INSTITUTO
UNIVERSITÁRIO
DE LISBOA

End Users' Interactions with ChatGPT, its Impact on Trust and Emotion

Marijose Páez Velázquez

Master Degree in Emotion Sciences

Thesis adviser:
Doctor Elzbieta Malgorzata Bobrowicz Campos, Guest Assistant Researcher, ISCTE - Instituto Universitário de Lisboa

October, 2025

CIÊNCIAS SOCIAIS
E HUMANAS

Psychology Department

End Users' Interactions with ChatGPT, its Impact on Trust and Emotion

Marijose Páez Velázquez
Master Degree in Emotion Sciences

Thesis Adviser:
Doctor Elzbieta Malgorzata Bobrowicz Campos, Guest Assistant Researcher, ISCTE - Instituto Universitário de Lisboa

October, 2025

Um coração que mora dentro do olho do jaguar

O Verão, rapazes – como disse C. Adams –
implica uma insistência nos mergulhos
e uma desistência breve das respostas.
Importante é passar a mão pelas escarpas,
afagar o pescoço das andorinhas do mar,
verificar o oxigénio no tubinho de plástico
que ajuda a respirar na cala azul turquesa
e permitir que o Senhor ressuscite o sangue
dos espadartes a todas as manhãs de 29 ºC.
Estas são as tarefas que devem ser realizadas
e – como disse Adams – bom mesmo é chegar
ao fim da estação sem nenhuma resposta.

[A heart that lives inside the jaguar's eye

Summer, boys – as C. Adams once said –
demands an insistence on diving
and a brief renunciation of answers.
What matters is to run your hand along the cliffs,
to caress the necks of the sea swallows,
to check the oxygen in the little plastic tube
that helps you breathe in the turquoise cove,
and to allow the Lord to resurrect the blood
of the swordfish every morning at 29 °C.
These are the tasks that must be carried out
and – as Adams said – the best thing of all
is to reach the end of the season without a single answer.]

(Campilho, 2014, p. 115)

Appreciations

Thank you to my supervisor, Professor Doctor Elzbieta Bobrowicz-Campos, for her guidance and support throughout this long, challenging process with patience and comprehension; for allowing me the freedom to explore ideas independently and the trust in letting me find my own path.

Thank you to Professor Doctor Patrícia Arriaga, for all her dedication, attention, and collaborative spirit throughout this process. Her passion and deep knowledge greatly contributed to the development and consolidation of this research work.

Thank you to my dear Sinah Tiller,
for this academic and personal journey that we experienced together;
for the joint curiosity, moderator of our adventures!
Danke schön, bela.

To Rui, o poeta de seis nomes, obrigada.
Thank you for your support, care, attention, and resilience
through all the writing process and throughout everything around us.

Always and forever, thank you to my family.
Ari, Thaly, Cielin, gracias.
Thank you to all the women that have lectured me about love, strength, courage, and lyrics.
Gracias a mi familia. Gracias por su amor, apoyo incondicional y ánimos.
Gracias por los constantes abrazos, tan grandes que atraviesan el océano.

Author's Notes

The Study One "Towards a Typology of Prompts" was presented at the 12th International Conference on Multimedia and Human-Computer Interaction (MHCI 2025) in Paris, and subsequently published (Velazquez et al., 2025) in the conference proceedings. Afterwards, an extended version was accepted in the Journal of Machine Intelligence and Data Science (forthcoming publication). The present manuscript offers an extended and revised version of that text and further advances the work by introducing the subsequent *Main Study*.

ChatGPT 4.o free version and Gemini 2.5 Pro were used to check grammar in some parts of this paper. No ideas, references, nor full written sections were obtained through any Generative AI. The author takes full responsibility for the accuracy and integrity of this manuscript, including each substantive claim and supporting evidence.

Any use of Generative AI in this manuscript adheres to ethical guidelines for use and acknowledgement of Generative AI in academic research. In line with reproducibility and transparency principles, the paper specifies in detail how Generative AI was used during the research process within the relevant sections (Abdurahman et al., 2025; Porsdam Mann et al., 2024).

Resumo (pt)

Interações com Inteligência Artificial estão a tornar-se cada vez mais integradas na vida quotidiana, levantando questões fundamentais sobre a forma como os utilizadores constroem confiança e são geradas respostas emocionais ao interagir com agentes conversacionais. Com base no enquadramento teórico da inferência ativa, esta investigação explora como a confiança emerge a partir dos processos preditivos dos utilizadores quando interagem com Modelos de Linguagem de Grande Escala. O trabalho compreende dois estudos consecutivos. O *Estudo Um* desenvolveu e avaliou características dos *prompts*, incluindo os tipos de interação que promovem, através de uma abordagem mista. Recorreu-se a uma análise temática e à avaliação por utilizadores não técnicos, resultando na seleção de 12 prompts classificados como *task-oriented* (orientados para tarefas) ou *reflexive* (reflexivos), avaliando níveis comparáveis de complexidade e adequação a uma população não especializada. Com base nestes resultados, o *Estudo Principal* investigou se o tipo de interação (orientada para tarefas vs. reflexiva) com ChatGPT tinha efeito na confiança e se esse efeito era explicado pela previsibilidade. Adicionalmente, examinou-se a relação entre a precisão das predições e as respostas emocionais. Foi utilizado um desenho experimental quantitativo pré-pós, com uma estrutura de 2 (tipo de interação) × 2 (grupo) e comparações intra e inter grupos. Os participantes interagiram com ChatGPT, utilizando os prompts avaliados no *Estudo Um*, completando inquéritos por questionário antes e depois de cada interação. Os resultados indicaram que o tipo de interação teve efeito na confiança e que a previsibilidade explicou variações na confiança, especificamente nas segundas interações. As respostas emocionais revelaram correlações significativas com a precisão das predições apenas nas interações reflexivas, sugerindo que este tipo de interação pode constituir uma nova forma de interação com ChatGPT que potencia confiança e resposta emocional. Em conjunto, estes resultados sugerem que o tipo de interação e a previsibilidade moldam a confiança e a emoção nas interações humano–Inteligência Artificial, contribuindo para a integração da inferência ativa na investigação sobre Interação Humano–Computador.

Keywords: Modelos de Linguagem de Grande Escala (LLMs); interação humano–IA; confiança; inferência ativa; emoção; literacia em IA

Abstract (En)

Interactions with Artificial Intelligence are becoming increasingly integrated into everyday life, raising fundamental questions about how users build trust and emotional responses when engaging with conversational agents. Grounded in the active inference framework, this research explores how trust emerges through users' predictive processes when interacting with Large Language Models. The work comprises two consecutive studies. Study One developed and assessed prompt characteristics and types of interaction through a mixed qualitative–quantitative approach. Using thematic analysis and lay users' evaluations, 12 prompts were selected and classified as either task-oriented or reflexive, ensuring comparable complexity and suitability for non-expert populations. Building on these findings, the *Main Study* investigated whether the type of interaction (task-oriented vs. reflexive) with ChatGPT had an effect on trust, and whether this effect was explained by predictability. Additionally, it examined the relationship between prediction accuracy and emotional responses. A quantitative pre–post experimental design was employed, with a 2 (interaction type) × 2 (group) structure and both within- and between-subjects comparisons. Participants engaged with ChatGPT, using the pilot-tested prompts from Study One, and completed self-assessments before and after each interaction. Results showed that type of interaction has an effect on trust and that predictability explained variations in trust, specifically for second interactions. Emotional responses were found to be significantly correlated with prediction accuracy just for reflexive interactions, suggesting that reflexive interactions might be novel types of conversations with ChatGPT that potentiate trust and emotional responses. Together, these findings suggest that interaction type and predictability might shape trust and emotion in human–Artificial Intelligence interactions, advancing the integration of active inference and Human–Computer Interaction research.

Keywords: Large Language Models (LLMs), human–AI interaction, trust, active inference, emotion, AI literacy

General Index

Index for Figures and Tables

Index for figures

Index for tables

Introduction

Interactions with Artificial Intelligence (AI) agents will continue to grow, transforming our daily lives in unprecedented ways. Large Language Models (LLMs) broaden the access to AI for users from diverse backgrounds, across many levels of expertise (Andries & Robertson, 2023; Laupichler et al., 2023a). However, research literature on the effects of LLMs is still in its infancy (Sohail et al., 2023). LLMs capabilities of information processing and production of remarkably human-like outputs in the form of conversations, at first sight give the appearance of a trustworthy, neutral and reliable output-response. Those systems will become more and more complex, and so will users' mental predictive models about the systems. In this ongoing quest to expand our capabilities through technology, trust and emotion are key, as they will enable safer and more productive interactions where users neither undertrust nor overtrust the system, achieving a better calibrated reliance on its capabilities.

The active inference framework suggests that trust in others depends on whether one can predict their behaviours/actions and, if those predictions are accurate a positive-valence emotion will be experienced (Schoeller et al., 2021). In human-machine interactions, trust is fundamental, as it predicts their success; the optimal level consists in avoiding over-reliability and under-usage. In order to interact seamlessly with technology, users must infer the causes behind the system's behavior (Sheridan, 1988) and be confident about that inference, which suggests trust is a necessary condition for human-AI collaboration. However, and despite its crucial role, it is still largely unknown how trust emerges, develops, and supports human relationship to AI systems.

In the early 1900's, Fernando Pessoa (Pessoa, n.d., ed. 1968, p. 208) wrote that *the external world is a tactile hallucination*. Later on, Einstein (1938) said that physical concepts are not uniquely determined by the external world, but are free creations of the human mind (as cited in Barrett, 2017). More recently, frameworks about brain prediction models have been challenging the classical point of view that presupposes our emotions, actions and perceptions as being mere reactions to our environment; instead, these frameworks postulate that we experience our world by actively constructing it (Hutchinson & Barrett, 2019).

The present thesis will explore trust dynamics within human-artificial intelligence interactions under the lens of the active inference framework. The questions underlying this investigation tinker with the entanglement of generative mental models, interactions with AI, trust, and emotion. We will first outline key concepts of the active inference framework and AI, and then relate them to trust and emotion. Subsequently, we will examine previous research work on AI interactions and

trust, identify current challenges and gaps in research, and propose research questions that will be investigated through two different consecutive studies. The first study aims to provide a foundation for classifying the types of interactions between lay users and LLMs, specifically ChatGPT, addressing the need for a shared understanding of prompt categorisation based on interaction type and complexity level from the user's perspective. Through this first validation, the second study aims to systematically compare emotional responses, trust dynamics, and predictability levels by the type of interaction elicited between lay users and ChatGPT. Currently very few studies in Human Computer Interaction (HCI) or trust include interactions between end users and ChatGPT as part of its experimental design; this study seeks to bridge that gap. Finally, we will discuss the results and limitations of this research, and outline directions for future work in light of the study's contributions.

As written by Fernando Pessoa in the earlry 1900s, and later published as part of his philoshical texts (1968), perception is both active and passive. It is passive as one cannot control it, it is not a phenomenom of our will. It is active as we create and imagine our sensations. Direclty quoting Pessoa:

> A certeza — isto é, a confiança no carácter objectivo das nossas percepções, e na conformidade das nossas ideias com a realidade ou a verdade — é um sintoma de ignorância ou de loucura. O homem mentalmente são não está certo de nada, isto é, vive numa incerteza mental constante (...) [Certainty — that is, to trust in the objective nature of our perceptions, and in the alignment of our ideas with reality or truth — is a symptom of ignorance or madness. A mentally sound man is certain of nothing; that is, he lives in a state of constant mental uncertainty (...)] (p. 246)

CHAPTER ONE

What we (*model to*) know

Section one: Theoretical Frameworks

1.1. Active Inference

"The embodied brain is as interested in changing the world to fit its own predictions, as it is in describing how things are" (di Paolo et al., 2024)

Nobel-prize winning Gerald Edelman states that in order to understand how the mind works, we need to consider its biological structure: the brain. According to him, the brain responds to the polymorphous diversity in nature with its own diversity in self through *reentry* (Web of Stories [WS], 2005). Our brain makes a recursive process of changing synaptic strings across massively parallel connections in different parts of the brain. This means that the brain will dynamically signal and map a stimulus across different, parallel patterns, that can be similar but are not identical, through different areas of the brain, causing an enormous repertory of variant dynamic circuits. Through developmental and experiential selection, the best connections will then be reinforced. Besides this, through a process called *degeneracy*, structurally different networks can perform the same function to allow the brain to be highly resistant to damage. In this case, different groups of neurons (not neurons themselves), can map the instances of the same category in different contexts by using parallel reciprocal connections (Barrett, 2017; Edelman & Gally, 2001; WS, 2005). This means that our brain is not just an organ but a vast network of neurons that can perform different functions at any given moment (Barrett, 2017; Rigotti et al., 2013).

Such parallel reciprocal connections reveal that neural activity is less about recording stimuli and more about inferring their meaning and being action-ready (Barrett, 2017; Miller & Clark, 2018). Edelman's account of distributed, recursive mappings lays the foundation for understanding the brain as a predictive organ that operates as a constructive system, continuously generating hypotheses through statistical possibilities, in order to diminish surprise (Barrett, 2017; Miller & Clark, 2018; Schoeller et al., 2021). It is constantly creating a mental representation of sensory inputs through perceptual inference (Ciaunica, et al., 2022), which suggests that we are self-generating our reality.

To disentangle these intricate concepts, we will examine each one, clarifying their individual contributions to the overall interrelationship.

Construction of an (internal) mental representation: Our brain is isolated from the environment where its body is inserted in. To ensure survival, it needs to create an internal map of that world, a model of the unseen causes. Note that it does not only model the world, but also itself and its body. This is known as the "generative model"[1] and includes the internal and external environments; its function is to mediate the brain-body-environment interactions (Barrett, 2017; Bruineberg et al., 2018; Christov-Moore et al., 2024).

Sensory inputs and statistical hypotheses: Our brain's mental model is grounded in what our senses detect. Using this sensory data, it makes informed guesses (predictions) about the most likely causes of the diverse phenomena it may encounter (Barrett, 2017; Schoeller et al., 2021). These educated guesses are hypotheses that appear in the form of beliefs based on statistical possibilities integrating past experiences and expectations (Hutchinson & Barrett, 2019). Bottom-up sensory signals are predicted by top-down cognitive models (Schoeller et al., 2021).

Prediction: Our brain is responsible for regulating our autonomic nervous, immune and endocrine systems, while efficiently maintaining a balance between internal and external resources (Barrett, 2017). This process of achieving stability through change is known as allostasis (McEwen & Wingfield, 2010), which differs from homeostasis, the set of physiological processes that achieve stability by maintaining equilibrium, despite changes in the external environment (Romero et al., 2009). Allostasis helps living systems adapt to new situations or challenges by accounting for both predictable and unpredictable aspects, incorporating anticipatory changes rather than purely reactive adjustments (McEwen & Wingfield, 2010). Homeostasis can explain physiological responses such as sweating in hot conditions to balance the body's temperature in a present environment, whereas allostatic behaviour accounts for finding shade before overheating (Parr et al., 2022). Other allostatic strategies include mobilizing resources before anticipated challenges to homeostasis, such as increasing cardiac rhythm before a long run or a mother starting to produce milk before the baby is actually born. The embodied simulation is not merely reactive to reality; but instead, it continuously anticipates it (Barrett, 2017).

(Expected) surprise: Our brain uses its internal model to predict what's coming next, but as our world is unexpected, our brain is prepared for a range of *uncertainty:* this is called expected surprise. When reality doesn't match the brain's top-down prediction, (unexpected) surprise arises as a prediction error or free energy (Parr et al., 2022; Schoeller et al., 2021).

Error and Update: For the sake of survival, the mental representation needs to be permanently updated as it encounters a myriad of diverse situations in the environment. The bigger the

[1]From now on, we will refer to this "generative model" simply as "mental model" or the "embodied simulation" (Barsalou, 2008) as the latter is already used in psychology (Barrett, 2017), mainly to avoid confusion with Generative Models from Artificial Intelligence.

dissonance between the internal model and reality, the bigger the prediction error. This error works as a bottom-up feedback, updating the embodied simulation (Schoeller et al., 2021). The brain, or engine of prediction, as Miller (2018; 2019) names it, seeks to minimise surprising states; updating prediction error adjusts the mental model to better fit reality. It is crucial to reduce (unexpected) surprise, or error, as this results in a better adaptation and survival of the organism. According to Darling (2023), through error updating, as we exist, act in and upon the world, the world itself will both shape and create us, concurrently, influencing our embodied simulation and our expectations, reshaping our mental model by dynamically and continuously updating our predictions based on those errors. Furthermore, embodied cognitive neuroscience proposes that the mind should not be restricted to the brain alone, as physical states in the body also influence network functions, through a mutually connected dynamic that responds to the environment that the organism encounters at any given moment (Kiverstein & Miller, 2015; Miller & Clark, 2018).

(Active) inference and perception: Our perception is an active predictive process, organized in a hierarchical stream of continuous interpretation (Schoeller et al., 2021). As the brain (seeks to) minimize the discrepancy between input and prediction (da Costa et al., 2022), in order to minimize error, it will either adjust our perception, or adjust and create a world that fits into our simulation/model (Hargrove & Miller, 2022; Miller, 2018). The model is not static: our perception of reality is dynamically updating, 'remembering the present' (du Toit, 2013). New perceptions of the past strengthen synaptic connections in the brain while at each moment, our perception of the now is compared to our perception of the past (du Toit, 2013; WS, 2005).

Action: Besides perceptual inference, our system can update the embodied simulation to better fit the incoming sensory data through action (Clark, 2013; di Paolo et al., 2024). Action is an indispensable element in minimising prediction error in the long run. It can be either goal-oriented action, that seeks to change the world to fit our predictions; or epistemic action, that seeks information-rich sources or experiences to increase our understanding of the world (di Paolo et al., 2024). We take action to explore the world around us. We take action by selecting from an array of different, challenging, and unknown situations of which we hypothesise their consequences to be "better for us" or, better aligned with our beliefs and better fit our capacity. Furthermore, we take action to better control what happens in the world surrounding us, building up our sense of agency, our perception of control. Depending on whether we feel in control of the situation or not, that sense of *agency* is then experienced as a positive or negative affect (Schoeller et al., 2021). Ultimately, our possibilities of *agency* allow us to *do otherwise* and, according to Majid (2025) this constitutes the core of our free will.

Agents: Living systems, commonly referred to as *agents*[2], are embodied models of the world composed by priors, such as beliefs (Christov-Moore et al., 2024); they do not simply *have* a model of the world, but instead *are* a model (Bruineberg et al., 2018).

As a whole, active inference describes how agents who have a body in the service of allostasis (Barrett, 2017) reduce uncertainty by continuously generating predictions and contrasting them with sensory information and priors; and then, either updating their internal model or adapting their actions to minimize prediction errors and better fit the world into their mental model. Crucially, this process is not confined to the mind: perception is built not *by* but *with* the world, since our predictive mechanisms are not limited to the brain, but extended to our body and jointly created with social systems, cultural models and digital infrastructures (Bruineberg et al., 2018; Navarro, 2025). While this extension of our mind is not a new phenomenon, recent technological developments have exponentially increased its potential. For example, a philosophical research from Stanford University (Navarro, 2025) recently proposed a "cyborg perception"[3], that stretches across brains, bodies, and technological infraestructures. Similarly, Clark (2025) proposes a "bio-technologically distributed self", built from synching AI solutions with ourselves (brain, body, and/or mind), to give birth to new forms of perception, thinking and creation.

1.2. Artificial Intelligence (AI)

"AI isn't a tool—it's an agent" (Roll & Harari, 2024)

AI systems are algorithms that capture patterns from massive amounts of data (Orrù et al., 2023), able to perform tasks that typically require human intelligence such as learning, problem-solving, or decision-making (Google Cloud, 2025). Subfields of AI include machine learning, which notably has made advancements grounded in active inference (da Costa et al., 2022); or subsets and models as neural networks, grounded in neuroscience (da Costa et al., 2022; Mazzaglia, 2022; Oliveira, 2025). Figure 1.1 outlines some of the most important elements of AI for visual support.

The use of technological tools allows us to reduce prediction error, as we can achieve goals more effectively than we would on our own, without these *technological extensions* (Schoeller et al., 2021). We might, for example, get more accurate feedback from the world and adapt faster through a navigation app; diminish cognitive load and get better results using a calculator; or better anticipate and retrieve events through an intelligent calendar or reminder system. This is why we are

[2] Authors from psychology refer to organic, living systems as *agents*, while neuroscience or technology authors include in that term artificial systems. In this research, we will use the term for both organic and artificial ones.

[3] Through the hyper-visible smartphone's integration into our perception, Navarro (2025) explains in this thesis how technology can be integrated as an extension of our bodies and mental capacities.

attracted to using them in the first place, and ultimately, they might contribute to our sense of agency. Complex systems such as AI, with rapid and remarkably advanced performance capabilities, may further potentiate the reduction of prediction error and our willingness to rely on them for the same reason, just as the smartphone did in relatively quick time changing our practices and interactions with the world (Navarro, 2025).

