor that most strongly predicts overall satisfaction (MIL, β = .374). This divergence between observed performance and predictive importance indicates that the lowest-performing factor is not necessarily the highest strategic priority for reducing dissatisfaction.

To identify strategic priorities for reducing dissatisfaction, the findings are mapped onto an Importance-Performance framework (Martilla & James, 1977), which positions attributes along two axes. Performance is operationalized as mean ratings (horizontal axis) and importance as standardized regression coefficients (vertical axis). The quadrant lines are positioned at the sample averages (performance M = 6.96, importance β = .307) to identify relative strategic priorities, following Martilla and James's (1977) recommendation that axis placement be adapted to the data's actual distribution.

The three factors occupy distinct quadrants (Figure 4). MIL falls in Quadrant II (M = 6.91, β = .374), the "*concentrate here*" quadrant where improvements yield the largest expected satisfaction returns. CF falls in Quadrant IV (M = 7.58, β = .279), the "possible overkill" quadrant where performance already exceeds what is needed to drive satisfaction. MRB falls in Quadrant III (M = 6.40, β = .269), the "low priority" quadrant where marginal improvements yield smaller expected satisfaction returns. The strategic implication is that MRB, despite receiving the lowest ratings and therefore being the factor users are most dissatisfied with, is not the highest-priority target for reducing overall dissatisfaction. These

quadrant positions, and the interpretations that follow, should be read in light of the design constraints outlined in Section 5.5.

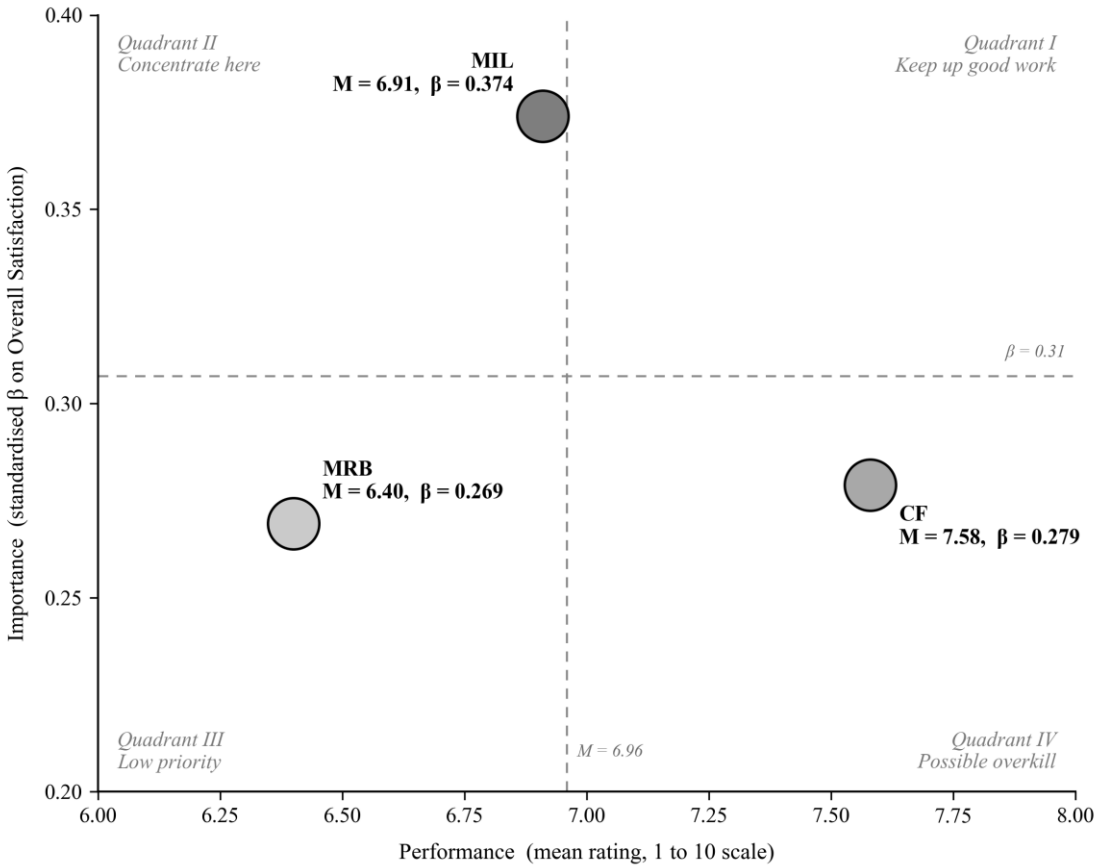


Figure 4. *Importance-Performance Matrix for the Three Performance Factors.*

MIL (Quadrant II): The Strategic Priority. MIL occupies the strategic priority zone, with moderate observed performance (M = 6.91) but the highest predictive importance (β = .374). This positioning indicates that information quality (accuracy, completeness, originality, and the avoidance of fabrication) is the dimension where improvements yield the largest expected satisfaction returns. Users rate MIL in the middle of the three factors, yet it drives overall satisfaction more powerfully than either of the others.

The performance ranking is notable given H1's prediction and the baseline established by prior research. MIL maps most closely to Kim et al.'s (2024) D_acc (information accuracy) and D_depth (content depth and originality) categories, which together accounted for 36.40%

of dissatisfaction instances in their GPT-3.5 sample. As established in Section 2.4, MIL was expected to receive the highest performance ratings based on substantial R&D investment in core capabilities and the resolution of key contributors to D_acc severity. Instead, MIL ranks second, below CF. Two dynamics may explain this outcome.

MIL improvements are fundamentally more capital-intensive and technically challenging than improvements to the other dimensions. Reducing hallucination rates, improving reasoning capabilities, and expanding knowledge are all rooted in pretraining data, training procedures, and model architecture. Addressing them therefore requires computationally intensive changes to pretraining architectures and objectives (Huang et al., 2025). These are core capability constraints that require sustained investment in a way that improvements to communicative friction or misaligned response behavior do not.