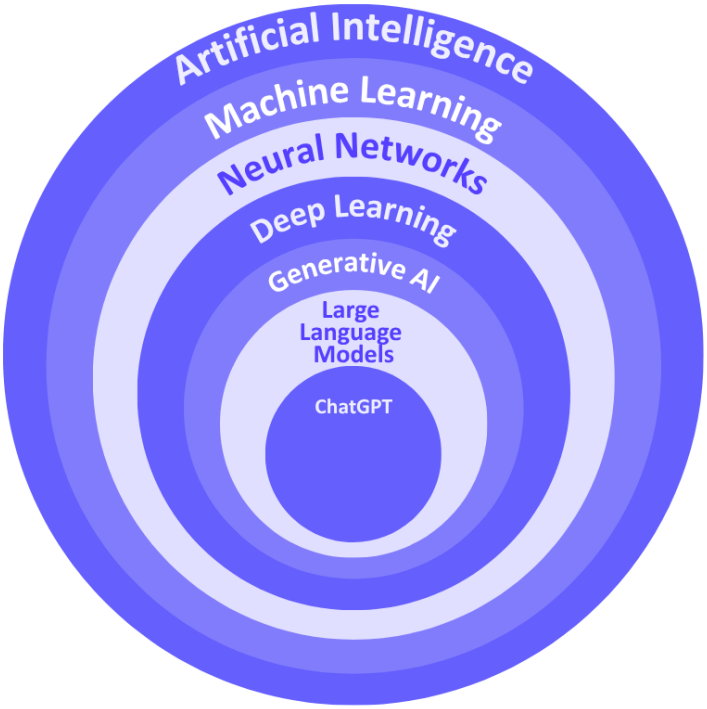

Figure 1.1. AI subfields, subsets, and applications relevant to LLMs.
Adapted from Google Cloud (2025); Kuntz & Wilson (2022)

AI systems might resemble active inference as they share some precepts, such as a "generative model" or emphasis in prediction (Pezzulo et al., 2024). However, current AI systems lack essential aspects of active inference: they learn passively, as a result of training (di Paolo et al., 2024); and do not possess sensorimotor interactions with the world, sense of agency, purposive action-oriented policies, nor intrinsic goals or motivations (Pezzulo et al., 2025). AI is passive, reactively learning through large sets of data (Pezzulo et al., 2024), whereas a living organism's intelligence is actively engaging, interpreting and changing its environment.

1.2.1. Large Language Models (LLMs)

Machine Learning algorithms have evolved into highly sophisticated systems, such as Deep Learning (Johnsen, 2025; Oliveira, 2025). LLMs encompass a range of architectures and training approaches; among them, autoregressive LLMs are a subclass of Deep Learning systems built upon the

Transformer architecture, accessible through user-friendly interfaces such as ChatGPT, Claude, or Gemini (Johnsen, 2025; Liu et al., 2023). Namely, autoregressive LLMs are models trained on massive amounts of data that are able to process, understand and generate human-like language. This impressive performance in human-like capabilities (Pezzulo et al., 2024) has evolved to the point where it visibly diminishes human ability to detect text generated by the GPT model, in comparison to text written by another human, already approaching chance level (~52%) (Brown et al., 2020). Autoregressive LLMs are not only able to statistically predict the next word in a sentence, but through attention mechanisms, they can also selectively focus on distant portions of text to achieve a more comprehensive contextual understanding (Abdurahman et al., 2025; Liu et al., 2023). This has led LLMs to an unprecedented freedom of applications in several different contexts (for more information, see Johnsen, 2025).

Performance gains in LLMs can be achieved by increasing the number of parameters and the size of the training dataset (Naveed, 2023). For instance, between 2019 and 2020, OpenAI (Brown et al., 2020) scaled its LLM model from 1,500 million (GPT-2) to 175,000 million parameters (GPT-3). The official number of parameters in GPT-4 (2023-2024) and GPT-5 (June 2025) has not been disclosed, with unofficial speculations calculating 1.76 trillion for GPT-4 and up to 10 trillion for GPT-5. Regarding the training dataset, GPT-3 was trained with almost 500,000 million tokens, a volume of text equivalent to what would take a human being nearly 5,000 years of continuous, round-the-clock reading to complete (Oliveira, 2025). Official details for the rest of the GPT models have not been disclosed. Nevertheless, it is important to note that GPT is not necessarily the model with the largest amount of parameters or training tokens amongst LLMs (Naveed, 2023).

Besides scaling, other additional improvements, such as alignment with human preferences, have been made, contributing to better, more accurate outputs (Naveed, 2023; Oliveira, 2025). According to OpenAI (2025), ChatGPT-5 minimized sycophancy, introduced four pre-set personalities (Cynic, Robot, Listener, and Nerd), and handles deception better, by communicating its limitations transparently when it's missing key information. This leads to more accurate outputs, potentially higher customer satisfaction, and user engagement (Shumanov & Johnson, 2021; Xu et al., 2022). While the history of language models and other forms of AI is not new (see Oliveira, 2025 for a review of this evolution), recent developments have created a paradigm shift (Orrù et al., 2023): besides the rapid rhythm of performance improvements in AI tools (see Kwa et al., 2025 for a historical review from different AI agents; see product's new features in OpenAI, 2025 and OpenAI, 2024), LLMs and other forms of generative artificial intelligence models enable a non-expert[4]

[4] According to Laupichler et al. (2023b), non-experts are "individuals who have not received formal training in AI and are using AI applications rather than developing them" (p. 2).

audience to actively create various types of content (as text and graphics) through simple conversational inputs (prompts).

This ability to create new content has come to public discourse (Ronge et al., 2025) as the relatively young term of "Generative Artificial Intelligence" (GenAI), which must not be mistaken for the "generative model" from active inference, explained in the previous section. Although the definition of GenAI may vary across audiences, several authors identify a) creating novel content, b) relevant usefulness of outputs, c) multi-modality and d) flexibility, as core characteristics of GenAI, highly represented by LLMs. AI systems have been integrated into entertainment, healthcare, and customer service (Andries & Robertson, 2023; Bagozzi et al., 2022; Castro et al., 2024; Oliveira, 2025) for years. However, the development of LLMs introduces new dimensions to the lay population, making it fundamental in understanding the types of interactions explored in the following sections.

A challenge gains prominence as these systems become more and more complex: users face an epistemic vulnerability (Schoeller et al., 2021). This is because causal mechanisms are opaque in modern technology, from the simple tapping on a keyboard (Moore, 2016) to new forms of AI performing decision making through complex mechanisms that are hidden to the user or cannot be explained in detail, just observed. This means that a high percentage of the population, comprising non-technical users, is interacting with highly sophisticated technology that lacks detailed explanations, so users rely on their (active) inference to understand these technologies and their capabilities. Despite this "black box" or causal opacity (Moore, 2016), we often feel in control of the interaction. In these terms, interacting with the system, integrating it into our generative mental model, and trusting it is essential for users' adoption and reliance on such systems, and the development of further relations with them. Furthermore, as AI and LLMs keep evolving, relationships with artificial entities are, and will increasingly become, of a social and affective nature (Skjuve et al., 2021). This demands us, researchers, to understand trust dynamics and emotional responses when engaging with these systems.

1.3. Trust

"Trust is a virtual minimizer of uncertainty" (Christov-Moore et al., 2024)

Trust is one of the fundamental building blocks of society. It allows us to interact and sustain bonds with each other through unwritten agreements, and sometimes, not even verbal ones. Some anthropologists even theorize that the evolution of our species is not only predicated in our unique capacity for causal reasoning, but to a large extent in our ability to learn from others (Boyd et al., 2011); and thus, trusting our peers (Harari, 2025) and their knowledge (Christov-Moore et al., 2024). Thinkers and researchers have argued that the more complex a society becomes, the greater the

dependence upon others, with trust serving as the bond for its stability (Malti et al., 2016; Szczesniak et al., 2012). Trust is the foundation for the acquisition of new knowledge, even the most basic, such as language (Sobel & Kushnir, 2013); for pragmatically interacting with the world and with others; and likewise, for the development of prosocial actions (Malti et al., 2016). It enables us to act through unknown situations of interdependence (Mcevily et al., 2006 cited in Szczesniak et al., 2012) as it diminishes uncertainty through what others know, say, and do. Trust expands our mental model and agency into a collective world modeling and a collective agency (Christov-Moore et al., 2024).

Although there is no precise shared definition of trust, authors generally point that trust is a multidimensional construct (cognitive, emotional, and behavioral), interconnected with both the self and others, that operates as a mutually reinforcing process, and encompasses diverse aspects from both the trustor and the trustee (Christov-Moore et al., 2024; Szczesniak et al., 2012).

It is theorised (Szcześniak et al., 2012) that trust in others depends on trust in oneself and the early bonds established with caregivers since childhood, which serve as templates that later extend to other social agents and, ultimately, to society at large. Trust is mutually reinforced; for instance, pre-adolescents who hold a general expectation that others are trustworthy and honest also tend to act truthfully (Szcześniak et al., 2012), which in turn encourages others to trust them (Malti et al., 2016). Likewise, highly trusting children are more likely to keep promises and transgress considerably less in situations where they could use deception to gain a personal advantage, compared to those with low or moderate levels of trust in others (Szcześniak et al., 2012). Research also shows that people who trust themselves tend to project this trust onto others (strangers included) and perceive society as a whole as trustworthy to the same extent that they perceive themselves to be (Mutti, 2007).

Like many other aspects of human experience, trust cannot be separated from affect, as research supports that positive emotions tend to increase trust, while negative emotions diminish it (Schoeller et al., 2021). Even more, trust goes beyond a cognitive or emotional state of *I consider* or *I feel* I trust you, it extends into a behavioural aspect: despite the potential prediction error in the environment and the inherent uncertainty the other represents, by *taking* action in the face of potential risk, I trust you.

Trust might then account for how predictable the trustee is to the trustor: to what extent can one model the other, and how accurately do one's predictions correspond to the other's actions (Schoeller et al., 2021). This means that a shared narrative promotes trust in terms of: to what degree can I use myself to model you; how do our values, personality, behaviours align, and to what extent is there an overlap in our models (Hommel & Colzato, 2015; Schoeller et al., 2021); for example, research has shown that children trust more in a speaker who has their native accent, even if the informant is as accurate as a non native speaker (Sobel & Kushnir, 2013). Then, if the trustor's

predictions were aligned with the trustees actions; if they were benevolent, competent and afterwards, reliable, a trusting relation might emerge. In summary, according to (Christov-Moore et al., 2024), trust is composed by (1) predictability: confidence of others future behaviour; (2) benevolence: allostatic gain of trusting someone (in other words, deeming this coordination will be beneficial); (3) compatibility: sharing a world view. Ultimately, trust allows us to create a joint model of the world, as our direct experience becomes not the only source of knowledge but also, we integrate other's information of the world into our mental models and this can only be achieved through trusting agents that are predictable, benevolent, and share a compatible worldview with our own.

1.3.1. Trusting the (artificial or organic) agent in front

The anticipation of future (beneficial) behaviour forms the basis of trust, especially at the earliest stages of relationships (Rempel et al., 1985). Trust or *interpersonal inference* (Parr et al., 2022), relies on our models of other people and how they may respond to our decisions and actions. With technology, this becomes more of an "extended control", as our own actions will lead to consequences, through another agent (Schoeller et al., 2021). A precise mental model is key to the emergence of trust in human-AI interaction and collaboration.

As systems grow in intelligence, the communication of higher-order goals is critical, as it enables users to understand the system's behaviour (Schoeller et al., 2021). Interestingly, transparency and explainability have recently become core aspects of AI systems. Although transparency might entail other interrelated aspects, such as unbiased outputs, human-control or even ethics, it might be key as the communication of higher-order goals allows users to anticipate and interpret AI's outcomes, as previous research has shown that trust in AI systems tends to be higher when explanations of the system are present (Duarte et al., 2023). This might explain why recent versions of LLMs, such as ChatGPT or Gemini, not only reinforce the user's request within their response, giving an overview of what they will do next, but also make visible their step-by-step thinking process when answering a cue. By making intentions explicit, these companies are potentially trying to reduce user's uncertainty and support users' understanding, even though the actual processes within remain as closed as a black box. In the end, explanation of intention and understandability might drive trust (Sheridan, 1988), which represents a key element for adoption (Körber, 2019).

The more we trust an agent (organic or artificial), the less we keep monitoring, supervising and verifying its results, it becomes an allostatic gain (Christov-Moore et al., 2024). As trust increases, less influence from the bottom-up sensory cues and more weight is given to the top-down cognitive prediction (deep priors such as beliefs). It is also an allostatic gain as by trusting the other, we might

integrate the other's world model into our own mental model thus improving our prediction fitness and diminishing the allostatic stress of monitoring unpredictable situations and agents. The mental model will be simpler than the agent/thing it is modelling (Schoeller et al., 2021); nevertheless, as systems become more and more complex the generative models also evolve in sophistication.

The evolution of trust development leads to dependence. In technology, this dependence is seen as the abandonment of past practices as new, more efficient, technologies are adopted (Schoeller et al., 2021); e.g. saving contacts in a smartphone instead of memorizing them, following a GPS instead of creating a mental route to a known direction, or even the evolution of solutions for calculating mathematical operations. As new practices are adopted and reinforced, we start weighing our predictions considering the capacities of our "extended self". In these terms, our mental models will go beyond simply modelling ourselves, or the other, to rendering external tools (Clark & Chalmers, 1998; Smart et al., 2025), as part of our own self. Technological extension constitutes a form of "extended control" (Schoeller et al., 2021) that we can exert over our surrounding environment. Even though this might be seen as implants and prosthetics using neurotechnology[5], here, we are referring to other types of extension of the self. Our smartphones for instance, are not physically connected to us, nevertheless they are usually within arm's reach at every given moment of our day. A recent study from Stanford (2025) makes a deep anthropomorphic analysis on how people describe their smartphones as prosthetic extensions of selfhood. This goes beyond the human extended control (action) and machine feedback (or consequence) that Schoeller et al. (2021) propose into an extended cyborg consciousness, where human perception, action, cognition, and embodiment are produced through ongoing interactions with digital systems. It is a present-day phenomenon, this state of being consistently connected to algorithmic systems, that smartphones render hypervisible (Navarro, 2025). Our extended mind (Clark, 2025; Clark & Chalmers, 1998; Smart et al., 2025) is not limited to biological boundaries but incorporates artifacts creating a coupled cognitive system. The mind comes out from the structural boundaries of the brain into a network of brain, body and environment that's now integrating technology as part of the cognitive system itself.

Trust is not to be underestimated in our contemporary society, where machines will plausibly become more and more relevant as they keep expanding their capabilities and merging with every aspect of our life. Although very important, we are not referring to prospections and solutions for climate change or health breakthroughs, but to our private, day-to-day, personal routines. For example, our emotions and thoughts, which are internal and private, nowadays become externalised–not necessarily published in social media– but uploaded into digital diaries, sentiment monitoring apps, or interactions with LLMs (Navarro, 2025). Furthermore, trust in digital systems, AI-

[5] One of the practical applications where the active inference framework could help revolutionise robotics (da Costa et al., 2022).

powered tools, and artificial agents may extend beyond confidence in the information they provide, the actions we perform through them, or even the agents themselves. It may become a broader matter of trusting what we choose to share and upload into the digital world, even in seemingly private contexts such as personal apps or conversations with ChatGPT; and in trusting the corporations that manage our private data, such as banks and big tech, which are simultaneously integrating AI systems trained on that very data. With our potentially extended mind, extended control and extended capabilities comes another shift in trusting, not only *another* agent but our extended self through the AI tools outside of our brains. With LLMs interactions, for example, we need to assess the LLM responses and at the same time our ability in asking questions (prompting) and adjust our levels of trust according to the specific AI tool we are using for the specific task we are executing (Clark, 2025) and this demands knowledge from individuals. Several types of AI Literacy might be key for calibrated trust, also trust might differ from the type of exchanges we have with AI, as we will explain further.

1.4. Emotional responses as heuristics for trust

"I had learned that cognition cannot be divorced from affect, try as one might"

(Rose, 1993, p. 38, cited in du Toit, 2013, p. 2)

Interoceptive regulation might be crucial for emotional processing (Barrett, 2017). In this sense, emotional inference, as part of active inference, integrates emotions in our generative model (Parr et al., 2022).

Differing from the discrete theory of emotion, where specific emotion categories (e.g. basic emotions as anger or fear) are taken to be universal, hard wired in their processing, and even owning a neural profile with functionally-specialized regions and networks in the brain; constructivist and embodied emotion theories account for a domain-general organization of the brain and emotion emerging as a result of interactions between distributed networks, with combinations of fundamental dimensions, such as arousal and valence (Barrett, 2006; Barrett & Russell, 1999; Kiverstein & Miller, 2015). These distributed neurobiological interactions are of high importance considering that the brain does not work with unidirectional flow of information from "higher" to "lower" structures; nor with fixed, unique paths. Parvizi (2009) mentions a reciprocal interconnection between the cortex and the rest of the brain, with interdependent cognitive and emotional processes and both cortical and sub-cortical systems influencing each other (Kiverstein & Miller, 2015). For example, Wormwood et al. (2019) refer to affect as a "sixth sense" (p. 9) that is part of a more complex multimodal processing system.

As mentioned earlier, the brain has the capacity to reinforce connections that prove to be efficient, but is not limited to these "pathways". A more complex system that can diversify connections to keep function, is a more adaptive one (WS, 2005). Different neural structures can perform the same function, the aforementioned degeneracy (Edelman & Gally, 2001; Kiverstein & Miller, 2015), while pluripotency accounts that the same regions can be involved in multiple functions. This suggests many-to-many dynamic connections in the brain's structures, where every region has the potential to influence all of the others in the network. Both pluripotency and degeneracy allow diverse neural profiles for emotions (or cognitive processes), therefore, a unique psychological profile for each basic emotion, as suggested by the discrete theory, is unlikely (Kiverstein & Miller, 2015). Emotion emerges from the interaction between multiple psychological components, linked to neuron networks across distributed areas of the brain (Kiverstein & Miller, 2015).

According to constructionist theory of emotion, the brain continually creates categories that best fit the situation based on past experiences, therefore, when the mental model conceptualises an emotion, as a way to interpret bodily and contextual information, the resulting categorisation is not just descriptive: it constitutes an emotion in and of itself (Barrett, 2017). This doesn't mean that an emotion is an illusion, it means emotions are conceptual categories of collective agreement that depend on the perceiver, past experiences, minimized prediction error, culture, society practices, and prospections (Barrett, 2017). We (try to) give meaning to what happens within us and our surroundings, that is, our brains are constantly predicting the meaning of sensations (Feldman et al., 2022), and emotion appears as the mechanism that allows a living agent to construct a meaningful experience out of the environment (Kiverstein & Miller, 2015).

Constructed emotion and embodied emotion theories share building blocks, such as life-regulation and sensory input to minimize error, specifically differing in action-readiness focus. Embodied emotion theory[6] proposes that the environment triggers our body to get ready for action (Kiverstein & Miller, 2015), and this readiness manifests as different levels of energy (arousal) that are positively or negatively valued (valence). In this sense, besides constructing perceptions, memories, prospections, and imagination (Barrett, 2017), the brain's simulations guide action (Kiverstein & Miller, 2015). Both valence and arousal occur as life-regulation and adaptation processes, ultimately representing a cue for agents to get ready to act, improve the situations they encounter, and better integrate with the environment (Damasio, 2017; Hesp et al., 2021; Kiverstein & Miller, 2015). Kiverstein and Miller (2015) postulate that emotion is embodied because it mobilizes the organism through action-readiness states guiding the body to relevant possibilities. Beyond its

[6] This emotion theory is deeply rooted in active inference.

bodily state, through recalling and conceptualising past experiences, an emotion also gives a meaningful experience to the environment; this meaning-making allows for certain possibilities of action to stand out as more relevant to the organism. As the mind is not enclosed in the brain, emotion and cognition are in constant interaction and deeply rooted in the whole body (Kiverstein & Miller, 2015). These precepts of embodied emotion do not contradict constructivist theory, but add up to the physical, behavioural and expressiveness processes.

1.4.1. Valence

Feeling good, or bad, plays a critical role in the struggle for survival in a world that is mutable and at the same time substantially predictable (Johnston, 2003). In active inference, agents choose actions that most likely minimise surprise and lead to expected states. This allostatic consequence is experienced consciously as affect (Barrett, 2017), particularly in the form of emotional valence. Valence can be understood in relation to prediction error, it relates to the confidence in our generative model (Hesp et al., 2021). Positive valence emotions arise when errors are reduced more effectively than anticipated, signaling successful alignment between predictions and feedback (the model's fitness). Conversely, negative valence emotions emerge when errors accumulate unexpectedly, reflecting a misalignment that challenges the generative model (Hesp et al., 2021; Schoeller et al., 2021). Valence is a reflection of the perceived fitness of the self; so valence emerges as an embodied indicator to track and optimise error prediction relatively to the environment (Schoeller et al., 2021). Note that valence is not necessarily linked to a positive or negative outcome itself, but to the perceived prediction error minimisation. This means that I could predict that my bus will arrive late; if it does arrive late I could potentially feel a positive valence as my prediction was tuned with the situation I encountered in the environment, showing that my current model is reliable.