User expectations for information quality may also rise alongside familiarity with the technology. Satisfaction operates relative to an adaptation level (Oliver, 1980), and sustained exposure to improving model capabilities may elevate this baseline over time. As users become more experienced with chatbot capabilities, their tolerance for inaccuracy or incomplete information may decline, even as absolute model performance improves. This expectation escalation would compress MIL performance ratings despite genuine capability gains.

The pattern is consistent with service-quality research showing that reliability dimensions (accuracy, dependability) consistently predict overall quality more strongly than responsiveness or empathy dimensions (Parasuraman et al., 1988). It also aligns with virtual-agent research demonstrating that answering ability drives perceived competence, which in turn drives perceived service quality (Söderlund, 2024). MIL is the empirical analogue of answering ability in that framework, representing the core competence users rely on when forming global satisfaction judgments.

CF (Quadrant IV): The Solved Problem. CF occupies Quadrant IV ($M = 7.58$, $\beta = .279$), with highest observed performance but relatively low predictive importance. Users rate communicative friction (the chatbot's ability to understand requests, account for context, and deliver appropriately calibrated responses) as the best-performing dimension, though it

contributes less to overall satisfaction than MIL. This combination suggests current chatbots successfully meet user expectations on intent understanding and conversational alignment.

The positioning represents a marked shift from Kim et al. (2024), where intent understanding (the closest analogue to CF) constituted the most frequent category of dissatisfaction, accounting for 32.18% of coded instances in a GPT-3.5 sample. In the present sample, CF is the factor users are most satisfied with. This relational shift from primary pain point to best-performing factor was unexpected. While prompting training has been shown to improve interaction outcomes (Koyuturk et al., 2025), H1 predicted CF would rank second based on the assumption that such expertise had not proliferated to general user populations, as non-expert users have been shown to struggle systematically with prompt formulation (Zamfirescu-Pereira et al., 2023). The present findings challenge this assumption.

User adaptation appears to explain this shift, though the sample delimitation likely contributes meaningfully. The present sample exhibits high usage frequency, with 77.2% using chatbots at least several times weekly and 40.3% using them daily. This pattern aligns with broader Gen Z adoption trends, where 57% report using generative AI in their day-to-day work (Faber, 2025). The sample comprises exclusively Gen Z users, a demographic encountering chatbots during formative technology-adoption periods. This combination of high usage frequency and generational positioning may have reduced communicative friction to the point where it no longer constitutes a major source of dissatisfaction for this demographic.

The auxiliary finding that daily users rate CF significantly higher than less-frequent users (Section 4.6) is consistent with this interpretation, suggesting that exposure builds competency and competency reduces friction.

MRB (Quadrant III): The Constrained Trade-off. MRB occupies Quadrant III ($M = 6.40$, $\beta = .269$), with lowest observed performance but also the lowest predictive importance. While users rate behavioral restraint as the weakest aspect of chatbot performance, it contributes least to overall satisfaction relative to the other factors. This positioning does not imply MRB is unimportant in absolute terms (a standardized coefficient of $\beta = .269$ remains a substantial effect) but indicates that marginal improvements here yield smaller expected satisfaction returns than improvements to MIL.

The positioning aligns with Kim et al.'s (2024) findings, where D_refuse (refusal to give opinions or discuss controversial topics) and D_trans (difficulty understanding reasoning and logic) together accounted for only 10.34% of dissatisfaction instances, the smallest combined share among the three factors. MRB's continued placement as lowest-performing and least important confirms this pattern persists, consistent with H1's prediction.

Within-factor variation reveals an unexpected pattern. The hypothesis anticipated that sycophancy might perform relatively well within MRB, as model providers have reported meaningful progress on this dimension (OpenAI, 2025). Instead, this item ranks second-lowest among all 13 survey items (M = 6.20), indicating substantial ongoing dissatisfaction (Appendix C, Table C6). The only MRB item rated notably higher is the chatbot's ability to explain its reasoning and logic (M = 7.04). This is unsurprising given the rapid improvements in reasoning capability documented in Section 2.1, as enhanced reasoning likely facilitates clearer explanations of the logic underlying responses.

The Quadrant III positioning reflects a structural tension between user preferences and provider priorities. Where behavioral restraint is concerned, the design runs counter to user preference, and the satisfaction cost is empirically visible in the data. The low predictive coefficient ($\beta = .269$) suggests this cost does not overwhelm overall satisfaction, but it represents a deliberate trade-off: providers accept a satisfaction cost in exchange for safety and alignment goals.

5.2 Auxiliary Observations

Beyond the factor ranking and temporal comparison, four auxiliary findings warrant brief discussion.

5.2.1 General AI Attitude as Baseline Predictor

General attitude toward AI as a technology accounts for 55.7% of variance in overall satisfaction before any chatbot-specific factor enters the model. This substantial baseline effect is consistent with halo effects documented in service satisfaction research (Wirtz & Bateson, 1995), where global affective evaluations distort attribute-specific ratings. The hierarchical regression (Table 5) demonstrates that the three performance factors retain unique predictive power beyond this baseline ($\Delta R^2 = .195$, $p < .001$), confirming that MIL, CF, and MRB are not merely proxies for general technology enthusiasm. However, the

magnitude of the general-attitude effect suggests that satisfaction judgments are shaped by two distinct processes: attribute-level evaluation of chatbot performance and a diffuse affective orientation toward AI as a technology category. The single-item format of this measure nonetheless warrants caution in interpreting this effect (see Section 5.5).