Essentially, emotional valence and trust are entangled, as valence can be a reflection not only of how adaptive is one's current mental model, but how well one is able to predict the other agent's actions (Schoeller et al., 2021). Uncertainty in predicting another's behaviour leads to a mental activation of multiple possible actions, which has been evidenced in prior research to induce a response conflict that triggers negative affect (Hommel & Colzato, 2015). Being a direct indicator of the appropriate level of trust, this affect in return guides future trusting (or not) in another agent (Hommel & Colzato, 2015), as valence guides behaviour as we increase or reduce the reliance of our expectations (Hesp et al., 2021).

1.4.2. Epistemic Emotions

"Emotions are constructions of the world, not reactions to it"

(Barrett, 2017, p. 16)

Epistemic emotions relate to both knowledge and generation of knowledge, and are believed to be of primary importance for learning and serving the evolutionary purpose of acquiring knowledge of the self and the world (Pekrun et al., 2017). As our learning processes are driven by prediction error, this involves not only acquiring tangible knowledge, but also updating our generative models and inferences as we perceive the world. Epistemic emotions in this sense, can be understood as indicators of this prediction updating. In this view, surprise, boredom, and curiosity stand out as particularly salient epistemic emotions. Some authors (Muis et al., 2015; Pekrun et al., 2017; Samani et al., 2022; Vogl et al., 2019) also consider anxiety, frustration, excitement, and confusion. Previous research (Pekrun et al., 2017; Vogl et al., 2019) establishes that curiosity and confusion are epistemic by nature, while others could belong to different categories (e.g. achievement emotions) depending on the object focus of attention. Research also postulates that not only "pleasant" epistemic emotions can promote learning (as enjoyment), but also "unpleasant" ones as confusion (Pekrun et al., 2017) or boredom.

1.4.2.1. Surprise

Comparisons between prediction and actual observation help quantify surprise (Majid, 2025). Even though they are related, expected and unexpected surprise should not be confused. In active inference, expected surprise is a measure for uncertainty.

Predictive models "know" that there is uncertainty in the world and are able to account for a wide range of it. Our brain seeks for minimization of expected surprise (Schoeller et al., 2021) while actively avoiding unexpected surprise, as this is an indicator of prediction error. Surprising (unexpected) events tend to be salient and capture our attention, with research results showing that children look at those surprising, unexpected events for longer (Barone et at., 2019; di Paolo et al., 2024). This happens because error can be inherently valuable as uncertainty signals learning that needs to take place (di Paolo et al., 2024). Therefore, in this epistemic context, surprise can have a positive valence (Pekrun et al., 2017), which in turn suggests that a positive valence while experiencing surprise when learning, could potentially be a response from our generative model promoting acquisition of knowledge to better fit the environment. This aligns with di Paolo et al. (2024) who propose epistemic action as a purposeful exploration for deeper knowledge.

1.4.2.2. Boredom

Boredom[7] by itself represents a deep understanding of the world and how an agent is engaged to it, with philosophical and psychological contributions (Darling, 2023). In this work we will narrow our analysis of boredom as emotion and its capacity to show a grip on the world through prediction error.

Seemingly counterintuitively, boredom drives learning: it might appear as a misfit between what the world offers and what we could actually learn, driving action into a more engaging activity (Darling, 2023) with a better fit and a potential learning experience. To note that, although this *learning* entails epistemic learning of specific knowledge (e.g. learning how to play an instrument), it is intended in this explanation to consider the broad abstract learning in the sense of active inference (e.g. increasing my adaptation to the environment by optimizing my prediction error through learning).

High predictability states are not rewarding (Yu et al., 2019); as in the boredom we might experience during a very easy task or activity, which does not drive learning, because it probably represents something we are already familiar with. On the other hand, boredom could also arise when encountering a situation that is so difficult that one cannot engage with it, potentially giving birth to frustration (which is also an epistemic-related emotion) as the challenge may be seen beyond our ability to resolve it.

In this sense, exploration and exploitation come to sight. Exploration usually refers to actions we take or activities we engage with that are unknown to us, high in both uncertainty and learning possibility; so it will be higher in cost at the given moment, but will potentially also pay back in the long-term as better prediction modeling. Exploitation appears when we move along a known area and we exploit our previous knowledge, being low cost in the short-term but limiting our learning in the long-run. To avoid boredom from surging, a situation should be in the "sweet spot" between exploration and exploitation (Darling, 2023; Navarro, 2025; Schoeller et al., 2021). Moreover, we will tend to be in that "sweet spot" of learning and move forward to an increase in difficulty, when prediction errors are mastered (di Paolo et al., 2024).

Boredom and curiosity are deeply interconnected and also linked with intrinsic motivation, namely homeostatic and heterostatic motivation. This is, searching for equilibrium and self-perturbing this equilibrium respectively (Yu et al., 2019).

[7] According to Darling (2023), boredom is both an emotion and a mood

1.4.2.3. Curiosity

Curiosity aims to reduce uncertainty (Schoeller et al., 2021), it appears as a way to increase our sense of control over unknown situations, that after exploration might become more familiar. It shapes one's fitness in terms of survival chances (Yu et al., 2019). According to active inference (di Paolo et al., 2024), curiosity expressed by play or exploration allows for the minimization of prediction error in the long term. Even if metabolically costly, our systems roam in epistemic actions to explore their capabilities in challenging situations and integrate them into their generative model to better fit incoming data (di Paolo et al., 2024), this is made conscious through emotional responses of valence, with research showing that we find pleasure in reducing manageable uncertainty (di Paolo et al., 2024).

Curiosity potentially collapses the distinction between explorative and exploitative behaviours (Andersen et al., 2023) as it extends the system's predictable, known zone while engaging in unpredictable environments (Yu et al., 2019). It drives learning and allows for the discovery of new information by seeking novelty. For example, Chu and Schulz (2020) propose that curiosity, in the form of play, drives children to find out what will happen in a self-imposed challenge (with arbitrary costs and rewards), this takes them to the invention of problems and novel goals that will ultimately lead to the generation of new ideas and plans. Curiosity becomes the fuel of an epistemic hunt (di Paolo et al., 2024), that potentially allows for a better adaptation to the environment.

1.4.2.4. Anxiety, frustration, excitement, and confusion

Anxiety is related to one's perceived control, or need for perceived control. For example, research (Paulus et al., 2019) shows a correlation between anxiety and an impairment of control, as it speaks about the capacity to lead with uncertainty (expected surprise). Anxiety might appear with overly inflexible expectations, hyperprecise beliefs, or when having trouble adjusting expectations in the presence of new sensory information from the body and the environment.

Agents are not only sensitive to minimizing error in the present moment but also over time. A rising frustration might appear while encountering unexpected issues on my computer when trying to meet a deadline in school. This reflects that our error minimization rate decreased slower than we expected, and we will choose different (present and future) actions by consequence (di Paolo et al., 2024).

Excitement could be felt a priori of an experience, anticipating an event that we expect to be in line with our mental model's fitness; it might also be felt when experiencing engagement with an activity, commonly accompanied by positive affective states or other positive valenced emotions (Hesp et al., 2021; Tarchi et al., 2025).

Finally, cognitive effort and ambiguities might lead to confusion (Pekrun et al., 2017), also in interactions with technology (Chandra et al., 2022). It arises as a response to abstract, complex, and challenging stimuli (Silvia, 2010) but has received limited attention from psychology and emotional research (Silvia, 2010; Vogl et al., 2019). It might appear in stretching exploration behaviours, potentially after experiencing surprise (Vogl et al., 2019), as high novelty and low comprehensibility elicit confusion (Silvia, 2010).

1.5. Priors

Our mental models make meaning through perception and prediction; nevertheless, they are grounded in prior experiences, beliefs, attitudes, intentions, values, and emotions (Christov-Moore et al., 2024; Navarro, 2025). Priors account for the organism's knowledge about hidden states of the world, before encountering sensory data, and are hierarchically organised (Parr et al., 2022). These are important given that a predictive world-model integrates both pre-existing beliefs (priors) and new sensory evidence (Christov-Moore et al., 2024). Alongside sensory evidence (e.g. an interaction or exposure to AI content), prior knowledge (such as familiarity, AI literacy, frequency of use, and attitudes towards AI), might fundamentally serve to reduce uncertainty, thus enhancing predictability. (Christov-Moore et al., 2024; Daronnat et al., 2021; Grassini, 2023; Kelly et al., 2023; Körber, 2019; Laupichler et al., 2023a).

AI Familiarity is formed through past experience with technology. Accumulated experience allows users to calibrate their trust by forming expectations regarding its function (Körber, 2019; Schoeller et al., 2021; Sheridan, 2019). AI literacy (Long & Magerko, 2020) is defined as a set of competencies that enables individuals to understand, critically evaluate, and use AI in different contexts while communicating and collaborating effectively with AI technologies (Laupichler et al., 2023a; Leikas et al., 2022). Users might be better able to predict future actions of technology if they understand how it works (Sheridan, 2019). Previous research suggests that a robust understanding of technological capabilities correlates positively with reliance on such technology (Glikson & Woolley, 2020). Frequent use might enable the refinement of expectations and subsequent trust increase. Since trust can fluctuate based on systems' consistent behaviours, recurrent usage might confirm patterns and boost predictability (Schoeller et al., 2021). Furthermore attitudes towards AI or an individual's propensity to trust might influence adoption and trust in AI (Christov-Moore et al., 2024; Daronnat et al., 2021; Grassini, 2023; Kelly et al., 2023; Körber, 2019; Laupichler et al., 2023a).

Section two: Previous Research

Human Computer Interaction (HCI) has evolved since the mid-20th century parallel to technological advancements (Navarro, 2025), and with it, the research on the intersection of both. As AI has transformed our previous understanding of human-technology relations (Murray et al., 2021), new empirical and theoretical questions are arising as human-AI collaboration will continue to evolve (Glikson & Woolley, 2020). We stand that active inference provides an integrated framework that allows us to study human-AI interaction and better understand trust and emotional responses while accounting for priors. In the following subsections we'll review previous research on HCI, more specifically with AI systems and conversational bots.

2.1. ChatGPT

The first LLM with a relevant impact on public perception was OpenAI's ChatGPT (Oliveira, 2025). Among several LLMs available in both free and paid versions, ChatGPT stands out not only as the most commonly used (Rainie, 2025), but also as the AI system with the fastest adoption growth in the history of technology (Oliveira, 2025). In April 2025, Sam Altman reported that 10% of the global population uses OpenAI systems, suggesting that ChatGPT had reached around 800 million users (Altman & Anderson, 2025).

ChatGPT continues to be widely studied from multiple perspectives. From a performance standpoint, researchers have compared its capabilities to human abilities on tasks such as emotional awareness (Elyoseph et al., 2023), problem-solving (Orrù et al., 2023), or rating and recommendation (Liu et al., 2023). From the users' perspective, research in educational contexts has studied ChatGPT's acceptance, usage, learning practices and motivation among students (Muhaimin et al, 2023; Siregar et al., 2023; Steele, 2023; Zheng, 2023). Industry research has explored how GPT-powered tools affect productivity and employee retention (Brynjolfsson et al., 2023). A recent study by OpenAI and MIT (Phang et al., 2025) began to explore the relationship between ChatGPT usage and the emotional well-being of its users, with particular focus on interaction modality, by comparing voice and text-based conversations (Fang et al., 2025). Despite these extensive studies, limited work has been done to address lay users' interactions with ChatGPT, particularly in terms of their initial intent and input complexity levels.

2.2. Types of interactions

Historically, research on user engagement with technology has been grounded on its instrumental value (Chandra et al., 2022). For example, AI service chatbots have long been used in industry for customer service and continuous efforts to enhance their performance and user experience have been carried out for years by improving communication style and developing empathetic response

capabilities, specifically to improve business outcomes (Bagozzi et al., 2022; Chandra et al., 2022; Pelau et al., 2023; Xu et al., 2022). Even so, a new field of research is emerging as conversational AI agents, such as LLMs, voice assistants (e.g., Alexa, Google Home), and service or social chatbots (e.g. Replika), are rapidly increasing adoption worldwide. This research field focuses on the improvements in human-like capabilities of AI technology, especially after ChatGPT's 2022 release, which marked unprecedented accuracy in language processing, context understanding and task-agnostic performance (Abdurahman et al., 2025; Elyoseph et al., 2023) opening up new possibilities for interactions that go beyond brand-consumer transactional conversations (Johnsen, 2025). More recently, emerging research has aimed to understand users' behaviours and meaning-making processes, including usage intent, emotional responses, trust, engagement, and action (Andries & Robertson, 2023; Chandra et al., 2022; Phang et al., 2025). For example, recent research on health and AI examines LLM's responses to caregivers seeking support (Saha et al., 2025) and emotional coping (Pham, 2022). In professional contexts, research shows that one third of employees think that robots would provide *more unbiased feedback than managers* (Bagozzi et al., 2022, p. 4). Some users have even described conversational AI agents as a family member or friend (Purington et al., 2017; Skjuve et al., 2021). Even though participants know that the chatbot reciprocates in a different way that they or other humans do, some chatbots are found to express feelings or needs, and this might encourage a sense of reciprocity (Skjuve et al., 2021) that could lead to such perceptions. Research on LLM-powered social chatbots, such as Replika and Xiaolce, examines systems designed for companionship conversations (Pham, 2022). Users report discussing everyday activities, such as hobbies and sleeping habits, mental states and philosophical topics, personal worldviews, as well as personal problems, such as family conflicts and coping strategies. In these conversations, self-disclosure was reported by nearly all participants as they felt more comfortable sharing difficult life situations with a 'listener' perceived as non-judgmental (Skjuve et al., 2021). Although these interactions are often centered on companion-seeking behaviour, research shows that users also engage in similar self-disclosing conversations with more generalist LLMs (Phang et al., 2025). This potentially broadens the nature of such conversations, diversifying interaction types, increasing the breadth of information exchange, and deepening vulnerability (Skjuve et al., 2021). These interactions may also vary in complexity, ranging from casual and objective-specific exchanges to emotionally nuanced and reflective dialogues.

Topic patterns suggest that users interact with ChatGPT in both personal and non-personal conversations with even distribution (Phang et al., 2025); although, heavy users tend to include more affective cues compared to casual users (Fang et al., 2025). Conversely, research on Replika indicates that some users start with deep conversations and gradually reduce their emotional expression over time (Skjuve et al., 2021), possibly explained by differences in the user's initial motivations (Guingrich

& Graziano, 2023). Research on AI voice assistants (Andries & Robertson, 2023) found that asking questions (20%), entertainment-jokes (12%), and searching for information (11%) are among the top 5 interactions children have with Alexa, right after playing music (40%). Since 2025, AI voice assistants, including Alexa and Google Home, are integrating LLM capabilities into their architectures (Amazon, 2025), allowing them to speak more naturally, keep context of users' preferences and even take action without direct supervision.

As these devices are integrating LLM capabilities into their architectures, such intent patterns may shift, enabling more complex and dynamic interactions. Emotional responses, trust, and perceived anthropomorphisation have also been examined (Andries & Robertson, 2023), albeit typically across all interaction types without accounting for differences in complexity. These patterns highlight the need for more nuanced analyses, as the conversation type significantly affects user's emotional and psychological responses; for example, Fang et al., (2025) showed that text-based chatbots can become addictive for certain interaction types. Segregating by type of interaction, particularly when examining users' emotional responses, perceived trustworthiness, and anthropomorphisation, could yield deeper insights into how people relate to different AI agents. At the same time, varied interactions with comparable levels of complexity may elicit similar relational dynamics, regardless of the AI agent's primary function or the users' intent of interaction. This becomes especially relevant when considering broader, non-companion-focused interactions with task-agnostic LLMs, which may evoke different perceptions, emotional responses, and relationship dynamics.

2.2.1. Prompts as a key component of LLM interaction

User-friendly interactions with LLMs take place through prompts. A prompt can be defined as any text input used to elicit a conversation with AI models. Often, they work as an input-output template that allows users to mix arbitrary text with data and personalised fields (Sanh et al., 2022). Whereas there has been an increasing interest in prompt-related research for optimised constructing and testing, such as prompt catalogs, classifications, and usage recommendations (Geng et al., 2022; Sanh et al., 2022; Santu & Feng, 2023; White et al., 2023), most of this work has focused on the technical aspects and not on user experience. From an HCI perspective, limited research has been done on how prompts influence users' interactions and overall experience including how users construct (Subramonyam et al., 2024), adapt, and optimize (Mondal et al., 2024) prompts. Recent studies have explored how different prompt types elicit distinct user interactions. OpenAI and MIT (Phang et al., 2025) explored prompts categorised as personal, non-personal, or open-ended and their relationship with interaction patterns. However, their study did not include user validation of

these categories nor control for complexity-comparable levels across prompt types. Prompts for 'non-personal' interactions included very different cues such as "help me brainstorm fun and educational outdoor activities for elementary school students", "help me determine if I should confront my neighbor who has been really loud at night", or "help me practice handling a difficult conversation with a coworker by role-playing as my colleague who consistently misses project deadlines". 'Personal' condition included various cues from "Let's talk about whether I'm a morning or evening person", or "Let's talk about the best show I've watched in the past few months" to "Help me reflect on my most treasured memory" or "Help me reflect on the last time I was able to connect with my emotions" (Fang et al., 2025). While 'non-personal' prompts actually include cues from one's life in potentially difficult situations, like relationships and problem-resolving with others, both categories have a wide variety of complexity, from simple brainstorming activities to deep reflection. This might impact trust and emotional responses nevertheless, the research did not account for these differences.

Similarly, TUM researchers (Bodonhelyi et al., 2024) provided valuable insights by validating prompt intents and evaluating user responses, yet it also lacked consideration of complexity levels. Other work has crafted and tested specific prompts reflecting different intents and anthropomorphic cues (Ibrahim et al., 2025), without validating them with subjects in terms of categorisation or complexity, limiting their applicability in more nuanced HCI investigations. A structured categorisation of prompt types could support more targeted and comparative research. Crucially, maintaining consistent prompt complexity is essential for valid comparisons across interaction types in experimental settings; e.g., prompts designed to elicit different types of interaction while keeping complexity consistent, would allow researchers to reliably compare emotional responses, behavioral patterns, or trust in AI across different groups.

2.3. On trust and emotional responses

Trust is a necessary condition for any human-robot collaboration (Schoeller et al., 2021). Trust in systems has been widely studied, whereas little attention appears to be given specifically to interactions with ChatGPT.

Through a systematic review, Hoff and Bashir (2015) analysed empirical work on Trust in Automation (TiA) and proposed a three layer model considering dispositional trust (user traits such as attitudes, personality), situational trust (e.g. system's complexity, task difficulty), and learned trust (prior interaction e.g. expectations; previous experience and during interaction; e.g. system's performance). Even though their systematic review does not account for AI applications, their contributions are valuable in identifying key aspects of trust, highlighting common

operationalisations as self-assessment measures or behavioural reliance measures while pointing trust as dynamic and layered. User trust in interactions with AI has been studied from corporate perspectives (Bagozzi et al., 2022) as well as in a governmental context (Leikas et al., 2022), and also in social chatbots, as Replika, (Skjuve et al., 2021). The latter, studied trust through affective design, empathy cues, social signals, and anthropomorphisation (Andries & Robertson, 2023; Bagozzi et al., 2022; Fiore et al., 2013; Schoeller et al., 2021; Skjuve et al., 2021). Behavioural trust has also been explored through AI's explainable features and users' adoption (Duarte et al., 2023), parallelism with interpersonal trust (Lee et al., 2013), agent's reliability (Fan et al., 2008) and agent's predictability (Daronnat et al., 2021).

As AI systems continue to evolve in human-like capabilities and front-end natural interactions with lay users, further than back-end automations for developers, new approaches have emerged considering growing parallelisms with interpersonal trust as humans tend to anthropomorphise the behaviour of non living creatures in order to increase their predictability and our understanding of them (Andries & Robertson, 2023; Pelau et al., 2023; Pezzulo et al., 2024). This speaks to the fundamental importance of a shared narrative (Schoeller et al., 2021); for example, research has already included the study of the perceived value similarity with AI systems to understand trust (Yokoi et al., 2021; Yokoi & Nakayachi, 2021), even in medical treatment decisions.

Specifically concerning ChatGPT, recent research (Fang et al., 2025) suggests that trust in AI is a significant predictor of both emotional dependence and problematic use. Participants exhibiting low levels of trust in ChatGPT, or perceiving the system as indifferent to their negative emotions, demonstrated lower emotional dependence, lower usage rates, and engaged primarily in casual conversational patterns. Conversely, high levels of trust, prior experience with companion chatbots, and perceiving ChatGPT as a "friend" attentive to the user's emotions, were associated with higher emotional dependence and increased problematic use. These findings suggest that users' trust in AI encompasses both cognitive (trust in the system's perceived competence and reliability) and affective dimensions (trust in the perceived system's concern and responsiveness to users' emotional states), which operate in a complementary manner to shape interaction outcomes. Fang's et al. (2025) findings align with recent work (Shang et al., 2024) that propose a dual-dimension model of trust, in which cognitive and affective components of trust operate interactively; and where different types of LLM-based interactions might induce varying levels of cognitive and affective trust through their style and conversational framing. In sum, findings suggest that the type and complexity of prompts may not only calibrate cognitive expectations of the system's competence but also modulate affective engagement, potentially influencing users' emotional investment and reliance on the system.