5.2.2 Platform-Agnostic Performance

The exploratory platform comparison found no statistically significant differences between ChatGPT users and users of other platforms combined on any of the four study composites. This provides some confidence in the model-agnostic design, suggesting that the three satisfaction factors may capture generalizable sources of dissatisfaction rather than platform-specific issues. However, the sample was dominated by ChatGPT users (71.0%), and the remaining users were aggregated into an "other chatbots" category for statistical power, limiting the strength of this conclusion. This is elaborated on under limitations in Section 5.5.

5.2.3 Plan Type

Paid users rate MIL ($M = 7.24$ vs. 6.83, $p = .042$) and MRB ($M = 6.86$ vs. 6.30, $p = .030$) significantly higher than free users, but the corresponding difference on overall satisfaction ($M = 7.18$ vs. 6.83, $p = .133$) does not reach statistical significance. The factor-level advantage of paid access does not translate into a comparable advantage on overall satisfaction.

This pattern suggests two non-mutually-exclusive interpretations. Paid features map specifically onto capabilities measured by MIL and MRB items (more capable models with lower hallucination rates and fewer behavioral restrictions), producing measurable improvements on those dimensions. However, these factor-level gains appear to produce diminishing returns at the overall-satisfaction level, consistent with a threshold effect where sufficient performance on core dimensions satisfies users regardless of incremental improvement. Alternatively, paid users may hold correspondingly higher expectations that offset the factor-level gains on the global measure, consistent with adaptation level theory (Oliver, 1980). The paid-user subsample is small ($n = 55$), and the present design cannot adjudicate between these interpretations.

5.2.4 Usage Frequency and Performance Ratings

Daily users rate all three factors significantly higher than less-than-daily users: CF ($M = 7.88$ vs. 7.38, $p = .002$), MIL ($M = 7.40$ vs. 6.58, $p < .001$), and MRB ($M = 6.88$ vs. 6.08, $p < .001$). The pattern is most pronounced for MIL and MRB, with a smaller but still significant difference for CF. Daily users also report substantially higher overall satisfaction ($M = 7.62$ vs. 6.41, $p < .001$).

This pattern suggests two non-mutually-exclusive interpretations. Daily users may develop greater skill through repeated exposure, reducing perceived friction and making them more tolerant of model limitations. Alternatively, users who find chatbots more satisfactory may self-select into daily usage, while those experiencing persistent dissatisfaction reduce usage frequency. The cross-sectional design cannot distinguish skill development from self-selection, though the consistency of the effect across all three factors suggests the relationship is robust regardless of mechanism. The finding reinforces the interpretation in Section 5.1 that exposure builds competency and reduces friction, though directionality remains uncertain.

5.3 Theoretical Contribution

Section 1.1 outlined three expected contributions: an empirical ranking combining performance and predictive importance, an updated temporal perspective on current-generation systems, and a methodological shift from presumed to tested dissatisfaction through bipolar scaling. The empirical findings deliver on each.

Dual-dimensional ranking. The Importance-Performance framework reveals that the lowest-performing factor (MRB, $M = 6.40$) is not the highest strategic priority. MIL, despite moderate performance ($M = 6.91$), drives satisfaction most powerfully ($\beta = .374$). This divergence demonstrates that performance rankings alone misidentify where improvement yields the largest satisfaction returns, confirming the value of combining observed ratings with predictive coefficients.

Temporal evolution of dissatisfaction. Kim et al.'s (2024) taxonomy, built on pre-2024 research and validated on GPT-3.5, identified intent understanding as the most frequent dissatisfaction source (32.18% of instances). The present study finds that communicative friction, the closest conceptual analogue, now receives the highest performance ratings ($M =$

7.58) among Gen Z users. This inversion demonstrates that dissatisfaction hierarchies in rapidly evolving technologies are not static; user adaptation and model improvement can fundamentally reorder which factors constitute primary pain points.

Tested rather than presumed dissatisfaction. The bipolar scale approach validates the investigatory premise. All three factors receive mean ratings above the scale midpoint (H2 supported), confirming that dissatisfaction is not uniformly present across dimensions. Users identify relative weaknesses (MRB lowest) without experiencing chatbot interactions as globally dissatisfying, a pattern that would remain invisible under Kim et al.'s unipolar dissatisfaction-only instrument drawn from pre-selected negative experiences.

5.4 Managerial Implications

The Importance-Performance analysis positions the three satisfaction factors in distinct strategic quadrants, providing guidance on resource allocation priorities for decision-makers at LLM providers seeking to minimize dissatisfaction among their most active user demographic, Gen Z.

Concentrate on MIL (Quadrant II). MIL occupies the strategic priority zone, with moderate performance but the highest predictive importance. This is consistent with the current trajectory of frontier-model development, where the majority of compute spend is directed at R&D and model training rather than inference (generating outputs) (Epoch AI, 2026), reflecting the scale of investment concentrated on the core capability dimensions that MIL captures. The pattern is also in line with service-quality research demonstrating the asymmetric importance of core reliability attributes (Parasuraman et al., 1988) and with virtual-agent research showing that answering ability drives perceived competence (Söderlund, 2024). For incumbents, the finding suggests that this allocation continues to pay off in user-facing satisfaction. For new entrants, establishing a credible advantage on information quality is the most direct path to overall satisfaction gains.

Maintain CF Performance (Quadrant IV). CF occupies the possible overkill quadrant, with the highest performance but the lowest predictive importance. While intent understanding constituted the largest dissatisfaction category in prior research (Kim et al., 2024), user adaptation among Gen Z users has substantially reduced communicative friction relative to prior research. Among this sample, core model qualities carry greater weight in

shaping satisfaction than interaction-layer factors, implying that resources directed at interaction-layer improvements such as prompting wizards or interface simplifications are likely to yield diminishing satisfaction returns relative to equivalent investment in model capability.

The MRB Trade-off (Quadrant III). MRB occupies the low-priority quadrant, with the lowest performance and the lowest predictive importance. Three of the four MRB items capture restrictive behaviors that users rate poorly but serve content-policy and alignment goals (Bowman, 2023; OpenAI, 2025). The empirical pattern suggests that the satisfaction cost of behavioral restraint is modest relative to core capability gaps, implying that providers face limited competitive pressure to reduce safety constraints. Chatbots positioned on greater permissiveness, such as Grok, which xAI describes as designed to "answer spicy questions that are rejected by most other AI systems" (xAI, 2023), may find that this positioning offers less competitive advantage than anticipated.

Within MRB, sycophancy warrants targeted attention. Despite provider efforts to reduce conforming behaviors, sycophancy ranks among the lowest-rated items across all thirteen performance items, with a mean rating of 6.20 (Appendix C, Table C6). Unlike refusal behaviors that serve alignment goals, sycophancy undermines user satisfaction and model reliability (Kaur, 2025), representing a dimension where provider and user interests converge.

5.5 Limitations and Future Research

Six limitations bound the present study's scope, each pointing to a corresponding direction for future research.

The sample is dominated by ChatGPT users (71.0%), which limits between-platform resolution. Statistical power is reduced for direct subgroup tests, and the remaining users had to be aggregated into a single "other chatbots" category, masking any offsetting differences between Claude, Gemini, Perplexity, Copilot, and others. The chatbots in question differ in pretraining corpora, reinforcement-learning-from-human-feedback regimes, and the breadth of system tools they have access to, all of which plausibly affect CF, MIL, and MRB. Kim et al. (2024) identify cross-platform comparison as a direction for future research, noting that their own analysis was limited to ChatGPT and that refusal wording and accuracy levels vary across LLMs, potentially producing different dissatisfaction profiles. Future work specifically

targeting users of Claude, Gemini, and other platforms in approximately balanced numbers, likely requiring samples on the order of several thousand, would test platform-level differences with greater resolution.

A second limitation concerns the reliance on self-reported performance and satisfaction measures, which are vulnerable to halo effects of the kind documented by Wirtz and Bateson (1995) and confirmed in the AI-attitude finding reported in Section 5.2.1. Behavioral measures, such as actual usage frequency, task completion rates, retention across model releases, and switching behavior between platforms, would complement the self-report instrument and would speak directly to whether perceived performance translates into observable user behaviour.

Several measurement caveats also warrant acknowledgement. The MIL1 and MRB1 cross-loadings reported in Section 4.2 identify boundary items at the conceptual seam between factors. As noted in Section 3.3, the three-factor structure is theoretically grounded in Kim et al.'s (2024) taxonomy but was operationalized through the present study's explained interpretation of how the 19 original codes aggregate into broader dimensions. While this interpretation is defensible, the cross-loadings suggest that the mapping between codes and factors is not uniform across all 13 items. Subsequent scale-development work could refine these items by partitioning them into multiple sub-dimensions or by reformulating them.

A related concern involves discriminant validity. The Fornell-Larcker criterion, which requires the square root of a factor's average variance extracted to exceed its correlations with other factors, was not formally assessed, as this procedure falls outside the quantitative scope of this thesis, which employed exploratory rather than confirmatory factor analysis. However, the inter-factor correlations, particularly CF-MIL at $r = .79$, exceed the threshold at which two constructs risk measuring the same underlying dimension. The three-factor structure is retained on theoretical grounds, given that CF, MIL, and MRB capture conceptually distinct sources of dissatisfaction rooted in communication, model capability, and behavioral alignment respectively. Findings comparing CF and MIL in particular should nonetheless be interpreted with this caveat in mind.

A related consideration concerns the presentation of performance items in three blocked pages, one per factor, each preceded by a framing sentence. While this structure reduced

cognitive load and was empirically motivated by the convenience pilot's poor CF reliability under unblocked presentation (α = .476), it cannot be ruled out that within-factor response similarity was partly inflated by common method variance arising from the blocked format rather than reflecting genuine construct coherence. Randomized item presentation would provide a stronger test of the factor structure, though at the cost of the contextual clarity that the framing sentences provided. Additionally, MRB's weak predictive coefficient may partly reflect a measurement artifact rather than purely an asymmetric-attribute effect. Users may have less coherent affective responses to behavioral attributes than to output-quality attributes, producing noisier ratings that attenuate the regression coefficient. The factor's slightly lower internal consistency (α = .789, compared with α = .893 for MIL and α = .841 for CF) is somewhat consistent with this alternative. The asymmetric-attribute interpretation and the measurement-noise interpretation are not mutually exclusive, and the present design cannot fully discriminate between them.

A final limitation concerns how the general AI attitude control was operationalized. The three substantive factors and the satisfaction measure are multi-item scales with established internal consistency (Section 4.2); the AI-attitude measure has no comparable reliability estimate, and a single-item measure is more vulnerable to interpretation drift and to response-set effects than a multi-item composite. The finding that AI attitude accounts for a large share of variance in overall satisfaction (Section 5.2.1) is therefore best read as a methodological signal and not as a finished theoretical claim. Replication with a multi-item AI-attitude scale would clarify how much of the residual baseline variance is genuinely attitudinal and how much is artifactual.

References

Anthropic. (2026). *Claude* [Large language model]. Retrieved May 16, 2026, from https://claude.ai

Bell, E., Bryman, A., & Harley, B. (2022). *Business research methods* (6th ed.). Oxford University Press.