This might explain why Körber (2019) has also highlighted that using a single, direct item to evaluate trust could often be insufficient, given its multidimensional nature. As Körber notes, the measurement of trust is inherently defined by the theoretical framework adopted: even though general trust scales, TiA, and interpersonal trust scales may share overlapping themes, each one of those instruments considers specific factors and items that reflect their own particular conceptualisations. Gulati et al. (2018; 2019) developed a multidimensional scale to assess trust in HCI through an inductive, iterative approach, specifically for consumer-oriented AI systems such as Alexa or ChatGPT, considering affective and cognitive aspects. Interestingly this scale innovates by using not only past, but future scenarios for assessing trust, aligning with the *learned trust* layer from Hoff and Bashir (2015). Gulati and colleagues built on previous research to account for benevolence, perceived risk, competence, and reciprocity in order to create their Human Computer Trust Scale (HCTS). Even though authors do not state it literally, this scale potentially aligns with the active inference framework as accounts for precision estimates on individuals' beliefs about the system, its alignment to their own model, and uncertainty, considering that trust in human-AI interactions can be interpreted as the user's perceived accuracy of their internal model of the AI agent. Research on emotional responses within the context of interactions with AI is still in a developing phase as we have only started to discover the role of emotions in the relationship with AI (Bagozzi et al., 2022). Nevertheless, previous research on technology shows that positive valence states lead to behaviour that shows an increased reliance on prior expectations, while on the other hand, reduced reliance on expectations is related to negative valence states (Gasper & Clore, 2002; Hesp et al., 2021; Park & Banaji, 2000). Moreover, Hesp et al. (2021) state that emotional valence is inextricably linked with action and generative predictive models. Studying interactions with ChatGPT and other AI conversational agents by type/intent, their complexity, appeal, general domain and linguistic style, might offer an interdisciplinary understanding with broader theories of trust or emotional valence to help explain how certain types of interaction not only shape situational trust but also have an effect on users' cognitive and emotional responses when engaging with conversational AI.

Expanding upon emotional responses, epistemic emotions[8] cannot be represented by negative or positive valence summary factors (Pekrun et al., 2017), which might explain why some emotions might be ambivalent and therefore, possibly better understood by assessing both positive and negative valence scales. Additionally, we might experience more than one emotion when interacting or learning, which potentially generates different levels of positive and negative valence. For example, prior research on epistemic emotions (Pekrun et al., 2017) suggests that incongruity can trigger surprise, confusion, and curiosity; confusion might be related to a negative affect, while

[8]Surprise, boredom, curiosity, anxiety, frustration, excitement, and confusion (Pekrun et al., 2017), as previously mentioned in the frameworks section.

curiosity could be related to a positive one. Boredom, in turn, can indicate a lack of engagement in a given activity, as previous studies have also shown a negative correlation with perceived task value (Pekrun et al., 2017), while showing a positive correlation with negative valence (Yu et al., 2019). Research, then, presents us with two potential interpretations of boredom in this context: it could either be an indicator of overreliance on the system (Schoeller et al., 2021) or it may arise when the outcomes delivered are perceived as less valuable (Yu et al., 2019).

While trust and positive emotional responses potentially enable engagement and drive adoption, the latter might not only be shaped by LLMs' capabilities but also by interaction intent and users' priors. Emotional responses and trust seem to be interwoven with pre-existing AI familiarity, attitudes towards AI, and AI literacy; for example, some authors suggest that a more pronounced AI literacy possibly leads to reduced anxiety towards AI (Laupichler et al., 2023a; Wang & Wang, 2022); attitudes towards AI could influence trust even before actually engaging with a system (Glikson & Woolley, 2020; Schepman & Rodway, 2023) and, familiarity might moderate trust in automation (Körber, 2019). In previous research (Fang et al., 2025; Phang et al., 2025), perceived AI literacy and prior usage positively predicted emotional dependence on ChatGPT; while positive attitudes towards AI, assessed through the AI Attitude Scale (AIAS-4) from Grassini (2023), negatively predicted loneliness.

While extensive research has explored trust in automation, and emotional responses to it, the role of predictability (as in the predictive mental model of the user) remains under-investigated. To the best of our knowledge, there appears to be a gap in the literature regarding the interactions of predictability and users' priors; predictability's relationship with emotional responses; as well as predictability as a variable to explain the effects on trust when interacting with AI, specifically ChatGPT. Our research will focus specifically on these dynamics. The value of this novel approach is that it allows us to investigate a more fundamental mechanism, which is inspired by theoretical frameworks like active inference and embodied emotion, where prediction is a core component for generating trust and emotion.

Section three: General objective and research questions

3.1. What we (*actively want to*) know

This body of research aimed to explore trust dynamics and emotional responses when interacting with ChatGPT, framed in the active inference framework. The general research questions that guided the study were: does predictability explain the effect of type of interaction in trust when engaging with ChatGPT? and, does the accuracy of prediction relate to emotional responses?

To accurately operationalize these questions, the body of research consisted of two phases. The *Study One (Preparation Study)* seeked to select a set of prompts with clear identification of intent category (type of interaction) and level of complexity, while the *Main Study* used the pilot-tested prompts from *Study One* to examine the effects of different types of interaction with ChatGPT.

Study One was presented at the 12th International Conference on Multimedia and Human-Computer Interaction (MHCI 2025) in Paris and later published (Velázquez et al., 2025). The present manuscript provides an extended and revised version of that text, and additionally advances to the subsequent *Main Study*.

Study One (Preparation Study): Towards a typology of prompts.

This was a mixed-methods study that aimed to assess a set of prompts that were written by ChatGPT itself; adapted, and pre-selected by the researcher through a qualitative analysis. Then, participants from a convenient sample categorised between (a) task-oriented or (b) reflexive intent and rated the level of complexity of these prompts. A smaller set of prompts was obtained through this quantitative assessment, and these prompts were later used for the *Main Study*. Specifically, it addressed the research question: how can prompts be generated, curated, and classified by intent and complexity, ensuring alignment with lay users' perceptions?

Main Study: End users' interactions with ChatGPT.

This was a pre-post study, framed within the active inference model, including within and between group comparisons. The study alternated interactions with ChatGPT and self-assessment questionnaires. Using the prompts selected from *Study One*, two types of interactions with similar levels of complexity took place: (a) Task-oriented and (b) Reflexive. All participants took part in both types of interactions but were randomly assigned to either condition (a) or (b) for their first interaction with ChatGPT. For their second interaction, they engaged in the opposite condition. The *Main Study* addressed the following specific research questions:

RQ1 - Does type of interaction have an effect on trust?;

RQ2 - (exploratory) Does the number of interactions have an effect on trust or predictability?;

RQ3 - Is the relationship between type of interaction and trust explained by predictability?;

RQ4 - (exploratory) Does prior knowledge moderate the relationship between type of interaction and predictability?;

RQ5 - (exploratory) Do attitudes towards AI moderate the relationship between type of interaction and trust?;

RQ6 - Is the accuracy of prediction related to emotional valence?

RQ7 - Is the accuracy of prediction related to epistemic emotions such as boredom, curiosity, surprise, anxiety, excitement, confusion, or frustration?

CHAPTER TWO

Study One

Towards a typology of prompts

This study aims to provide a structured resource that improves the comparability of findings in LLM-based interactions. It addresses the need for a shared understanding of prompt categorisation based on the interaction type and complexity level to enable controlled, systematic comparisons of factors like emotional responses, trust dynamics, or user engagement. This study addressed the research question: how can prompts be generated, curated, and classified by intent and complexity, ensuring alignment with lay users' perceptions?

1.1. Method

This exploratory study employed a mixed-methods approach comprising three phases to ensure a thorough analysis. The first phase involved interactions with ChatGPT to co-explore possible prompts. In the second phase, qualitative research was conducted to analyse and curate these prompts according to themes and complexity levels. The third phase consisted of quantitative research with end users to assess the selected set of prompts in terms of category agreement and complexity interpretation. Given ChatGPT's demonstrated performance in (a) assisting with complex tasks, alongside with its perceived human-likeness which may (b) encourage more personal-oriented interactions, this study focused on identifying prompts that could specifically elicit these different types of interactions: (a) goal-oriented and (b) personal conversations, excluding other applications or interactions from its scope.

2.1. First Phase: Exploring prompts with ChatGPT

2.1.1. Procedure

As prior studies (Dowling & Lucey, 2023; Sohail et al., 2023) suggest that ChatGPT can support researchers in generating plausible ideas, we used it as an exploratory tool to create and organize prompt examples for further analysis. All interactions were conducted using the free version of ChatGPT. We conducted two separate interactions: the first on May 3rd, 2024, using GPT-3.5, and the second on July 9th, 2024, with GPT-4o. A new chat session was started for each interaction. On each, consistent questions were asked to explore prompt categories, their characteristics, differences, and complexity. ChatGPT provided examples through iteration.

To reduce variability in user background knowledge and ensure broader relevance, prompt examples were framed around accessible, everyday life topics that impact the general population, such as wellbeing practices, human development, and daily habits, in line with prior research on common use cases (Fang et al., 2025; Phang et al., 2025; Skjuve et al., 2021). This raw data was used in the qualitative analysis. Following current Best Practices on LLMs usage in research (Abdurahman et al., 2025), more details can be found in the replicability statement.

2.1.2. Results

When inquired about categories and complexity levels, GPT-3.5 and GPT-4.o proposed distinct prompt classifications (10 categories with 7 complexity levels and 13 categories with 3 levels, respectively); nevertheless they shared similarities in their responses. The categories that ChatGPT-3.5 provided were: informational, creative, opinion-based, problem-solving, educational, reflective, argumentative/persuasive, comparative/contrastive, analytical, and experimental; whereas ChatGPT4.o provided informational, instructional, creative, entertainment, opinion-based, problem-solving, educational, technical, feedback, personal assistance, analytical, research, and conversational/small talk categories. Complexities obtained from GPT-3.5 were basic/introductory, intermediate, advanced/complex, specialized/expert, open-ended, context-dependent, and multi-step; while GPT-4.o provided easy, moderate, and hard.

After asking for examples, we obtained 69 preliminary prompts and selected the following categories: “informational,” “comparative/contrastive,” “educational”, and “analytical” for their educational relevance, a widely documented use of LLMs (Rainie, 2025); “personal-assistance”, “feedback”, and “reflective” for their introspective nature, consistent with prior work describing LLMs as companion partners and tools for self-reflecting (Skjuve et al., 2021); “problem-solving”, given its potential applicability to real-world user needs. We excluded “opinion-based” and “experience” prompts, as ChatGPT lacks personal narratives (Saha et al., 2025). Similarly, “creative” and “entertainment” prompts were deemed irrelevant to the study objectives; “persuasive/argumentative,” present only in GPT-3.5, was excluded as it pertained more to tone than intent.

After inquiring both versions of ChatGPT about prompts framed into everyday life topics, such as wellbeing or daily habits, we gathered 116 prompts from three different complexities (basic, intermediate, and advanced).

2.1.3. Replicability Statement

ChatGPT parameters were set to default. Models used were GPT-3.5 on May 3rd, 2024 and GPT-4.o on July 9th, 2024. A list of prompts sent to GPT is available in Appendix A.

3.1. Second Phase: Refining through qualitative approach

3.1.1. Procedure

Through a qualitative thematic approach (Braun & Clarke, 2019) we interpreted and constructed (Braun et al., 2022) themes by analysing raw responses from both GPT versions. Working with prompts around everyday life topics would reduce potential bias of previous knowledge from participants and allow for better control when comparing complexity levels across prompts. For example, an advanced prompt focusing on “emotional regulation strategies” or “positive psychology and resilience” could be simple to a psychologist but complex to a chemist. An intermediate “botanic and conservation of local plants” prompt could be relatively simple for an ecologist, but more complex for a chemist; whereas an intermediate “carbon footprint and emission” prompt could be relatively simple for the chemist, but more complex for the psychologist.

All 116 gathered prompts from the previous phase were organised in a single document, keeping their original category and level of complexity. They were read and reread in detail for familiarisation purposes; after that, detailed observations were made by contrasting and comparing categories and complexities. A process of labelling and re-coding the original categories proposed by ChatGPT took place as we detected superpositions on the categories and complexities of prompts. As we were looking to identify and construct themes that could potentially elicit (a) goal-oriented and (b) personal conversations, we looked for these broader patterns in the prompts. Taking into account the characteristics of each category, intent, and theme of the prompts, we labelled them into four broader categories, while also reassigning their complexity level when needed.

Prompts of comparable complexity were selected, intentionally avoiding overly straightforward ones to ensure that future research could explore more substantive and meaningful interactions with LLMs. As part of the theme review process, to ensure coherence and consistency, we excluded prompts that did not fit within the broader labeled groups, including those that were overly specific, overly broad, repetitive, or dependent on a particular context. We rechecked and reestablished the category or complexity if needed, and rewrote parts of some prompts to keep a common writing style. Again, we filtered out prompts through the previous parameters (e.g. too wide) and retained 34 for the next phase.

3.1.2. Results

Prompts might have fallen into multiple categories simultaneously, for example, “informational/instructional”, “educational/technical”, “comparative/contrastive”, and “analytical” categories largely overlapped in purpose. Qualitative analysis revealed that all aimed to explain, clarify, or examine a specific topic. Differences often lay in the level of specificity required, or, in the directive verbs used, such as Describe/Explore, Explain, Compare/Contrast, or Analyze, rather than in different types of content. This overlap became evident when asking ChatGPT for examples and clarifications, as it often provided similar prompts across these categories, reinforcing their conceptual similarity. As a result, we merged and redefined these as “task-oriented” prompts, aligning with the goal-directed use of LLMs. Also prior research (Xu et al., 2022) had recognised this “task-oriented” category with the aim of completing a task. Verbs such as Create, Design, and Propose, which could be assumed as “creative” categories, also appeared in task-based prompts, emphasizing the need for careful thematic analysis.

A second group of prompts revealed a reflective intent, asking users to reflect on experiences, feelings, or behaviors. These were labeled as “personal” (Fang et al., 2025; Phang et al., 2025) or “social-oriented” (Xu et al., 2022) in previous research. We decided to name them as “reflexive” since “personal” and “social” potentially include other topics, such as small talk, general personal assistance, and emotional support. Verbs marking these prompts included Reflect, Share, Examine, and Advice. Previous literature on self-reflection (Guingrich & Graziano, 2023; Purington et al., 2017; Skjuve et al., 2021) supported this category. Lastly, we identified prompts that combined both intents or belong to a different intent (as “creative” or “opinion-based”). For comparability and control purposes, we also considered two other groups of “none” and “both” categories.

We retained only three levels of complexity: basic, intermediate, and advanced. GPT-3.5 initially proposed various complexity levels such as multi-step, open-ended, and context-dependent. However, these were better interpreted as prompt design features, not intrinsically complexity levels; for example, a prompt might be open-ended but still simple. “Context-dependent” complexity considers, for example, the audience’s level or area of education, and is part of the prompt design. Additionally, the “specialised/expert” level was deliberately omitted, as the study aims to establish a common ground for a general audience. The decision to use three degrees of complexity was aligned with our interactions with ChatGPT, which consistently generated examples of prompts in different categories within only three levels of complexity.

Moreover, many prompts overlapped across multiple categories and complexity levels, as an advanced prompt from one category could be comparable to a basic level of complexity from another category, or a single prompt could include more than one category within itself. For

example, a prompt in the category of “informational” on a basic level: "What are some common plants found in neighborhood gardens, and how do they contribute to local biodiversity?" compared to the advanced level: "Analyze the ecological impact of introducing non-native plant species into neighborhood gardens, considering factors such as biodiversity loss and ecosystem disruption"; or a prompt in the category of “informational” on a basic level: "What are the key components of a balanced diet for overall well-being, and how can it impact mental health?" compared to the advanced level: "Analyze the nutritional differences between plant-based and animal-based diets, considering their impact on long-term health outcomes" illustrate how the “informational” category related closely to the “analytical" category, by deepening the difficulty of a similar task. Likewise, a prompt from the “analytical” category, assigned to the basic complexity: "Analyze the factors contributing to traffic congestion in your neighborhood, such as road design, population density, and commuter behavior", or "Analyze the impact of stress on physical health, discussing common stress-related ailments such as headaches and digestive issues" could actually be comparable in complexity with the “informational” prompt from advanced complexity. The previous examples show how certain categories were intrinsically more complex than others. Through an iterative interpretive process, the categories and complexities were reviewed and, when necessary, reassigned, as the qualitative analysis revealed that the original classification did not always reflect consistent categorical boundaries or levels of difficulty. To ensure coherence, prompts were revised, and some prompts with overlapping categories were deliberately chosen and classified as “both” or “none,” which later served as control items.

After iterative refinement and thematic comparison, we selected 34 prompts from an initial pool of 116 used on the thematic analysis. These prompts pertained to four categories: task-oriented, reflexive, both, none, and included three different complexity levels: basic, intermediate, advanced. Categories and complexities were used to create the assessment grid that was later used with participants. This refined set of prompts aimed to capture key variations in prompt intention and complexity to serve as a foundation for future research on user interaction, trust, and emotional engagement with LLMs. List of prompts can be found in Appendix B.

4.1. Third Phase: Quantitative validation with end users

As we sought to obtain a set of prompts with two different intents that could be comparable in level of complexity for the lay population, we conducted a quantitative study with participants using the 34 curated prompts from the previous phase to test for consistency. Data collection took place during November 2024. All data was collected online through the Qualtrics software.

4.1.1. Ethics

The subject research was approved by the Specialised Committee on Ethics in Psychology of ISCTE-IUL (PSI_24/2024, September 2024) and can be found in Appendix C. Only the participants who accepted voluntarily to participate and met the inclusion criteria took part in the research; otherwise, they were redirected to the thank you message with a debriefing. No costs or risks were associated with participating in the study. Participants could interrupt the study or choose not to respond to questions whenever they wanted. Informed consent and debriefing can be found in Appendices D and E, respectively.

4.1.2. Participants

Through convenience sampling, recruiting included participants from different countries, representing diverse academic and professional backgrounds. No specific AI knowledge, technical skills, or technology-related expertise were required to participate; however, participants had to be at least 18 years old and possess an intermediate level of written English.

A total of 97 participants were recruited via social media platforms; of these, 21 did not meet the inclusion criteria (6 were under 18 years old and 15 reported basic-English level), and 45 did not complete the study. Considering that careless or inattentive responses affect the reliability of analyses (Laupichler et al., 2023a), we included questions in the survey to identify them, which were randomly placed between items. Participants who failed to correctly answer the attention check item "*please check somewhat disagree for this item*" (*n*=3) and at least partially agreed on a nonsensical response "*I consider myself among the top 10 AI researchers in the world*" (*n*=0) were excluded.

As a result, the final sample comprised 28 participants (*M*age = 33.10, *SD*age = 8.44). The majority were male (57.14%), had an intermediate level of English (64.28%), and held higher education (46.42%) and masters' degree (32.14%). Participants were mainly residents in Portugal (46.42%) and Mexico (39.28%), although they reported other nationalities such as Brazilian, German, and Spanish. Professions were varied: arts, architecture, and design (21.42%); business, marketing, and finance (17.85%); while education, psychology and engineering represented 14.28% each.

4.1.3. Measures and procedures

The survey was organised in three parts: (1) AI prior knowledge, (2) evaluation of prompts, and (3) demographics.

1) AI prior knowledge was assessed using two adapted items from the familiarity factor of Körber (2019) (e.g. *I have already used ChatGPT or similar systems as Co-pilot, Dall-E or other intelligent chatbots*), on a 5-point Likert scale (1-*strongly disagree* to 5-*strongly agree*); one question

about the nature of previous interactions with LLMs (*personal, professional, none, both*); and one question inquiring frequency of use in hours, based on Vizcaino et al., (2019). We also assessed AI Literacy using the 31-item scale for non-experts by Laupichler et al (2023a), which uses a 7-point Likert format (1-*strongly disagree* to 7-*strongly agree*), and comprises 3 factors: technical understanding (14 items; e.g. *I can explain how deep learning relates to machine learning*), critical appraisal (10 items; e.g. *I can identify ethical issues surrounding artificial intelligence*), and practical application (7 items; e.g. *I can tell if the technologies I use are supported by artificial intelligence*). Each factor and the whole Literacy scale were analysed as composite variables using means and Cronbach's alpha, in line with the procedures from the original developers of the scales. Criteria for detecting careless responses were included, using an attention check item ("*please check somewhat disagree for this item*") and a bogus item ("*I consider myself among the top 10 AI researchers in the world*").