Bowman, S. R. (2023). *Eight things to know about large language models* (arXiv:2304.00612). arXiv. https://arxiv.org/abs/2304.00612

Burtsev, M., Reeves, M., & Job, A. (2024, January 29). The working limitations of large language models. *MIT Sloan Management Review*. https://sloanreview.mit.edu/article/the-working-limitations-of-large-language-models/

Chatterji, A., Cunningham, T., Deming, D., Hitzig, Z., Ong, C., Shan, C., & Wadman, K. (2025). *How people use ChatGPT* [Working paper]. OpenAI; Duke University; Harvard University. http://dx.doi.org/10.2139/ssrn.5487080

Cuadra, A., Wang, M., Stein, L. A., Jung, M. F., Dell, N., Estrin, D., & Landay, J. A. (2024). The illusion of empathy? Notes on displays of emotion in human-computer interaction. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*, Article 3642336. Association for Computing Machinery. https://doi.org/10.1145/3613904.3642336

DeepSeek. (2026). *DeepSeek* [Large language model]. Retrieved May 16, 2026, from https://chat.deepseek.com

Eliasson, A. (2018). *Kvantitativ metod från början* (4th ed.). Studentlitteratur.

Epoch AI. (2026). *Data on AI companies* [Data set]. Retrieved May 15, 2026, from https://epoch.ai/data/ai-companies

Espinosa, C. (2025, December 17). *Generation Z in the UK: Statistics and facts*. Statista. Retrieved May 17, 2026, from https://www.statista.com/topics/11047/generation-z-in-the-uk/

Faber, E. (2025, June 2). Gen Zs and millennials at work: Pursuing a balance of money, meaning, and well-being. *Deloitte Insights.* https://www.deloitte.com/us/en/insights/topics/talent/2025-gen-z-millennial-survey.html

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41(4), 1149-1160. https://doi.org/10.3758/BRM.41.4.1149

Field, A. (2024). *Discovering statistics using IBM SPSS Statistics* (6th ed.). SAGE Publications.

Fornell, C. (1992). A national customer satisfaction barometer: The Swedish experience. *Journal of Marketing, 56*(1), 6-21. https://doi.org/10.1177/002224299205600103

Google. (2026). *Gemini* [Large language model]. Retrieved May 16, 2026, from https://gemini.google.com

Herzberg, F., Mausner, B., & Snyderman, B. B. (1959). *The motivation to work*. Wiley.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems, 43*(2), Article 42. https://doi.org/10.1145/3703155

Kaur, A. (2025). Echoes of agreement: Argument driven sycophancy in large language models. *Findings of the Association for Computational Linguistics: EMNLP 2025*, 22803-22812. https://aclanthology.org/2025.findings-emnlp.1241/

Kim, Y., Lee, J., Kim, S., Park, J., & Kim, J. (2024). Understanding users' dissatisfaction with ChatGPT responses: Types, resolving tactics, and the effect of knowledge level. In

Proceedings of the 29th International Conference on Intelligent User Interfaces (IUI '24) (pp. 385-404). Association for Computing Machinery. https://doi.org/10.1145/3640543.3645148

Koyuturk, C., Theophilou, E., Patania, S., Donabauer, G., Martinenghi, A., Antico, C., Telari, A., Testa, A., Bursic, S., Garzotto, F., Hernandez-Leo, D., Kruschwitz, U., Taibi, D., Amenta, S., Ruskov, M., & Ognibene, D. (2025). Understanding learner-LLM chatbot interactions and the impact of prompting guidelines. In A. I. Cristea, E. Walker, Y. Lu, O. C. Santos, & S. Isotani (Eds.), *Artificial intelligence in education* (Lecture Notes in Computer Science, Vol. 15878, pp. 364-377). Springer. https://doi.org/10.1007/978-3-031-98417-4_26

Malhotra, N. K. (2019). *Marketing research: An applied orientation* (7th ed., Global ed.). Pearson Education.

Martilla, J. A., & James, J. C. (1977). Importance-Performance Analysis. *Journal of Marketing, 41*(1), 77-79. https://doi.org/10.2307/1250495

Oliver, R. L. (1980). A cognitive model of the antecedents and consequences of satisfaction decisions. *Journal of Marketing Research, 17*(4), 460-469. https://doi.org/10.2307/3150499

OpenAI. (2023). *GPT-4 technical report* (arXiv:2303.08774). arXiv. https://arxiv.org/abs/2303.08774

OpenAI. (2024). *Introducing ChatGPT search*. https://openai.com/index/introducing-chatgpt-search/

OpenAI. (2025). *GPT-5 system card.* https://cdn.openai.com/gpt-5-system-card.pdf

OpenAI. (2026). *ChatGPT* [Large language model]. Retrieved May 16, 2026, from https://chatgpt.com

Parasuraman, A., Zeithaml, V. A., & Berry, L. L. (1988). SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality. *Journal of Retailing, 64*(1), 12-40.

Peer, E., Brandimarte, L., Samat, S., & Acquisti, A. (2017). Beyond the Turk: Alternative platforms for crowdsourcing behavioral research. *Journal of Experimental Social Psychology, 70*, 153-163. https://doi.org/10.1016/j.jesp.2017.01.006

Prolific Academic Ltd. (2026, January 22). *Participant terms.* Prolific. https://prolific.notion.site/Participant-Terms-700436ee4d6141b2a22335ea7e450baf

Rapp, A., Di Lodovico, C., & Di Caro, L. (2025). How do people react to ChatGPT's unpredictable behavior? Anthropomorphism, uncanniness, and fear of AI: A qualitative study on individuals' perceptions and understandings of LLMs' nonsensical hallucinations. *International Journal of Human-Computer Studies, 198*, 103471. https://doi.org/10.1016/j.ijhcs.2025.103471

Skjuve, M., Følstad, A., & Brandtzæg, P. B. (2023). The user experience of ChatGPT: Findings from a questionnaire study of early users. In *Proceedings of the 5th International ACM Conference on Conversational User Interfaces (CUI '23)* (pp. 1-10). Association for Computing Machinery. https://doi.org/10.1145/3571884.3597144

Söderlund, M. (2018). *Experiments in marketing.* Studentlitteratur.