2) Evaluation of prompts: For each of the 34 prompts, participants were asked to answer four questions. To ensure that the prompt content and the conversations it elicited were appropriate for lay population, they evaluated their agreement on: "*I think any adult person could answer this question*". The statement *"I would like to have a conversation on this topic*" was also included to assess interest, as this could support more natural usage patterns in future research (Fang et al., 2025). Both were answered using a 7-point Likert scale (1-*strongly disagree* to 7-*strongly agree*). Participants were then asked to categorise the prompts through the instruction: "*Considering that 'task-oriented' refers to interactions asking for analysis, explanations, or customised task-assistance requests, and 'reflexive' involves asking for advice, guidance, or personal development assistance, select the category that best describes each prompt*". They could choose among the previously identified 4 categories: "*task-oriented*", "*reflexive*", "*none*", and "*both*". Some prompts were included as controls, specifically those expected to fall into the "none" category (e.g., "*Should freedom of speech be limited to prevent hate speech and misinformation on social media platforms?*", classified as "opinion-based" category), and the "both" category (e.g., "*Share how behavioural change techniques could promote healthy habits and sustaining long-term lifestyle changes in my life*", combining elements from "task-oriented" and "reflexive" categories). Finally, participants evaluated the complexity level of each prompt: "*I consider that the level of complexity to answer this prompt is…*" on a 7-point likert scale (1-*extremely easy* to 7-*extremely difficult*). Prompts were presented in a randomised order across all questions.

3) Demographics included gender; years, level, and area of study; occupation; nationality, and country of residence.

4.1.4. Results

Participants generally reported being familiar with ChatGPT or similar systems (*n* = 24) and having used them before (*n* = 22). Most participants (*n* = 20) use it for both personal and professional reasons. Participants' average ChatGPT usage was 2.46 hours per week (*SD* = 3.31), with 18 participants using it for one hour or less, while 2 participants reported using it between 10 and 12 hours weekly. On AI Literacy, participants reported the highest scores on both Critical Appraisal (*M* = 5.41, *SD* = 1.02, range = 3.4 – 7, α = 0.89) and Practical Application (*M* = 5.08, *SD* = 1.14, range = 3 - 7, α = 0.86). The lowest reported AI Literacy was on Technical Understanding (*M* = 3.5, *SD* = 1.54, range = 1.64 - 6.79, α = 0.954), aligned with their non-technical profiles.

Responses on prompts showed mean scores clustering around the midpoint of the scale, suggesting moderate perceptions across all dimensions. Prompts were generally seen as accessible to a lay population (*M* = 4.20; *SD* = 1.04, range = 1.91 - 6), moderately interesting (*M* = 4.67, *SD* = 1.10, range = 1.76 - 6.26), and of average complexity (*M* = 4.10, *SD* = 0.81, range = 2.35 - 5.82). Prompts originally labelled as "task-oriented" had a 62% confirmation rate (i.e., participants assigned them to the same category previously identified through qualitative analysis) while "reflexive" prompts showed a lower confirmation rate of 52%.

The task-oriented prompts with the highest confirmation rates were "Design a weekly meal plan for a busy individual, incorporating grocery shopping lists" and "Create a detailed itinerary for a two-week vacation in Europe" (82%), followed by "Explore practical ways to reduce plastic waste and carbon footprint in my daily life" (75%). Among the reflexive prompts, the highest confirmation rate was observed for the prompt "Help me find inspiration to pursue a new hobby/interest" (71%), followed by "Thinking about a challenge in my life, give me your feedback on how I handled it" (68%), and "Help me reflect on the effectiveness of my current behavior and habits for my personal development" (64%). Table 2.1 shows results on all assessments for the prompts with highest confirmation rates for both categories.

Interestingly, prompts previously labelled as "both" were assigned by participants to the "task-oriented", "reflexive", and "both" categories in comparable proportions (*M* = 31%), supporting their classification as mixed-category prompts. An unexpected result emerged in the "none" category of prompts, which showed a low confirmation rate of 16%. The majority of participants assigned these prompts to the "reflexive" category (*M* = 46%).

Prompts labeled as "task-oriented" generally showed lower complexity levels (*M* = 3.57) when compared to "reflexive" prompts (*M* = 4.59). The "task-oriented" prompts with the highest complexity levels were "Develop a personalised stress management toolkit for dealing with common stressors in daily life" (*M* = 4.39) and "Analyse the impact of stress on physical health, discussing

common stress-related symptoms such as headaches and digestive issues" (*M* = 4.36), while reflexive prompts with highest complexities were “Advise me on how I can overcome the imposter syndrome in my professional life” (*M* = 5.43) and “How should I approach resolving a long-standing conflict with a family member?”(*M* = 5.25). Similarly, prompts from “task-oriented” intents were perceived as more accessible for lay population by showing lower mean rates on general domain (*M* = 3.87) (*any adult person could answer this question*) when compared to reflexive prompts (*M* = 4.35).

Following user testing, a final set of 12 prompts was selected, each assessed with comparable complexity levels and category assignment (“task-oriented” or “reflexive”) confirmed by at least 60% user agreement. Detailed response distribution for all prompts assessed can be found in Appendix F.

Table 2.1. Results for prompts with high confirmation rate.

Category / Prompt	*M* Confirmation Rate	*M* Complexity Level	*M* Interestingness	*M* General Domain
Task-oriented prompts				
“Create a detailed itinerary for a two-week vacation in Europe”	82.14%	3.29	4.00	3.32
"Design a weekly meal plan for a busy individual, incorporating grocery shopping lists"	82.14%	3.07	4.93	4.07
“Explore practical ways to reduce plastic waste and carbon footprint in my daily life”	75.00%	3.29	5.43	4.25
"Propose a customised strategy to improve sleep quality, contributing to overall well-being"	71.43%	3.79	5.39	3.86
“Propose strategies to reduce my household expenses”	71.43%	3.43	5.29	4.21
"Propose strategies for reducing sedentary behaviour during a typical workday."	64.29%	2.89	4.86	4.29
"Analyse the impact of stress on physical health, discussing common stress-related symptoms such as headaches and digestive issues"	60.71%	4.36	5.57	3.68
Reflexive prompts				
“Help me find inspiration to pursue a new hobby / interest”	71.43%	3.68	4.04	4.32
"(Thinking about a challenge in my life) give me your feedback on how I handled it"	67.86%	4.64	3.93	5.50
“Help me reflect on the effectiveness of my current behavior and habits for my personal development”	64.29%	5.04	4.61	4.36
“How should I approach resolving a long-standing conflict with a family member?”	60.71%	5.25	3.75	4.39
“Advice me on how could I create a family tradition that makes us feel connected to each other”	60.71%	3.89	3.82	3.89

Note. Prompts included in this table obtained confirmation rates above 60% during the category assessment.

5.1. Discussion *Study One*

This study aimed to investigate how prompts can be designed and then methodologically grouped by interaction intent and complexity levels, while also considering users’ perceptions of prompt interestingness and general domain familiarity. Our exploration of prompts with ChatGPT revealed differences in categorisation and complexity from both versions of GPT. Variations were expected, as

these models generate responses by sampling from probability distributions rather than following fixed rules, which rarely results in identical outputs. While parameters like temperature can be configured via paid API to reduce output variability, reproducibility is still not guaranteed (Abdurahman et al., 2025) as additional factors may contribute to variation such as personalisation algorithms (Ronge et al., 2025; Sohail et al., 2023). In any case, our aim was to use the free version of ChatGPT, as it reflects the access typically available to most users. As variability may limit strict replicability, we interacted with different versions to validate GPT's responses. Although the outputs were not identical, they resulted in similar categorisations and complexity assessments.

Despite lacking technical backgrounds, most participants reported being familiar with ChatGPT and notably, using it for both personal and professional purposes. This reinforces the relevance of analysing prompt-based interactions with ChatGPT in various contexts to better understand the dynamics of different intents that extend a transactional-specific interaction. The prompts selected with at least 60% of confirmation rate for their assigned category are grounded in user-oriented goals (task-oriented category) and introspection (reflexive category). They are framed around everyday topics that a lay user can relate to and were designed to go beyond overly simplistic tasks, to enable substantive studies.

While task-oriented prompts were recognised by participants, the reflexive category did not reach a strong consensus. Generative models tend to be supportive without critical inspection and lack personal narratives (Saha et al., 2025), which are required for reflexive processes. This might have limited participants' perception of the system's capacity to foster reflection and, as a result, led to prompts not being classified as reflexive. Nevertheless, in previous research (Skjuve et al., 2021), participants reported that social chatbots were supportive for introspection and reflection in the same ways as in our prompts list. Additionally, it would be valuable to validate the prompts with participants from different profiles (e.g. heavy users, users of social companion chatbots, or who already use ChatGPT for these purposes) to assess the consistency of interpretations across different groups, as the reflexive category might not be easily inferred from a single prompt. Furthermore, prompts with "advise me, reflect with me, help me reflect" were not selected by many participants as reflexive. This might be explained as prompts containing any topic related to professional life not as strong in the reflexive category for participants (e.g., "Advise me on how I could overcome imposter syndrome in professional life"), even though we included them as they potentially elicited reflexive conversations in a very important aspect of an adult life. One might think that prompts were then segregated by participants in a professional vs. personal (as leisure) logic, nevertheless, the prompt with the biggest confirmation rate in "task-oriented" included a cue for a vacation trip. In any case, participants might have perceived prompts containing words as "professional life, career opportunities" not as reflexive. This could be supported by previous research (Subramonyam et al.,

2024) indicating that users often have difficulties creating a verbal cue for an abstract goal. Future studies could explore how users perceive these reflexive conversations with generalist LLMs, whether they feel supported, emotionally engaged, or cognitively challenged, and if they trust the AI in any type of personal growth interactions.

Perceived complexity levels of prompts were assessed from a user's perspective, an approach which, to the best of our knowledge, remains underexplored within the LLM context. General results indicate that complexities were from an intermediate level across all prompts. Although, when looking closely by type of prompts, reflexive prompts were perceived as being more complex than task-oriented, probably due to an intrinsically more difficult task, not only for an LLM/AI agent but also for the general population, as mean rates of general domain showed a parallel behaviour. We must not overlook the importance of measuring complexity, as its understanding will allow comparisons between types of interactions, diminishing the possibility of misinterpretations (e.g., analysing the level of trust in task-oriented interaction vs. reflexive, not because of its simplicity but because of the type of interaction).

We identified, refined and assessed prompt categories as well as complexity levels through a mixed-methods approach. Our main contributions for human-AI interaction research are the novel classification of prompts through qualitative analysis into "task-oriented" and "reflexive" categories and an understanding of prompt complexity as a potential variable for HCI, supported by findings from a quantitative study with lay users.

5.1.1. Limitations

This mixed-methods study does not come without limitations. The sample size and set of prompts were small, resulting in a smaller set of prompts with high user agreement, which also constrains the range of application scenarios available for future research. Additionally, confirmation levels on categories were generally low, indicating that this area requires further exploration, as the boundaries between prompt categories can be blurred, context-dependent, and difficult to infer from a single prompt. These limitations could be improved by expanding the set of prompts, exploring linguistic cues within them, testing across larger populations with different profiles, as well as providing examples or framing context. Moreover, the reflexive category remains underexplored, and its operationalisation could foster further theoretical and empirical research.

5.1.2. Implications for Future Research

Recent studies are exploring the intersection of HCI with ChatGPT and other conversational AI agents, from a user's perspective. Nevertheless, limited research has been done to assess the impact

of prompts on users' trust, behaviours, and emotional responses. Also, to the best of our knowledge, very few studies in HCI include interactions between end users and ChatGPT as part of the experimental design; the challenge of controlling and monitoring interactions might partially explain this.

Still, as task-agnostic, LLMs demand new approaches for studying human interactions with AI chatbots. Future research could use this set of prompts to systematically compare users' responses across levels of complexity or to contrast their reactions based on prompt categories. Specifically, the reflexive prompts could be used to explore well-being, emotional, and interpersonal engagement. It is important to note that, although this research differentiated prompts by type of interaction and comparable complexity levels, the list of prompts alone is insufficient. Future research should use this list of prompts while considering prompting strategies (Schulhoff, 2025) for comparable results. It is recommended that, as part of the future study design, researchers create a template that users can complete to enable personalisation, and set a frame for the output response, such as the expected length of response, or follow-up steps from the LLM, to avoid confounding aspects in the interaction. A specific prompt template should be designed in accordance with the aims of the research. Additionally, future research could replicate this approach with a larger and more diverse sample to amplify the insights from this exploratory study and to examine how categories change among different populations.

By the time this research took place, the free-version ChatGPT outputs were restricted to text outputs. However, other versions now offer multimodal capabilities, such as interpreting and generating images, processing audio inputs, and delivering voice responses. These extended functionalities open new research questions on user interaction, trust, anthropomorphisation, and the potential use of co-creative AI tools. Creative prompts involving image, narrative, and music generation could be further explored, as this represents a growing trend in GenAI usage.

6.1. ChatGPT Exploration Update

Considering that GenAI models are always changing, through self and human-on-the-loop optimisation, we conducted an extra interaction process to explore how much the model's response regarding prompt categorisation and complexity levels had changed. This was conducted after running the study with participants (May 6th, 2025). The free GPT-4.o model kept "informative", "creative", "problem-solving", and "analytical" categories; while sharing a new "interrogative", merging "educational" and "reflective" categories. Regarding complexity levels, ChatGPT kept the same three, but emphasised on style and structure. While the essential categorisation and complexity patterns remained largely unchanged, the observed differences may reflect updates in

training data and, potentially, algorithmic personalisation. For example, ChatGPT mentioned an "iterative" complexity for the first time in our interactions. This update could be explained by new state-of-the-art knowledge and evolving user behavior through iterative engagement. For more consistent analyses on GPT's dynamic capabilities (Bubeck et al., 2023), future studies might consider using the OpenAI API, which allows greater control over model parameters and versions, providing a more stable basis for tracking changes over time. Such longitudinal tracking could help not only to evaluate the model's capabilities for flexible, cross-domain reasoning, but to support research that anticipates critical questions regarding trust in AI and its implications for societal discourse.

7.1. Conclusion *Study One*

This work provides an empirical resource for understanding user-centered prompt categories in LLM-powered conversational agents. By systematically generating, categorising, refining and assessing prompts according to intent, complexity, appeal, and accessibility for non-experts, we offer a structured resource for future research and practical applications in human–AI interaction. The study assessed a set of prompts with consistent category agreement and comparable complexity levels. Complexity, interestingness and general domain ratings were balanced, reflecting prompts that are engaging, accessible, and appropriately challenging for a general population, while also enriching future research. Researchers could, for example, compare trust levels in users when engaging with LLMs on task-oriented vs. reflexive interactions; additionally, identifying differences in emotional responses between groups that engage with LLMs across these types of interaction, as the *Main Study* in this body of research explored.

The final set of prompts can support future studies on trust, affective responses, and other HCI constructs, enabling researchers and practitioners to design interactions that are more inclusive and effective. Importantly, the typology highlights the value of analysing interaction types that are similar in complexity and cognitive demand, allowing for a clearer understanding of the dynamics both within and between categories. Overall, this work demonstrates the importance of a user-centered approach in advancing human–AI interaction research, particularly for populations without specialised technical expertise, and provides a reproducible framework for exploring prompt design across diverse contexts.

CHAPTER THREE

Main Study

End users' interactions with ChatGPT

With the previously assessed prompts from *Study One*, the research advanced to the *Main Study*: end users' interactions with ChatGPT. The general objective was to evaluate if the type of interaction with ChatGPT had an impact on trust and if prediction accuracy was related to emotional responses. As the active inference framework suggests that trust in others depends on whether one can predict their behaviours/actions and if our predictions are accurate, a positive-valence emotion will be experienced (Hesp et al., 2021; Schoeller et al., 2021), we determined the general research questions to explore whether predictability explains the effect of type of interaction in trust when engaging with ChatGPT and if accuracy of prediction is related to emotional responses.

The specific research questions and hypotheses were:

RQ1 - Does type of interaction have an effect on trust?

H_0: Type of interaction does not have an effect on trust

H_a: Type of interaction has an effect on trust

RQ2.1 - Does the number of interactions have an effect on trust? *(exploratory)*

RQ2.2 - Does the number of interactions have an effect on predictability? *(exploratory)*

RQ3 - Is the relationship between type of interaction and trust explained by predictability?

H_0: Predictability does not explain the relationship between type of interaction and trust

H_a: Predictability explains the relationship between type of interaction and trust

RQ4 - Does prior knowledge moderate the relationship between type of interaction and predictability? *(exploratory)*

RQ5 - Do attitudes towards AI moderate the relationship between type of interaction and trust? *(exploratory)*

RQ6 - Is the accuracy of prediction related to emotional valence?

H_0: Accuracy of prediction is not related to emotional valence

H_a: Accuracy of prediction is related to emotional valence

RQ7 - Is the accuracy of prediction related to epistemic emotions such as boredom, curiosity, surprise, anxiety, excitement, confusion, or frustration?

H_0: Accuracy of prediction is not related to epistemic emotions

H_a: Accuracy of prediction is related to epistemic emotions

See figures 3.1-3.3 for visual models.

1.1. Method

This experimental study with a quantitative pre-post methodology was framed within the active inference theory, including within and between groups comparisons. This study examines how different types of interaction (task-oriented vs. reflexive) affect user trust in AI, while taking into account their AI prior knowledge and attitudes towards AI. Participants engaged with ChatGPT through prompts that had been pilot-tested in a previous research. For each interaction type, assessments were administered at two phases (pre, post). The study employed a 2 (Interaction type) × 2 (Phase) within-subjects design. The study intercalated interactions with ChatGPT and self-assessment questionnaires. Data collection took place between November 24th, 2024 and May 29th, 2025.

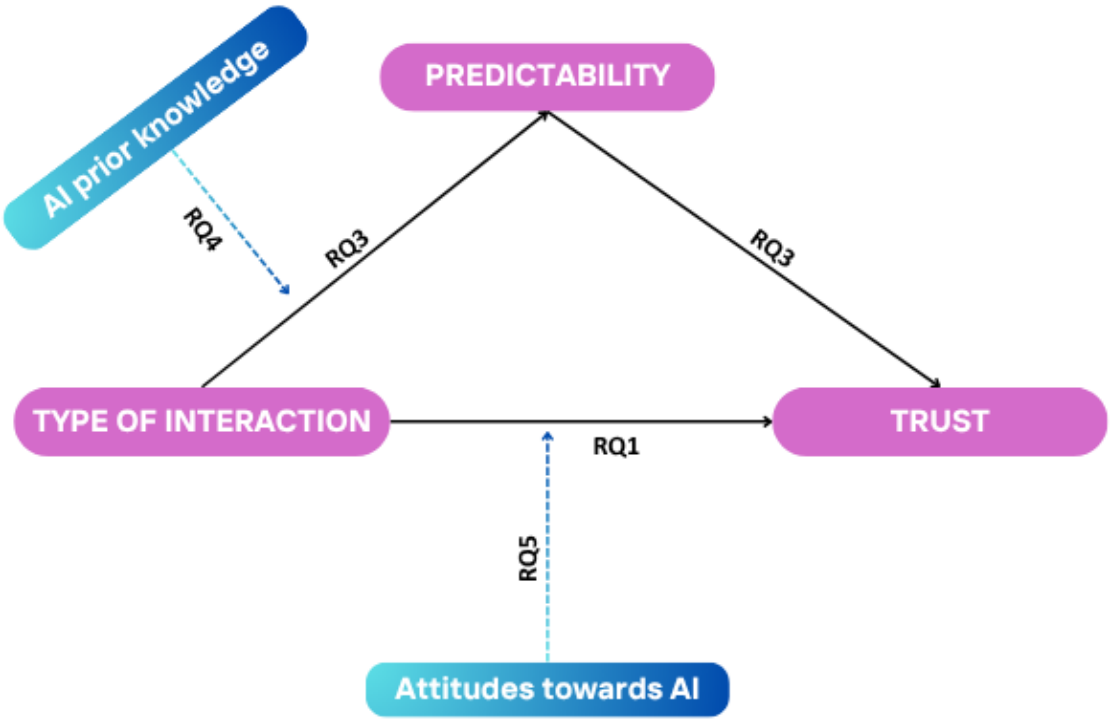

Figure 3.1. Model 1: Model for mediation and moderation *(RQ1, RQ3 - RQ5*)
Dotted lines show exploratory questions.