Söderlund, M. (2024). 'I do not know': An examination of reactions to virtual agents that fail to answer the user's questions. *The International Review of Retail, Distribution and Consumer Research, 34*(2), 228-250. https://doi.org/10.1080/09593969.2023.2253003

Wirtz, J., & Bateson, J. E. G. (1995). An experimental investigation of halo effects in satisfaction measures of service attributes. *International Journal of Service Industry Management, 6*(3), 84-102. https://doi.org/10.1108/09564239510091358

xAI. (2023, November 3). *Grok*. https://x.ai/news/grok

Zamfirescu-Pereira, J. D., Wong, R. Y., Hartmann, B., & Yang, Q. (2023). Why Johnny can't prompt: How non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the*

2023 CHI Conference on Human Factors in Computing Systems (CHI '23), Article 437. Association for Computing Machinery. https://doi.org/10.1145/3544548.3581388

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., Liu, P., Nie, J.-Y., & Wen, J.-R. (2023). *A survey of large language models* (arXiv:2303.18223). *arXiv*. https://doi.org/10.48550/arXiv.2303.18223

Appendix

Appendix A: Survey Instrument

The table below shows every item presented to respondents, in the order they appeared, grouped by the Qualtrics page. Each dark bar marks a page break in the live instrument.

Question	Survey Item / Wording	Response Format
page 1 / 11 Introduction and Consent		
Welcome screen text introduces the study (Stockholm School of Economics authorship, approx. 4 minutes, anonymity, GDPR, withdrawal by closing the browser). The consent question below appears on the same page.		
Q1	Do you consent to participate in this study?	1 = I agree and wish to participate 2 = I do not wish to participate (terminates the survey)
page 2 / 11 Section 1: Screening (chatbot use + most-used selection)		
Q2_1 ... Q2_7	Which general-purpose AI chatbot(s) have you used in the past 30 days? (Select all that apply.) Q2_6_TEXT records the free-text entry for "Other".	Multi-select. Each option coded 1 = Selected, 0 = Not selected. Q2_1 ChatGPT Q2_2 Claude Q2_3 Gemini (formerly Google Bard) Q2_4 Microsoft Copilot Q2_5 Perplexity Q2_6 Other (free text) Q2_7 None of the above
Q3	Which of the general-purpose AI chatbots you selected above have you used the MOST in the past 30 days? Please select only one. All questions from this point onward will be about this specific chatbot. (Q3_6_TEXT records the free-text entry for "Other".) Display logic: shown only if more than one chatbot was selected in Q2; respondents who selected exactly one chatbot skip this item and the single selected chatbot is carried forward automatically.	1 = ChatGPT 2 = Claude 3 = Gemini 4 = Microsoft Copilot 5 = Perplexity 6 = Other (free text)

page 3 / 11 Access Type		
Q4	How do you access the general-purpose AI chatbot you selected?	1 = Free version (no subscription) 2 = Paid subscription, self-paid 3 = Paid subscription, covered by employer or school 4 = Not sure
page 4 / 11 Communicative Friction (CF) matrix and Attention Check		
Block stem: "The following questions are about how well the chatbot you selected communicates with you. Please rate its performance on each aspect, based on your experiences over the past 30 days. Use the scale from 1 to 10, where: 1 = Very low performance, 10 = Very high performance."		
Q5_CF1	The chatbot's ability to understand what I am asking for	1 to 10 (1 = Very low performance, 10 = Very high performance)
Q5_CF2	The chatbot's ability to account for my specific context	1 to 10 (1 = Very low performance, 10 = Very high performance)
Q5_CF3	The chatbot's tone and communication style	1 to 10 (1 = Very low performance, 10 = Very high performance)
Q5_CF4	The appropriateness of the length and detail in the chatbot's responses	1 to 10 (1 = Very low performance, 10 = Very high performance)
Q6	Please select 5 on the scale below to show that you are paying attention. (Attention check, shown on the same page as the CF matrix.)	1 to 10 (correct response: 5)
page 5 / 11 Model-Inherent Limitations (MIL) matrix		
Block stem: "The following questions are about the quality and accuracy of the chatbot you selected. Please rate its performance on each aspect, based on your experiences over the past 30 days. Use the scale from 1 to 10, where: 1 = Very low performance, 10 = Very high performance."		
Q7_MIL1	The chatbot's ability to give specific rather than vague responses	1 to 10 (1 = Very low performance, 10 = Very high performance)

Q7_MIL2	The originality of the chatbot's responses	1 to 10 (1 = Very low performance, 10 = Very high performance)
Q7_MIL3	The completeness of the information the chatbot provides	1 to 10 (1 = Very low performance, 10 = Very high performance)
Q7_MIL4	The accuracy of the information the chatbot provides	1 to 10 (1 = Very low performance, 10 = Very high performance)
Q7_MIL5	The chatbot's ability to avoid fabricating information	1 to 10 (1 = Very low performance, 10 = Very high performance)
page 6 / 11 Misaligned Response Behavior (MRB) matrix		
Block stem: "The following questions are about how the chatbot you selected acts in conversations. Please rate its performance on each aspect, based on your experiences over the past 30 days. Use the scale from 1 to 10, where: 1 = Very low performance, 10 = Very high performance."		
Q8_MRB1	How well the chatbot explains the reasons and logic behind its responses	1 to 10 (1 = Very low performance, 10 = Very high performance)
Q8_MRB2	The chatbot's willingness to give its own opinion when asked	1 to 10 (1 = Very low performance, 10 = Very high performance)
Q8_MRB3	The chatbot's willingness to discuss difficult or controversial topics	1 to 10 (1 = Very low performance, 10 = Very high performance)
Q8_MRB4	The chatbot's ability to remain objective rather than conforming to my views	1 to 10 (1 = Very low performance, 10 = Very high performance)
page 7 / 11 Overall Satisfaction (Fornell 1992 barometer)		
Block stem: "The following questions are about your overall experience with general-purpose AI chatbots broadly, not just the one you selected above. Please answer based on your experiences."		
Q9_SAT1	Considering all your experiences with AI chatbots, how satisfied are you overall?	1 = Not at all satisfied; 10 = Very satisfied
Q9_SAT2	To what degree do AI chatbots fulfil all your expectations?	1 = Much less than expected; 10 = Much more than expected