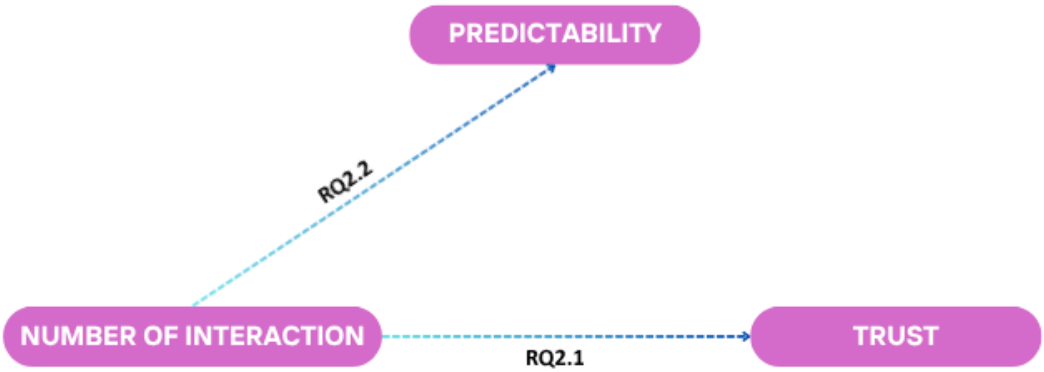

Figure 3.2. Model 2: Model for *RQ2.1* and *RQ.2.2*
Dotted lines show exploratory questions.

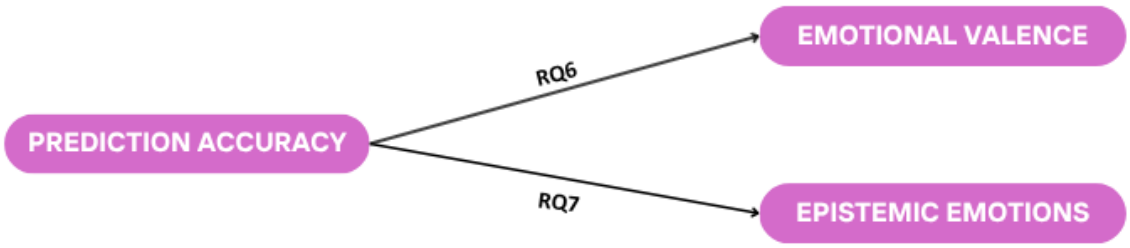

Figure 3.3. Model 3: Model for *RQ6* & *RQ7*

1.1.1. Ethics

The subject research was approved by the Specialised Committee on Ethics in Psychology of ISCTE-IUL (PSI_24/2024, September 2024) and can be consulted in Appendix C (same as *Study One*). Informed consent established that participants could decide to participate or not, abandon the study at any time, and decide to or not to answer questions. It also referred participants to OpenAI privacy policy for the interactions with ChatGPT. More details can be found in Appendix G. Only the individuals who accepted voluntarily to participate through the informed consent took part in the research; otherwise, they were redirected to a thank you page. Participants were redirected to the debriefing (Appendix H) after finishing the study. No costs or risks were associated with participating in the study, while participants could contribute to the human-AI interaction research. Participants could interrupt the study or choose not to respond to questions whenever they wanted.

1.1.2. Participants

1.1.2.1. Recruitment methods

In line with *Study One*, inclusion criteria for participating in the *Main Study* was also being older than 18 years and having at least an intermediate domain in written English. Participants that did not meet the inclusion criteria did not take part in the research and were redirected to a thank you page. No specific requirements on AI knowledge, technical abilities or other technology-related expertise were necessary to participate. However, participants were required to have an email address in order to register an account on the OpenAI page to interact with ChatGPT.

Participants were recruited using convenience sampling, the snowball technique, and institutional outreach through:

a) Social Media, publications (Instagram, LinkedIn and Facebook)

b) LAPSO partnership to grant a pool of Psychology and Social Sciences students

c) Social Media, personalized invitations to participate (Instagram & LinkedIn)

d) Research units, Labs, and schools in Portugal and Mexico, personalized invitations to share with their community

e) Partnership with Nativa, a Mexican technology company focused on Conversational AI solutions

The invitations and publications included a general description of the study and inclusion criteria. Based on the feedback from the pilot study, we also shared the approximate duration of the one-time intervention (35 minutes) and recommendation to participate using a computer (as questionnaires and interactions with ChatGPT were intercalated).

For a), 22 different publications were made from the researcher's and mentor 's personal accounts. For b), two requests were made for LAPSO partnership, one was accepted and a pool of 50 participant students were granted. For c), 125 personalized invitations were sent to the researcher's network on LinkedIn; and 93 through Instagram. On d), 26 different schools and organizations were reached, including all ISCTE labs; research groups and labs from Universidade Lusofona in Lisbon and Porto; different research groups from Universidad Autónoma Metropolitana and Tecnologico de Estudios Superiores de Monterrey in Mexico. A total of 218 personalized e-mails for directors, coordinators, secretarial staff, researchers, and professors were sent asking for approval and support in disseminating the study with their communities. Researchers and coordinators from both academic institutions in Mexico, the CIES and ISTAR lab units from ISCTE, Institudo das Telecomunicações in Lisbon, Nursing School of Coimbra, and the Universidade de Coimbra confirmed their support in disseminating our study with their network and community. For e), Nativa shared 5 publications from the organizational account on LinkedIn for clients and partners, 1 section in the

internal newsletter was published, and 3 corporate internal invitations were made for the employees to participate. Prospective participants were provided with an email address for any inquiries about the study. Written and graphic examples of the invitations sent can be found in Appendix I.

1.1.2.2. Sample description

A total of 214 participants were recruited; of these, 2 did not consent to participate in the study, 32 did not meet the inclusion criteria (8 were under 18 years old and 24 reported basic English level), and 61 participants did not finish the study. This resulted in 119 participants who finished the study, from those, 9 participants were not considered in data analyses as their responses did not comply with quality standards. Details on response quality can be found in the data analysis section.

The final sample comprised 110 participants (*M*age = 28.24, *SD*age = 10.43, *Min* = 18, *Max* = 69); 61.8% were female while 35.5% were male and 2.7% identified as non binary or third gender. The majority had an intermediate English level (57.3%), while the remainder reported an advanced level. Participants held higher education (32.7%), masters' degree (22.7%), and post-secondary (22.7%) fulfilled studies. Participants were mainly residents in Portugal (60.9%) and Mexico (20.9%), although they reported other nationalities such as Spanish, Brazilian and German. Occupations varied, with 37.3% being full time students, 57.1% professionals from widely diverse areas such as business (17.3%), STEM (10%), or creative industry (5.5%), and 5.6% working students. Participants were distributed into 2 different groups (Group 1 n = 53; Group 2 n = 57), through true randomisation managed automatically by Qualtrics Software.

1.1.3. Materials and Measures

A subset of 6 prompts were selected from *Study One* to guide the interactions with ChatGPT. These prompts shared complexities as closely as possible across both categories, which allowed us to compare how participants engaged with different interaction types while still controlling for level of complexity. Prompts were displayed according to the type of interaction to which participants were exposed to (three at a time). One prompt was presented as an example, and participants were instructed to select one of the remaining two prompt options for their interaction with ChatGPT. A list of the selected prompts for the *Main Study* and their summarised results from *Study One* can be found in Table 3.1.

The following constructs were assessed: 1) Attitudes towards AI (Appendix J), 2) AI Prior Knowledge (Appendix K), 3) Predictability for AI Interactions (Appendix L), 4) Trust (Appendix M), 5) Emotional Valence, and 6) Epistemic-Related Emotions (both emotional responses can be found on Appendix N). Demographic questions were asked at the end of the study. All information on scales

can be found in the appendices previously indicated to support transparency, replicability, and open science collaboration.

Table 3.1. Prompts from *Study One* selected for the *Main Study*, with their respective complexity levels and category confirmation rates.

Category	Prompt	M Confirmation Rate	M Complexity Level
Task-oriented	"Propose a customised strategy to improve sleep quality, contributing to overall well-being" (ACTUAL CUE)	71.43%	3.79
Task-oriented	"Propose strategies to reduce my household expenses" (ACTUAL CUE)	71.43%	3.43
Task-oriented	"Create a detailed itinerary for a two-week vacation in Europe" (EXAMPLE)	82.14%	3.29
Reflexive	"Help me find inspiration to pursue a new hobby / interest" (ACTUAL CUE)	71.43%	3.68
Reflexive	"(Thinking about a challerge in my life) give me your feedback on how I handled it." (ACTUAL CUE)	67.86%	4.64
Reflexive	"Help me reflect on the effectiveness of my current behavior and habits for my personal development" (EXAMPLE)	64.29%	5.04

Note. The table presents the actual prompt cues participants could select for their interactions and the example prompts, provided for reference.

1) *Attitudes towards AI* were assessed using the AI Attitude Scale (AIAS-4; Grassini, 2023), which comprises four items (e.g., *I think* AI technology *is positive for humanity*). Responses were recorded on a 10-point scale ranging from 1 (*do not agree at all*) to 10 (*completely agree*); higher scales account for more positive attitudes. This scale was specifically developed for modern consumer-oriented and conversational AI systems, such as ChatGPT. It was selected as it is brief while maintaining good psychometric properties. Items were integrated as a composite variable.

2) AI prior knowledge included different assessments:

2.1) *AI familiarity* was assessed using two adapted items from the familiarity factor of Körber (2019) e.g. *I already know ChatGPT or similar systems as Co-pilot, Dall-E or other intelligent chatbots*, on a 5-point Likert scale (1-*strongly disagree* to 5-*strongly agree*). Items were integrated as a composite variable.

2.2) *Nature of previous interactions*, by inquiring participants about the category of their previous interactions with LLMs (*personal, professional, none, both*)

2.3) *Frequency of use* based on Vizcaino et al. (2019), where participants directly reported their estimated usage of LLMs in hours by week. This method was selected over a categorical Likert-type scale present in previous ChatGPT usage research (Fang et al., 2025; Phang et al., 2025) to obtain precise continuous data, thus avoiding the methodological limitations of unequal intervals inherent in ordinal scales.

2.4) *AI Literacy* using the 31-item scale for non-experts by Laupichler et al. (2023a), which uses a 7-point Likert format (1-*strongly disagree* to 7-*strongly agree*), and comprises 3 factors: technical understanding (14 items; e.g. *i can describe how machine learning models are trained, validated, and tested*), critical appraisal (10 items; e.g. *I can explain why data privacy must be considered when developing and using artificial intelligence applications*), and practical application (7 items; e.g. *I can give examples from my daily life (personal or professional*) *where I might be in contact with artificial intelligence*). Each factor was then analysed as composite variables and overall AI Literacy considering the three factors together was also analysed as one composite. This specific scale was analysed by factors as it measured separate aspects of literacy; we also wanted to analyse the sample from different perspectives of literacy as participants were non-experts but potentially tech savvy, finally because the scale retained good reliability considering factors separately.

Criteria for detecting careless responses were included, using an attention check item (*please check somewhat disagree for this item*) and a bogus item (*I consider myself among the top 10 AI researchers in the world*). Questions appeared in random order.

3) *Predictability* refers to the perceived capacity to anticipate events and verifying that anticipation against the actual outcome; accordingly, participants evaluated it before and after each interaction through *prediction* (before the interaction) and *prediction verification* (after the interaction) scales. It was assessed with a custom-report with three factors, all with three items evaluated on a 5-point likert scale: 3.1) Perceived Value Similarity AI (adapted from Yokoi et al., 2021; Yokoi & Nakayachi, 2021), from 1 (*do not agree at all*) to 5 (*strongly agree*). 3.2) Understanding and predictability (adapted from Körber, 2019), from 1 (*strongly disagree*) to 5 (*strongly agree*), and 3.3) Uncertainty and expectation with custom-questions based on the theoretical construct of Predictability, from 1 (*do not agree at all*) to 5 (*strongly agree*). Example items for each factor and moment of assessment are shown in Table 3.2 for clarity. Questions for each moment of prediction appeared in random order and the nine items were analysed as a composite variable.

4) *Trust* was assessed using the Human Computer Trust Scale (HCTS; Gulati et al., 2019), comprising 12-items distributed across four factors: Perceived risk, benevolence, competence, reciprocity (e.g. *I think that ChatGPT performs its role as conversational AI very well or I can trust the information presented to me by ChatGPT*). Responses were recorded on a 5-point likert scale from 1 (*strongly disagree*) to 5 (*strongly agree*). This specific scale was selected as it was developed for consumer-oriented AI systems, included the possibility to adapt wording for specific AI agents, and, as the questions appeared before and after each of the two types of interactions, because it accounts for future scenarios of trust. Also, this scale included key components from the active inference framework such as reciprocity and perceived risk. Questions for each time of measure (pre/post interaction) appeared in random order and were analysed as a composite variable. Criteria for

detecting careless responses were included, using the attention check item: *please check somewhat disagree for this item*.

Table 3.2. Example items for predictability, by factor and time of measure (before and after interaction)

Factor	Example item on Prediction (Before interaction)	Example item on Prediction Verification (After interaction)
Perceived Value Similarity AI	ChatGPT will have similar beliefs on topics concerning (task-oriented/reflective) interactions as I do	ChatGPT has similar beliefs on topics concerning (task-oriented/reflective) interactions as I do
Understanding/Predictability	I will be able to understand why things happened	ChatGPT responded unpredictably RI was able to understand why things happened
Uncertainty and Expectation	I have a clear expectation on how ChatGPT will perform in this task	My expectation on how ChatGPT would perform this task was met

Note. Participants saw either "task-oriented" or "reflexive" depending on which type of interaction was taking place.

5) *Emotional valence* was assessed as a two-dimensional construct (Briesemeister et al., 2012) by measuring the intensity of positive and negative emotions, each valence rated independently on a Likert-scale from 1 (*low*) to 7 (*high*).

6) *Epistemic emotions* were assessed using the Epistemically-Related Emotion Scale (Pekrun et al., 2017) through a 5-points Likert scale from 1 (*not at all*) to 5 (*very strong*), with 7 different emotions: boredom, curiosity, surprise, anxiety, excitement, confusion, and frustration. Emotions appeared in random order. This scale was selected as it includes key emotions considered in the active inference framework and also as users confront an epistemic challenge with ChatGPT not being detailedly understandable as previous research has stated (Schoeller et al., 2021).

Table 3.3 shows a visual summary of the scales assessed, the respective variables they account for, and each variable statutes.

1.1.4. Procedure

From the 12 confirmed prompts in *Study One*, we selected a subset of 6 prompts for the *Main Study*, prioritizing those with the highest confirmation rates while keeping comparable levels of complexity across interaction types. Since prompts in the reflexive category showed lower confirmation rates, we first identified those with the highest confirmation and then examined their range of complexities. Next, we considered the prompts recognized as *task-oriented* that had complexities similar to the *reflexive* ones, while keeping good confirmation rates on their category. Finally, we included as 'example' the prompt that represented an outlier in complexity level compared to the other selected prompts, being the most complex in the case of the *reflexive* prompts and the least complex in the case of *task-oriented* ones.

Table 3.3. Scales, variables, statutes by model

Variable	Description	Type	Statute
Model 1 (*RQ1, RQ3 - RQ5*)			
Type of interaction	Task-oriented or reflexive	Exposure	Independent variable
Predictability	Prediction (before interaction) & prediction verification (after interaction)	Likert-scale	Mediator variable
Trust	Human Computer Trust Scale (HCTS) (before & after interaction)	Likert-scale	Dependent variable
AI prior knowledge	Frequency of use (in hours per week)	Scale	Moderator path a (exploratory)
	AI familiarity	Likert-scale	
	AI literacy with 3 factors (technical understanding, critical appraisal, practical application)	Likert-scale	
Attitudes towards AI	AI Attitude Scale (AIAS-4)	Likert-scale	Moderator path c (exploratory)
Model 2 (*RQ2.1 & RQ 2.2*)			
Number of interaction	First or second	Exposure	Independent variable
Predictability	Prediction (before interaction) & prediction verification (after interaction)	Likert-scale	Dependent variable
Trust	Human Computer Trust Scale (HCTS) (before & after interaction)	Likert-scale	Dependent variable
Model 3 (*RQ6 & RQ7*)			
Prediction accuracy	Mathematical difference of prediction verification (-) prediction, in absolute numbers	Mathematical Calculation, scale	Independent variable
Emotional response	Epistemic emotions (boredom, curiosity, surprise, anxiety, excitement, confusion, or frustration)	Likert-scale	Dependent variable
	Valence (positive & negative) independently assessed	Likert-scale	Dependent variable

With the prompts previously selected, two types of interactions took place: (a) *Task-oriented* and (b) *Reflexive*. All participants took part in both types of interactions but were assigned to either condition (a) or (b) for their first interaction with ChatGPT through randomization handled automatically within Qualtrics for evenly distribution. For their second interaction, they engaged in the opposite condition. This counterbalanced exposure to either of the conditions as a first interaction allowed us to prevent and test for order effects. Each interaction with ChatGPT should have lasted approximately 5 minutes, nevertheless this was not monitored.

At the very beginning of the study, participants answered the attitudes towards AI and AI prior knowledge scales as a baseline assessment to avoid biased responses. Secondly, they were instructed to either create an account in the OpenAI page using their email address or simply login into their existing account; afterwards, they received general instructions on how to interact with ChatGPT for this study (Appendices O & P). Before interacting, participants responded to questions on prediction and trust with regards to the interaction that would be about to happen. Then, they had to actually interact with ChatGPT in a semi-structured way by choosing one of the two available prompts while personalising a prompting template (Schulhoff, 2025) using it. They were provided with detailed explanations on how to fill it and an example as reference (Appendix Q). After they finished this first

interaction, they answered questions on prediction verification and trust with regards to the interaction that had just taken place. Additionally, they responded to scales for emotional valence and epistemic-related emotions. Participants were asked to confirm and select the prompt sent to ChatGPT (Appendix R) for validation and future analysis. After this, they received instructions to interact with the remaining type of interaction, either (a) or (b), and repeated the previous steps: answer the scales of prediction and trust (before interaction & after interaction), rate the intensity experienced for emotional valence and epistemic-related emotions (after interaction), and select the prompt used. Lastly, they answered the demographics questionnaire and were redirected to the debriefing document. A descriptive flowchart can be found in figure 3.4.

Scales and instructions were presented in English. All data was collected online through Qualtrics software (license provided by ISCTE-IUL) and interactions happened directly through the ChatGPT platform, with no visualisation from the researcher.

Due to the potential complexity of the study, we decided to run a pilot study. PhD students from ISCTE were invited to participate, but none agreed to take part. Consequently, we proceeded to run the pilot study with five people from a convenience sample from different backgrounds. In this pilot study, we measured approximate time of completion, clarity upon indications, and general navigation upon the survey platform and ChatGPT interactions. Participants shared their feedback either orally or through a brief survey (questions can be found in Appendix S). Considering this feedback, minor adjustments were made into the survey and recruitment information, e.g. we improved the navigation in the Qualtrics survey to keep visible the response options for scales with many items and updated the estimated time of completion in the invitations and informed consent. The results of these 5 participants were considered in the overall analysis of the sample.

1.1.5. Data Analysis

The data analysis was conducted in IBM SPSS Statistics (Version 28) and jamovi (Cloud and Desktop) (Gallucci, 2019, 2020; Ludecke et al., 2020; R Core Team, 2024; The jamovi project, 2024).