Q9_SAT3	Imagine an AI chatbot which is perfect in all respects. How close to this ideal do you consider AI chatbots to be?	1 = Very far away from ideal; 10 = Very close to ideal
page 8 / 11 Demographics (age, gender, education, frequency, AI attitude)		
Q10	What is your age?	Numeric, years (eligibility filter: 18 to 28)
Q11	What is your gender?	1 = Male 2 = Female 3 = Non-binary / gender diverse 4 = Prefer not to say
Q12	What is the highest level of education you have completed or are currently pursuing?	1 = Secondary school / high school (or equivalent) 2 = Some university or college (no degree yet) 3 = Bachelor's degree 4 = Master's degree or higher 5 = Other
Q13	How often do you use AI chatbots?	1 = Daily 2 = Several times a week 3 = About once a week 4 = A few times a month 5 = Less than once a month
Q14	How would you describe your general attitude toward AI as a technology?	1 = Very negative; 10 = Very positive
page 9 / 11 Topic Recall Check		
Q15	What was this survey primarily about? (Topic-recall attention check.)	Single-select: Online shopping experiences / AI-powered fitness coaching / AI chatbot user experiences / Social media usage habits. Correct response: AI chatbot user experiences. Used as exclusion filter alongside Q6.
page 10 / 11 Prolific ID		

Q16 / PROLIFIC_PID	What is your Prolific ID? Please note that this response should auto-fill with the correct ID. (Auto-populated from the Prolific piped field ${e://Field/PROLIFIC_PID}.)	String, auto-populated from Prolific.
page 11 / 11 Debrief and End of Survey		
End-of-survey screen thanks the participant ("We thank you for your time spent taking this survey. Your response has been recorded.") and triggers the redirect back to Prolific. No screenshot of this page was captured; structure inferred from the Qualtrics End-of-Survey block default.		

Note. Question labels (Q1, Q2_1 etc.) follow the analysis-file variable names verbatim. Qualtrics editor-side question numbers differ from analysis-file numbers because hidden timing-question placeholders inflate the editor sequence; the data dictionary recodes them. Page breaks for Pages 1 through 10 are confirmed by direct inspection of the deployed survey (one screenshot per page in dump/survey screenshots updated/). Only Page 11 (End of Survey) is inferred.

Appendix B: Kim et al.'s (2024) Dissatisfaction Taxonomy

This appendix presents the complete taxonomy of user-side dissatisfaction with general-purpose AI chatbots developed by Kim et al. (2024), which served as the empirical foundation for scale construction in this study (Section 3.3.1). The taxonomy was derived from a systematic literature review of 59 papers and applied to categorize dissatisfaction instances from 94 participants' chatlogs. The taxonomy identifies seven broad categories encompassing 19 specific dissatisfaction codes.

Table B1. *Kim et al.'s (2024) Seven Categories of User-Side Dissatisfaction*

Category	Code	Description
Intent Understanding (D_intent)	C1	Response does not meet users' intent or instruction.
Intent Understanding (D_intent)	C2	Response is not aligned with the user's context.
Intent Understanding (D_intent)	C17	The tone or communication style is disappointing.
Content Depth and Originality (D_depth)	C3	Response is too general.

Content Depth and Originality (D_depth)	C4	Response lacks originality.
Content Depth and Originality (D_depth)	C5	Response lacks information.
Information Accuracy (D_acc)	C6	The response contains incorrect information.
Information Accuracy (D_acc)	C7	Response is based on training data cut off at a certain date, and has limited access to newly created data.
Information Accuracy (D_acc)	C8	Response is inconsistent.
Information Accuracy (D_acc)	C9	ChatGPT struggles with reasoning.
Information Accuracy (D_acc)	C10	(Hallucination) ChatGPT fabricates contents that conflict with the source content or cannot be verified from existing sources.
Information Accuracy (D_acc)	C19	(Sycophancy) ChatGPT excessively conforms to the user.
Transparency (D_trans)	C11	It's difficult to understand the reasons, criteria, logic, and evidence behind the responses.
Refusing to Answer (D_refuse)	C12	ChatGPT avoids giving its own opinion by saying something similar to "As a language model, I am not capable..."

Refusing to Answer (D_refuse)	C13	ChatGPT avoids talking about difficult or controversial issues by saying something similar to "As a language model, I am not capable..."
Refusing to Answer (D_refuse)	C7	Response is based on training data cut off at a certain date, and has limited access to newly created data.
Content Ethics and Integrity (D_ethic)	C14	Response contains unlawful content.
Content Ethics and Integrity (D_ethic)	C15	Response contains unethical, harmful content.
Content Ethics and Integrity (D_ethic)	C16	Response contains biased content.
Response Format and Attitude (D_format)	C17	The tone or communication style is disappointing.
Response Format and Attitude (D_format)	C18	Response is overly detailed or too long.
Response Format and Attitude (D_format)	C19	(Sycophancy) ChatGPT excessively conforms to the user.

Note. Adapted from Kim et al. (2024). Some codes (C7, C17, C19) appear in multiple categories in the original taxonomy, reflecting their cross-cutting nature across dissatisfaction dimensions.