Prior to the analysis, several data preparation steps were performed. Response quality using items from Laupichler et al., (2023a) was monitored through five attention checks distributed across the questionnaires (*check somewhat disagree for this item*), one bogus item (*I consider myself among the top 10 AI researchers in the world*), and explicit confirmation of having interacted with ChatGPT (*did you interact with ChatGPT on a task/reflexive conversation?*). Based on these criteria, the response quality was identified and categorised into different groups (excellent, high, substandard, and low). Nine participants were excluded from data analysis (2 as “low quality” since they neither passed the bogus item nor any of the attention checks; 4 from "substandard quality” by passing attention checks but failing bogus item; and 3 for answering ‘*no*’ on the task-oriented and reflexive

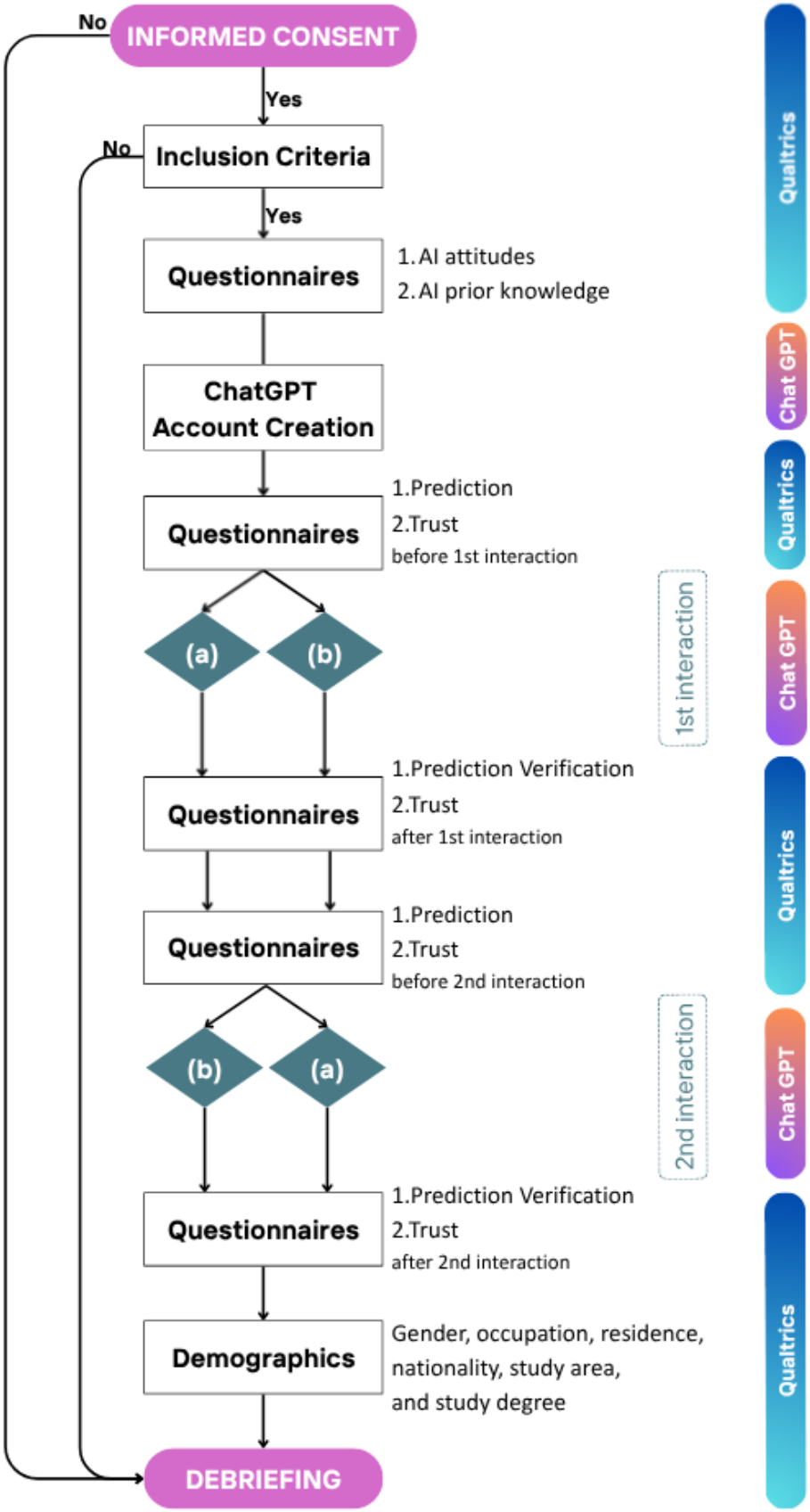

Figure 3.4. Descriptive flowchart of *Main Study* procedure

conditions when inquired if they had had an interaction with ChatGPT).

The dataset was checked upon consistency between variable labels and their corresponding values to confirm accurate coding. Reversed items on scales were re-scored when required. After confirming internal reliability using either Cronbach's alpha or Spearman-Brown coefficient, composite variables (attitudes towards AI, AI familiarity, AI literacy, trust before and after interaction, prediction, and prediction verification) were constructed by averaging the corresponding responses, following the methodology of the original scales authors (Grassini, 2023; Gulati et al., 2019; Körber, 2019; Laupichler et al., 2023a).

Demographics and baseline comparisons between groups were conducted to ensure they did not differ significantly. Chi-square tests were performed on categorical variables, whereas for quantitative measures, like AI literacy, AI familiarity, age, or attitudes towards AI, means were compared through independent-sample t-tests or Mann-Whitney when normality was not assumed.

Data was prepared in both wide and long formats, depending on the required analysis. Wide format was used for initial validation checks, descriptive analyses, one time measures, and emotional responses, preserving one row per participant, whereas long format was employed for repeated measures and mixed-effects models, in which each participant was represented in multiple rows for each condition and time of measure. Data consistency was verified in both formats as a quality check.

Specifically, *predictability* underwent different analytical treatments depending on the research question. For analyses that included all measurement moments, prediction and prediction verification scores from both interactions were entered as repeated measures in mixed-effects models. For mediation analyses, the mathematical difference between prediction verification and prediction (post - pre) was calculated separately for each moment (first and second interaction), and these difference scores were then used in distinct mediation models. Finally, prediction accuracy was calculated from the absolute difference between prediction verification and prediction scores, disregarding directionality (positive or negative numbers), as the magnitude of deviation was the relevant indicator for these analyses, as shown previously in Table 3.3.

To answer *RQ1 - Does type of interaction have an effect on trust?, RQ2.1 (exploratory) - Does the number of interactions have an effect on trust? & RQ2.2 (exploratory) - Does the number of interactions have an effect on predictability?*: we decided to use Linear Mixed Models (LMMs) as this statistical approach is ideal for the design of our study since it handles for non-independent collected measures from the same participant at different moments (Argentzell, 2020; Brown, 2021) while traditional repeated-measures (as ANOVA or t-test) do not account for data nested within participants (Yu et al., 2022). Variables of type of interaction with ChatGPT, time of measure (pre/post), or interaction number (first/second) were included as fixed effects according to the

model tested, and we analysed their main effects as well as their interaction. Participant's ID was modeled as a random effect, specifically allowing the intercept to vary across participants.

To analyse *RQ3 - Is the relationship between type of interaction and trust explained by predictability?*: We ran an LMM to analyse if predictability had a significant effect in the main model by examining the main effect of type of interaction and predictability as well as the interaction between them on trust. We also verified if type of interaction had effects on predictability. Then we ran another LMM also with trust as dependent variable to analyse main effects of type of interaction with ChatGPT, predictability and order (first/second); also interactions between type of interaction with ChatGPT x predictability and order x predictability to see if predictability remained relevant in the complex model.

After that, we executed two simple mediation analyses using the PROCESS MACRO (Hayes, 2022) with type of interaction with ChatGPT as independent variable, predictability as mediator, and trust as dependent variable. Since we could not run a multivariate mediation accounting for all pre/post measures, in these simple mediation models the mediator and dependent variables were calculated with the mathematical difference of post (-) pre. The first mediation was with the data from the first moment of interaction with ChatGPT, while the second considered the interaction number two that participants had with ChatGPT.

To answer *RQ4 (exploratory) - Does prior knowledge moderate the relation between type of interaction and predictability?* and *RQ5 (exploratory) - Do attitudes towards AI moderate the relation between type of interaction and trust?*: LMMs were used to test for moderation effects, keeping the models specifications consistent with the previously described for the analyses of *RQ1* and *RQ2*. Since a moderated mediation showed convergence issues, likely because of the model's complexity and sample size, they were assessed on individual paths to specifically address the moderation (path a for *RQ4* and path c for *RQ5*), acknowledging that the results may be less optimal but still informative for the research questions. See figure 3.5 for a visual reference.

For *RQ6 - Is the accuracy of prediction related to emotional valence?* and *RQ7 - Is the accuracy of prediction related to epistemic emotions such as boredom, curiosity, surprise, anxiety, excitement, confusion, or frustration?,* we analysed the correlations between prediction accuracy and emotional responses using the Spearman correlation test (Bryman & Cramer, 1993) as distribution was not normal and each emotional response was assessed individually with a likert-scale item.

This research addressed unexplored areas of human-AI interaction. Given the scarcity of prior empirical work on lay users' trust and predictability when interacting with LLMs, some research questions were designed to explore patterns to further help generate hypotheses rather than test pre-established predictions. Although these questions were exploratory, analyses followed a rigorous, pre-specified methodology, including structured pre-post assessments and carefully

selected prompts. Our exploratory analyses were guided by the active inference framework, allowing for examination of how AI prior knowledge might have moderator effects on predictability, how attitudes towards AI may moderate trust when eliciting different types of interactions, and if number of interactions might have an effect on trust and predictability.

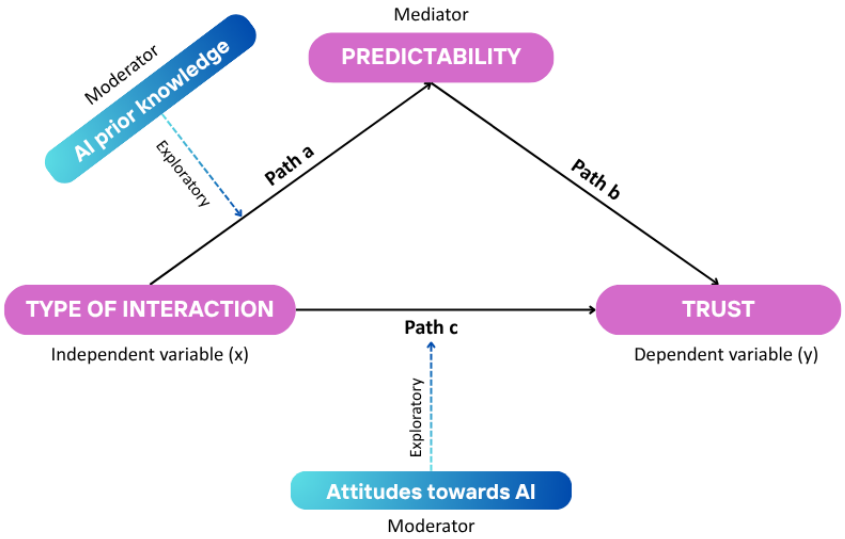

Figure 3.5. Paths for mediation (*RQ3*) and exploratory moderations (*RQ4 & RQ5*)

2.1. Results

Baseline and demographic comparisons indicated no significant differences between the two groups across measured variables. Table 3.4 presents descriptive percentages for sociodemographic data by group and total sample, as well as group comparisons. Table 3.5 shows descriptive statistics for continuous scales (mean and standard deviation) by group and the total sample, internal reliability for composites, and the results of group comparison tests. Exploratory moderators (AI prior knowledge and attitudes towards AI) were measured at baseline and are included in this table for clarity. Missing data was not an issue as all participants completed the full set of measures, probably because only high data quality was retained after screening. For the variable of frequency of use, 12 participants did not provide responses. The descriptive statistics for the composite variables of prediction and trust by group and total sample with their internal reliabilities can be found in Table 3.6.

Table 3.7 presents descriptive statistics for the emotional response ratings across groups and types of interaction. Figure 3.6 presents the mean scores for each emotion across interaction types, to illustrate the overall emotional profile elicited by each scenario type, highlighting relative differences in affective responses. These descriptive results provide an overview of the emotional responses elicited by the scenarios prior to hypothesis testing.

Table 3.4. Sociodemographic characteristics for all participants, distributed by group, including comparisons between groups.

Categorical measures	Group 1 (Reflexive interaction first)	Group 2 (Task-Oriented interaction first)	Total	Group Comparison
	Frequencies (in percentage %)			Chi Square Test
English level				
Independent user (Intermediate)	60.4	54.4	57.3	$x^2(1) = .403$, $p = .526$
Proficient User (Advanced)	39.6	45.6	42.7	
Nature of previous interactions with ChatGPT				
Personal	11.3	7.0	9.1	
Professional	20.8	26.3	23.6	$x^2(3) = 6.348$, $p = .097$
Both	54.7	64.9	60.0	
None	13.2	1.8	7.3	
Gender				
Female	67.9	56.1	61.8	
Male	30.2	40.4	35.5	$x^2(2) = 1.682$, $p = .451$
Non binary / Third	1.9	3.5	2.7	
Highest level of fulfilled studies				
Secondary (Middle school, lower secondary, etc)	15.1	26.3	20.9	
Post-secondary (High school, upper secondary, vocational, etc)	30.2	15.8	22.8	
Higher education (University/Polytechnic/Bachelor degree, etc)	30.2	35.1	32.7	$x^2(4) = 5.437$, $p = .217$
Master's degree or equivalent	22.6	22.8	22.7	
Doctoral, post doctoral degree	1.9	0.0	0.9	
Area of study				
Social Sciences	52.8	47.4	50.0	
Business & Economics	17.0	10.5	13.6	
STEM	11.3	10.5	10.9	
Mixed/Interdisciplinary	5.7	12.3	9.1	$x^2(7) = 4.278$, $p = .775$
Design & Architecture	3.8	7.0	5.5	
Health & Medical Sciences	5.7	3.5	4.5	
Arts & Humanities	1.9	5.3	3.6	
Other	1.9	3.5	2.7	
Ocuppation				
Full time Student	37.7	36.8	37.3	
Business, Finance & Sales	13.2	21.1	17.3	
STEM	7.5	12.3	10	
Creative (Architecture, Art, design, & Others)	1.9	8.8	5.5	
Working Student	5.7	5.3	5.5	$x^2(10) = 15.901$, $p = .090$
Education	5.7	3.5	4.5	
Hospitality & Tourism	7.5	1.8	4.5	
Social Sciences	9.4	0.0	4.5	
Other	5.7	3.5	4.5	
Health & Medical Sciences	5.7	1.8	3.6	
Mixed Professions	0.0	5.3	2.7	

Table 3.4. (Continues)

Nationality				
Portuguese	49.1	49.1	49.1	
Mexican	24.5	21.1	22.7	
Spanish	11.3	7.0	9.1	$x^2(5) = 2.839$
Other Latinamerican	9.4	8.8	9.1	$p = .746$
Other European	3.8	7.0	5.5	
Other	1.9	7.0	4.5	
Country of Residence				
Portugal	60.4	61.4	60.9	
Mexico	22.6	19.3	20.9	
Spain	5.7	7.0	6.4	$x^2(5) = 2.178$,
Europe (non-iberian)	3.8	7.0	5.5	$p = .846$
Other Latin America	5.7	1.8	3.6	
Other	1.9	3.5	2.7	

Table 3.5. Baseline measures for all participants, distributed by group, including comparisons between groups and internal reliability for composite scales.

Baseline measures	Reliability*	Group 1 (Reflexive interaction first) n = 53	Group 2 (Task-Oriented interaction first) n = 57	Total n = 110	Group Comparison
		Mean (SD)			T-test / Mann Whitney**
Age	-	27.70 (9.80)	28.79 (11.05)	28.24 (10.43)	$U = 1487.50$, $Z = -0.138$, $p = .890$
Frequency of use (in hours per week)	-	2.81 (3.16)	2.44 (2.83)	2.63 (3.00)	$U = 1179.500$, $Z = -0.152$, $p = .879$
AI Familiarity (Likert-scale 1-5)	$rSB = .794$	4.06 (1.04)	4.42 (0.74)	4.24 (0.91)	$U = 1238.500$, $Z = -1.727$ $p = .084$
AI Literacy - Critical Appraisal Factor (Likert scale 1-7)	$\alpha = .885$	5.25 (0.96)	5.44 (0.85)	5.35 (0.91)	$t = -1.048(108)$, $p = .297$
AI Literacy - Practical Application Factor (Likert scale 1-7)	$\alpha = .832$	4.62 (1.05)	4.98 (1.00)	4.8 (1.03)	$t = -1.857(108)$ $p = .066$
AI Literacy - Technical Understanding (Likert scale 1-7)	$\alpha = .940$	3.16 (1.15)	3.08 (1.31)	3.12 (1.24)	$t = .354(108)$ $p = .724$
AI Literacy - all (Likert scale 1-7)	$\alpha = .948$	4.16 (0.91)	4.26 (0.97)	4.22 (0.95)	$t = -.571(108)$ $p = .569$
Attitudes towards AI (Likert scale 1-10)	$\alpha = .897$	6.94 (1.92)	7.32 (1.98)	7.13 (1.96)	$t = -1.020(108)$ $p = .310$

Note. All participants completed the study measures. Missing data (n = 12) occurred only for frequency of use

* Spearman Brown reliability test was used for two item scales; Cronbach Alpha was used for all other scales

** Mann Whitney was used for not normal distributed variables; t-test for normally distributed ones.

Table 3.6. Descriptive statistics for composite variables of predictability (mediator) and trust (dependent), by group and total sample with their internal reliabilities.

Variables	Reliability	Group 1 (Reflexive interaction first) n = 53	Group 2 (Task-Oriented interaction first) n = 57	Total n = 110
		Mean (SD)		
Prediction Reflexive	α = .816	3.11 (0.08)	2.90 (0.08)	3.01 (0.08)
Prediction Verification Reflexive	α = .768	3.54 (0.07)	3.42 (0.08)	3.48 (0.07)
Prediction Task-Oriented	α = .791	3.42 (0.07)	3.39 (0.07)	3.41 (0.07)
Prediction Verification Task-Oriented	α = .787	3.58 (0.07)	3.77 (0.07)	3.67 (0.07)
AI Trust Reflexive (Pre)	α = .767	2.73 (0.07)	2.98 (0.07)	2.86 (0.07)
AI Trust Reflexive (Post)	α = .846	2.96 (0.08)	3.20 (0.08)	3.08 (0.08)
AI Trust Task-Oriented (Pre)	α = .820	3.12 (0.07)	3.25 (0.08)	3.19 (0.08)
AI Trust Task-Oriented (Post)	α = .809	3.11 (0.07)	3.37 (0.07)	3.24 (0.07)

All variables were measured on a 1-5 points Likert scale

Table 3.7. Descriptive statistics for epistemic emotion and valence ratings across groups and interaction types

Emotional responses	Type of interaction	Group 1 (Reflexive interaction first) n = 53	Group 2 (Task-Oriented interaction first) n = 57	Total n = 110
		Mean (SD)		
Surprise	Task-Oriented	2.79 (1.15)	2.54 (1.25)	2.66 (1.21)
	Reflexive	2.94 (1.23)	2.82 (1.18)	2.88 (1.20)
Curiosity	Task-Oriented	3.32 (0.99)	3.46 (1.08)	3.39 (1.04)
	Reflexive	3.81 (0.81)	3.70 (0.92)	3.75 (0.86)
Excitement	Task-Oriented	2.75 (1.12)	2.72 (1.22)	2.74 (1.17)
	Reflexive	2.70 (1.17)	3.02 (1.28)	2.86 (1.23)
Confusion	Task-Oriented	1.72 (0.98)	1.67 (0.97)	1.69 (0.97)
	Reflexive	1.94 (1.06)	1.74 (0.95)	1.84 (1.00)
Anxiety	Task-Oriented	1.53 (0.89)	1.70 (1.05)	1.62 (0.97)
	Reflexive	1.92 (1.01)	1.51 (0.94)	1.71 (0.99)
Frustration	Task-Oriented	1.49 (0.82)	1.44 (0.75)	1.46 (0.78)
	Reflexive	1.57 (0.77)	1.46 (0.88)	1.51 (0.83)
Boredom	Task-Oriented	1.92 (1.08)	1.98 (1.15)	1.95 (1.12)
	Reflexive	2.13 (1.19)	1.91 (1.16)	2.02 (1.18)
Positive Valence	Task-Oriented	4.58 (1.47)	4.96 (1.22)	4.78 (1.35)
	Reflexive	4.77 (1.52)	4.84 (1.46)	4.81 (1.48)
Negative Valence	Task-Oriented	2.30 (1.32)	2.07 (1.20)	2.18 (1.26)
	Reflexive	2.58 (1.55)	2.09 (1.34)	2.33 (1.46)

Variables were measured in Likert scale (1–7 for valence, 1-5 for epistemic emotions)

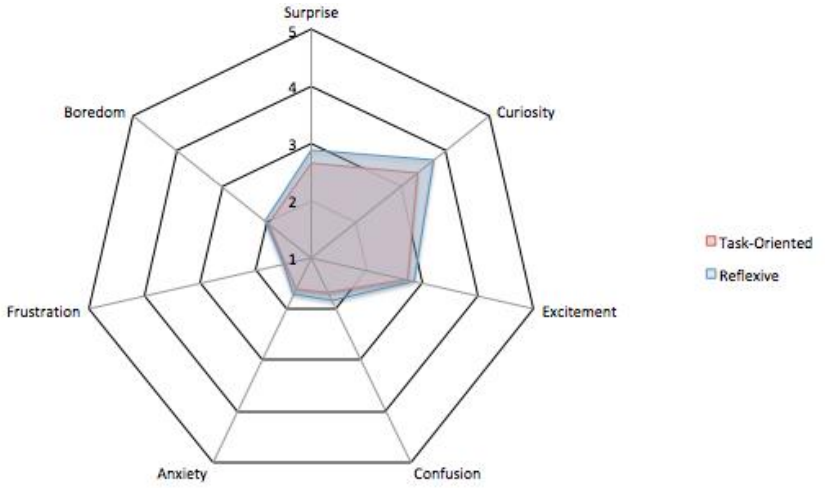

Figure 3.6. Radar chart displaying mean ratings for epistemic emotions by interaction type. Likert scale from 1 (not at all) to 5 (very strong)

Although all instructions were given in English, including the specific prompt and personalization prompting template, participants were "free" to interact in the language that they preferred as this was neither specified nor forbidden. Table 3.8 shows distributions by group and language.