Appendix C: Measurement Properties and Descriptive Statistics

This appendix provides complete descriptive statistics and psychometric properties supporting the analyses in Chapter 4. Tables C1-C2 describe sample demographics and chatbot usage. Tables C3-C5 document scale reliability, factor structure, and inter-factor correlations. Table C6 provides item-level descriptive statistics for all survey measures.

Table C1. *Demographic Profile (N = 293)*

Variable	Category	n	%
Gender	Male	**146**	**49.8%**
	Female	**146**	**49.8%**
	Non-binary / gender diverse	**1**	**0.3%**
Education	Secondary school / high school	**47**	**16.0%**
	Some university or college (no degree yet)	**76**	**25.9%**
	Bachelor's degree	**120**	**41.0%**
	Master's degree or higher	**50**	**17.1%**
Usage frequency	Daily	**118**	**40.3%**
	Several times a week	**108**	**36.9%**
	About once a week	**26**	**8.9%**
	A few times a month	**25**	**8.5%**
	Less than once a month	**16**	**5.5%**

Note. Percentages may not sum to 100 due to rounding.

Table C2. *Most-Used Chatbot (CHATBOT_USED, N = 293)*

Chatbot	n	%
ChatGPT	**208**	**71.0%**
Claude	**24**	**8.2%**
Gemini	**29**	**9.9%**
Microsoft Copilot	**25**	**8.5%**
Perplexity	**3**	**1.0%**
Other	**4**	**1.4%**

Note. Percentages may not sum to 100 due to rounding.

Table C3. *Cronbach's Alpha for Each Scale: Pre-Pilot (N = 43) and Main Study (N = 293)*

Scale	Items	Pre-Pilot α	Main Study α
Communicative Friction (CF)	4	.476	**.841**
Model-Inherent Limitations (MIL)	5	.817	**.893**
Misaligned Response Behavior (MRB)	4	.770	**.789**
Overall Satisfaction (SAT)	3	.719	**.902**

Table C4. *Pattern Matrix (Principal Axis Factoring, Promax Rotation, 3 Factors Fixed)*

Item	F1 (CF)	F2 (MIL)	F3 (MRB)
CF1, understand what I am asking	**.850**		
CF2, account for context	**.750**		
CF3, tone and style	**.634**		
CF4, appropriateness of length	**.702**		
MIL1, give specific not vague (cross-loads)	**.471**		
MIL2, originality		**.643**	
MIL3, completeness		**.703**	
MIL4, accuracy		**.885**	
MIL5, avoid fabricating		**.667**	
MRB1, explains reasons (no primary > .40)			
MRB2, gives opinion			**.764**

MRB3, discusses controversy			**.773**
MRB4, remains objective			**.460**

Note. Pattern-matrix loadings below .40 are suppressed in line with SPSS default reporting (Field, 2024). MIL1 and MRB1 are discussed in Section 4.2 as cross-loading caveats.

Table C5. *Factor Correlation Matrix*

	F1 (CF)	**F2 (MIL)**	**F3 (MRB)**
F1 (CF)	**1.000**	**.792**	**.592**
F2 (MIL)	**.792**	**1.000**	**.690**
F3 (MRB)	**.592**	**.690**	**1.000**

Table C6. *Item-Level Descriptive Statistics (N = 293)*

Item	**Factor**	**M**	**SD**
CF1, understand what I am asking	CF	**7.73**	**1.50**
CF2, account for specific context	CF	**7.37**	**1.68**
CF3, tone and communication style	CF	**7.67**	**1.77**
CF4, appropriateness of length	CF	**7.57**	**1.79**
MIL1, give specific not vague	MIL	**7.56**	**1.72**
MIL2, originality	MIL	**6.37**	**2.04**
MIL3, completeness	MIL	**7.29**	**1.72**
MIL4, accuracy	MIL	**6.93**	**1.77**
MIL5, avoid fabricating	MIL	**6.39**	**2.12**
MRB1, explains reasons	MRB	**7.04**	**1.92**

MRB2, gives opinion when asked	MRB	**6.30**	**2.34**
MRB3, discusses controversy	MRB	**6.06**	**2.32**
MRB4, remains objective	MRB	**6.20**	**2.27**
SAT1, overall satisfaction	SAT	**7.25**	**1.95**
SAT2, fulfils expectations	SAT	**7.05**	**1.97**
SAT3, closeness to ideal	SAT	**6.40**	**2.07**

Table C7. *Pearson Correlations Between Composites and AI Attitude (N = 293)*

	CF	**MIL**	**MRB**	**SAT**	**AI Attitude**
CF	**1**	**.751**	**.565**	**.712**	**.545**
MIL		**1**	**.683**	**.767**	**.586**
MRB			**1**	**.682**	**.582**
SAT				**1**	**.746**
AI Attitude					**1**

Note. All off-diagonal correlations are significant at **p < .001**.

Appendix D: AI Disclosure

What tools were used and how. ChatGPT and Claude were used as language and structuring tools throughout the writing process. This included rephrasing sentences for clarity, checking grammar, improving readability, and helping organize text within sections. AI was also used to assist with brainstorming during the early stages of the project or when feeling stuck.

How they improved the work. The tools helped improve the overall readability and flow of the text. They were useful for catching awkward phrasing and for getting a second perspective on whether a paragraph communicated its point clearly. This saved time on editing and allowed us to focus more on the substance of the research.

Risks identified and how they were handled. AI tools are unreliable when handling academic sources and can fabricate references or misrepresent the content of articles. For this reason, AI was never used to generate or verify sources. All references were independently found, read, and assessed by the authors. All factual claims and statistics in the thesis were verified against the original source material. No AI-generated content was included without review and revision by the authors.

What we learned. AI is a useful support tool for language and structuring but cannot replace the analytical thinking required in academic work. It works best when used to refine ideas that already exist rather than to generate new ones. We found that maintaining a critical approach to all AI output was essential throughout the process.