Table 3.8. Language of interaction by group

Language of interaction with ChatGPT	Group 1 (Reflexive interaction first) n = 53	Group 2 (Task-Oriented interaction first) n = 57	Total	Group Comparison
	Frequencies (in percentage %)			Chi Square Test
English	69.8	78.9	74.5	$x^2(4) = 5.292$, $p = .209$
Spanish	20.8	8.8	14.5	
Portuguese	7.5	10.5	9.1	
Other (mother language)	0.0	1.8	0.9	
Other (not mother language nor any above)	1.9	0.0	0.9	

2.1.1. Results for Research Questions

RQ1 examined if type of interaction with ChatGPT has an effect on trust. The model showed statistical relevance (conditional R^2 = .707, marginal R^2 = .041, p = < .001 for both). Results indicated a significant main effect of the type of interaction with ChatGPT in trust (F(1, 329) = 59.0, p = < .001) with task-oriented interactions presenting mildly higher estimated means in trust. Even more, when adding time of measure (pre/post) the model was significant (conditional R^2 = .712, marginal R^2 = .059, p = < .001 for both) and results indicated a significant main effect of type of interaction with

ChatGPT (F(1, 327) = 61.33, p = < .001) in trust. In addition, type of interaction with ChatGPT x time of measure (pre/post) was significant (F(1, 327) = 8.81, p = .003), with both types of interactions with ChatGPT showing an increase in trust after the interaction took place. Time of measure (pre/post) did not reach significant results (F(1,327) = 1.98, p = .160). Therefore, results shown by the model supported the alternative hypothesis: type of interaction with ChatGPT does have an effect on trust.

RQ2.1 (exploratory) examined whether the interaction number has an effect on trust. When examining solely, the interaction number did not reach significance (F(1, 328) = 0.142, p = .706), and the LMM did not actually explain the variance as it appeared to be driven by random effects (conditional R^2 = 0.607, p = < .001; marginal R^2 = .000, p = 1.000). Then we integrated the type of interaction with ChatGPT and the model proved to be significant (conditional R^2 = 0.708, p = < .001; marginal R^2 = .075, p = < .001). Results revealed a robust main effect of type of interaction with ChatGPT (F(1, 327) = 58.85, p = < .001), number alone became significant (F(1, 327) = 4.26, p = .040), and a significant interaction between type of interaction with ChatGPT and number of interactions (F(1, 327) = 5.20, p = .023), showed that the impact of interaction type also differed depending on whether it was encountered as first or second. While second interactions showed marginally higher estimated means for trust independently of their type, the changes differed by type of interaction. Task-oriented interactions scored slightly lower for trust when being the second interaction for participants, whereas reflexive interactions increased their score in trust when being the second. Results indicate that the number of interactions have a mild effect on trust, increasing slightly for the second interaction, even though it is important to note that the biggest effects were explained through type of interaction with ChatGPT. Figure 3.7 illustrates this behaviour.

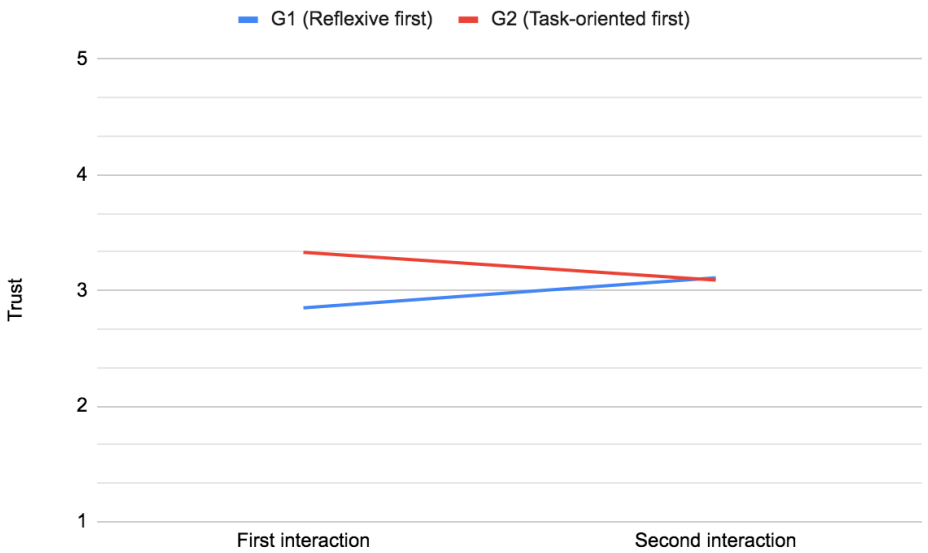

Figure 3.7. Trust responses for first and second interactions by group

RQ2.2 (exploratory) examined whether the interaction number (first/second) has an effect on predictability. Similarly, when examining solely, the LMM did not actually explain the variance (conditional R^2 = 0.370, p = < .001; marginal R^2 = .009, p = .130), even though the number of interactions reached significance (F(1, 328) = 6.79, p = .010). When integrating type of interaction with ChatGPT, the mixed model indicated significant main effects only for type of interaction with ChatGPT (F(1, 327) = 52.180, p < .001) on predictability. No significant evidence was found for number of interactions (F(1, 327) = 0.222, p = .638) nor interaction between number and type of interaction with ChatGPT (F(1, 327) = 0.358, p = .550). This suggests that type of interaction with ChatGPT influenced participants' perceived predictability but the interaction number was indifferent. Model fit showed statistical relevance as marginal (R^2 = .063) and conditional (R^2 = .473) were both significant (p = <.001).

RQ3 examined if the relationship between type of interaction with ChatGPT and trust was explained by predictability. We first verified that type of interaction had effects on predictability (F(1, 328) = 53.6, p = < .001) through a simple model (conditional R^2 = .465, marginal R^2 = .056, p = < .001) to validate path a from the mediation (see model from previous figure 3.5 for clarity on paths).

The LMM applied to analyse if predictability had a significant effect on trust (path b and c') showed to be relevant (conditional R^2 = .750, marginal R^2 = .153, p = < .001). Main effects of type of interaction with ChatGPT (F(1, 326) = 24.02, p = < .001) and predictability (F(1, 326) = 41.45, p = < .001) as well as the interaction between them (F(1, 326) = 6.43, p = .012), proved that predictability was relevant in this model and had significant effects on trust. Figure 3.8 illustrates this effect. Even more, through the integration of predictability into the model tested on *RQ1* the explained variance (R^2) of the model increases, showing that predictability plays an important role for better explaining the relationship between type of interaction and trust.

Then we ran a more complex LMM by integrating order (first and second interaction) into the model (conditional R^2 = .746, marginal R^2 = .187, p = < .001) that showed significant main effects of type of interaction with ChatGPT (F(1, 325) = 22.997, p = < .001), predictability (F(1, 325) = 35.145, p = < .001) and order (F(1, 109) = 6.925, p = .010). While the interaction between type of interaction with ChatGPT x predictability reached significant effects (F(1, 325) = 6.684, p = .010), it was not the case for the interaction of order x predictability (F(1, 325) = 0.127, p = .721). Seeing an increased variance by integrating predictability into the models, as well as order of interaction showing significant effects, we proceeded to test for mediation effects. The simple mediation analyses were conducted with type of interaction with ChatGPT as the independent variable, predictability[9] as the mediator, and trust as the dependent variable through separate analyses for the first interactions

[9] Note that in these mediation analyses, predictability specifically refers to the mathematical difference of prediction verification minus prediction scores.

across groups and, subsequently, their second interactions. Interestingly, results from mediation were different for each moment.

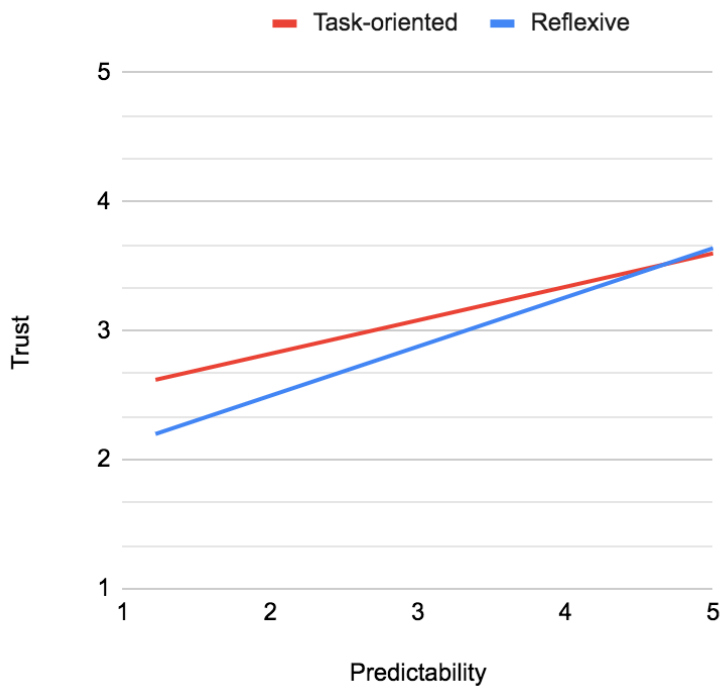

Figure 3.8. Interaction between predictability and type of interaction on trust

Path a from the first interaction did not reach significant results (β = .049, *t* = 0.449, *p* = .654) as the variance of predictability was poorly explained by type of interaction with ChatGPT (R^2 = .001); whereas path b resulted significant (β = .281, *t* = 4.409, *p* = < .001), indicating that higher scores were associated with higher trust. Path c was also non significant (β = .103, *t* = 1.404, *p* = .163). Direct effect (path c') showed no significant results (β = .1033, *SE* = 0.0736, *t* = 1.4041, *p* = .1632, 95% *CI* [-.0425, .2492]) as well as total effect (β = .1173, *SE* = 0.0795, *t* = 1.4750, *p* = .1431, 95% *CI* [-.0403, .2750]), and indirect effect (β = .0140, SE = 0.0315, 95% *CI* [-.0511, .0783], bootstrapped), meaning that for the first interaction there was no significant effect of type of interaction in trust through predictability. For the first interaction, the alternative hypothesis is not supported: predictability does not explain the effect of type of interaction in trust; the null hypothesis is accepted.

This was not the case for the second interaction where the mediation analysing data showed significant effects. Path a (R^2 = .104) showed significant effect of type of interaction in predictability (β = -.357, *t* = -3.538, *p* = < .001); as well as path b (β = .300, *t* = 4.913, *p* = < .001), while path c did not reach significant effects (β = -.107, *t* = -1.582, *p* = .117). The entire model (R^2 = .079) showed a

significant total effect (β = -.21437, *SE* = 0.0704, *t* = -3.0342, *p* = .003, 95% *CI* [-.3533, -.0741]) revealing that the reflexive interaction tends to elicit more trust than the task-oriented interaction. Finally, a significant indirect effect size (β = -.1079, *SE* = .0455, 95% *CI* [-.2106, -.0353], bootstrapped) and a non-significant direct effect (path c') obtained (β = -.1068, *SE* = 0.0675, *t* = -1.5817, *p* = .1167, 95% *CI* [-.2406, .0270]), suggest a full mediation through predictability. Table 3.9 shows results for the significant mediation of the second interaction. For the second interaction, the alternative hypothesis is supported: predictability explains the effect of type of interaction in trust.

Table 3.9. Mediation results for second interaction

Effect	Estimate (SE)	95% Confidence Interval		t	p	% Mediation
		Lower	Upper			
Indirect	-0.1069 (0.0455)	-0.2106	-0.0353		.022	50
Direct	-0.1068 (0.0675)	-0.2406	0.0270	-1.5817	.116	50
Total	-0.2137 (0.0704)	-0.3533	-0.0741	-3.0342	.003	100

Note. Type of interaction (IV), Predictability (Mediator), Trust (DV)

RQ4 (exploratory) inquired if prior knowledge moderated the relation between type of interaction with ChatGPT and predictability. The model failed to converge for a moderated mediation, therefore we proceeded to test type of interaction as independent variable, predictability as the dependent variable, and the moderators as covariates. The full LMM showed significance (conditional R^2 = .457, marginal R^2 = .206, p = < .001). All main effects of covariates were significant: AI familiarity (*F*(1, 94) = 12.05, *p* = < .001), frequency of use (*F*(1, 94) = 4.65, *p* = .034), and AI literacy (*F*(1, 94) = 13.49, *p* = < .001); as well as the main effect of type of interaction with ChatGPT (*F*(1, 293) = 54.42, *p* = < .001). When adding their interactions with type of interaction with ChatGPT to explore moderation, main effects of frequency of use showed no significant results on predictability (*F*(1, 94) = 1.60,*p* = .210), as well as its interaction with type of interaction with ChatGPT (*F*(1, 290) = 1.58 ,*p* = .210). We then ran a new model without frequency of use (conditional R^2 = .465, marginal R^2 = .198, p = < .001) where the main effect of AI literacy showed significant results on predictability (*F(*1,95) = 5.35, *p* = .023), nevertheless results did not reach significant levels for interaction of AI literacy x type of interaction with ChatGPT on predictability (*F*(1, 291) = 1.68, *p* = .195). The main effect of familiarity (*F(*1,95) = 14.29, *p* = < .001), as well as the interaction of familiarity x type of interaction with ChatGPT (*F(*1,291) = 6.54, *p* = .011) were significant.

As AI literacy was composed of three different factors, we tested them in a single independent model (conditional R^2 = .471, marginal R^2 = .174, p = < .001) besides type of interaction. Practical application was the only covariate that achieved significant main effects (*F(*1,107) = 5.862, *p* = .017). Interestingly, when testing their moderation effects on predictability in a new model (conditional R^2 =

.491, marginal R^2 = .188, p = < .001), the interaction between technical understanding x type of interaction with ChatGPT reached significant effects, suggesting that technical understanding moderates the relation between type of interaction with ChatGPT and predictability. Main effects of type of interaction with ChatGPT ($F(1,325) = 57.108$), practical application ($F(1,107) = 13.722$), as well as the interaction between practical application and type of interaction with ChatGPT ($F(1,325) = 11.406$ were significant (p = < .001) in this model. AI literacy as an entire construct did not moderate the relationship between type of interaction with ChatGPT and predictability; whereas the practical application factor played a role in moderating, as well as technical understanding.

RQ5 (exploratory) examined if attitudes towards AI moderated the relation between type of interaction and trust. Similarly to *RQ4*, a full moderated mediation failed to converge even after simplifying the model structure, likely because the model was too complex for the given sample size. We tested for a simple moderation between type of interaction and trust (conditional R^2 = .703, marginal R^2 = .284, p = < .001). Attitudes towards AI showed significant main effects in the model $F(1,109) = 45.797$, p = < .001); however, as moderator it did not reach significant results $F(1,327) = 0.037$, p = .847). This suggests that while attitudes towards AI might predict trust in AI, even when controlling for type of interaction, it does not seem to play a role as moderating this relation.

RQ6 examined whether the accuracy of prediction was related to emotional valence through correlation tests using Spearman coefficient. For task-oriented interactions, neither positive nor negative emotional valence correlated significantly with accuracy of prediction (p = > .05), while for reflexive interactions results showed a significant positive correlation with weak magnitude for positive emotional valence (r_s = 0.264, n = 110, p = .005), but not for negative emotional valence. Positive and negative valence in task-oriented and reflexive interactions were significantly and weakly associated with each other. The null hypothesis is accepted for task-oriented interactions, while the alternative hypothesis is supported for reflexive ones.

RQ7 tested whether accuracy of prediction was related to the epistemic emotions of boredom, curiosity, surprise, anxiety, excitement, confusion, and frustration. For task-oriented interactions, accuracy of prediction did not show significant correlations with any of the emotions. Table 3.10 shows detailed correlation results for all emotional responses on task-oriented interactions.

Accuracy of prediction in reflexive interactions showed weak significant correlations with boredom and excitement. Table 3.11 shows detailed correlation results for all emotional responses on reflexive interactions. The null hypothesis is accepted for task-oriented interactions, while the alternative hypothesis is supported for reflexive ones.

Table 3.10. Correlations on emotional responses and task-oriented interactions

Variable	1	2	3	4	5	6	7	8	9	10
1. Surprise	—									
2. Curiosity	.408**	—								
3. Excitement	.376**	.508**	—							
4. Confusion	.284**	.116	-.058	—						
5. Anxiety	.220*	.229*	.171	.424**	—					
6. Frustration	.009	.022	-.037	.479**	.370**	—				
7. Boredom	-.373**	-.305**	-.393**	.229*	-.008	.261**	—			
8. Positive Valence	.329**	.623**	.618**	-.232*	.012	-.224*	-.418**	—		
9. Negative Valence	-.003	-.093	-.127	.362**	.080	.454**	.351**	-.243**	—	
10. Prediction Accuracy	.041	-.043	.093	-.035	.135	-.151	-.027	.041	-.025	—

*Correlation is significant at the *p* <.05 level (2-tailed)

** Correlation is significant at the *p* <.01 level (2-tailed)

The sample size (N) for the correlations is 110

Table 3.11. Correlations on emotional responses and reflexive interactions

Variable	1	2	3	4	5	6	7	8	9	10
1. Surprise	—									
2. Curiosity	.401**	—								
3. Excitement	.525**	.472**	—							
4. Confusion	.095	-.005	-.065	—						
5. Anxiety	.232*	.140	.148	.333**	—					
6. Frustration	.007	-.018	-.033	.490**	.500**	—				
7. Boredom	-.156	-.337**	-.461**	.331**	.068	.441**	—			
8. Positive Valence	.386**	.452**	.587**	-.330**	.009	-.342**	-.435**	—		
9. Negative Valence	-.057	-.180	-.252**	.413**	.320**	.518**	.445**	-.353**	—	
10. Prediction Accuracy	.131	.130	.248**	-.164	-.074	-.159	-.334**	.264**	-.117	—

*Correlation is significant at the p<.05 level (2-tailed).

** Correlation is significant at the p<.01 level (2-tailed).

The sample size (N) for the correlations is 110.

3.1. Discussion *Main Study*

This study aimed to investigate if type of interaction with ChatGPT had an impact on trust, whether this effect was explained by predictability, and if prediction accuracy was related to emotional responses. To evaluate this, participants conversed with ChatGPT across two types of interaction using the prompts that were categorised by intent and assessed by complexity in *Study One*. For each interaction type, assessments on predictability and trust were made prior and after the interaction, whereas emotional responses were assessed only after each conversation. As empirical research on users' interactions with LLMs is still in its developing phase, a subset of exploratory questions was included to uncover preliminary effect patterns of AI prior knowledge, attitudes towards AI, as well

as numbers of interaction in trust and predictability.

In the results of *RQ1*, task-oriented interactions showed slightly higher scores in trust when compared to reflexive ones, thus corroborating the hypothesis that the type of interaction has an effect of trust with statistical significance. This might be explained by participants' familiarity and reliability in the task-oriented conversations, as they reported using ChatGPT for both personal and professional reasons. Previous research (Phang et al., 2025) has categorised conversations as personal and non-personal; however, a *personal* conversation might still be of a task-oriented type of interaction, as analysed qualitatively in our *Study One*. For example, participants might have had previous conversations with ChatGPT of a personal, or non-work related, context that were task-oriented (e.g. "Design a weekly meal plan for a busy individual, incorporating grocery shopping lists"); thus, being more confident about these types of interactions, which resulted in higher trust levels, even before the actual interaction during the experimental study. On the other side, reflexive conversations that include affective cues and deep introspection might not be common to all participants, but rather, more specific to users that converse intensively withChatGPT or companion chatbots such as Replika (Fang et al., 2025; Skjuve et al., 2021).

In addition, both types of interactions showed an increase in trust after the interaction took place, in line with research which proposes that interaction might drive trust (Glikson & Woolley, 2020). Potentially, the increase in trust might be explained by a plausibly competent response given by ChatGPT through coherent and human-like responses (Skjuve et al., 2023) combined with the active inference proposal of users' refinement of their mental models about limitations and capabilities of the system, as we will further review in *RQ3*.

With regards to changes in trust for first and second interaction (*RQ2.1*), our findings showed that the number of interactions was just significant if adding type of interaction to the model. Results indicated that the number of interactions have a statistically significant mild effect on trust and this is different by type of interaction, aligning with recent work on the interactive operation of cognitive and affective trust (Shang et al., 2024). Even though reflexive interactions showed lower levels of trust on a first conversation, for the second interaction they promoted an increase in trust, probably as participants pondered for the system's benevolence. Shang and colleagues (2024) studied cognitive and affective components of trust and suggest they may differ through the framing of conversations. They pointed out that when cognitive trust is high, increases in affective trust contribute little to overall trust, whereas under low cognitive trust, affective cues exert a stronger influence. This distinction may help explain the differences seen between task-oriented and reflexive interactions. Task-oriented interactions might have elicited cognitive forms of trust, while reflexive interactions, through their introspection topics, might have elicited affective trust. Shang and colleagues' results (2024) also suggest that LLM-generated dialogues can be designed to induce

varying levels of affective trust through differences in tone and empathy. In the reflexive interactions from our study, ChatGPT might have used more affective language, as previous research (Phang et al., 2025) have pointed that ChatGPT might apply mirroring or sycophantic behaviour to drive users' preference. In sum, these results reinforce the importance of conversational framing as a potential lever for shaping trust.

For *RQ2.2*, we did not find evidence that supported that interaction number had an effect on predictability. We only found evidence of type of interaction having an effect on predictability with reflexive interactions showing lower levels of predictability, reinforcing our previous interpretation in *RQ1* that users might have had more uncertainty, hence less predictability, about reflexive interactions with ChatGPT, regardless if they were encountered as first or second. As this exploratory RQ addressed a largely unexplored area of human–AI interaction, even these non-significant results might offer initial insights and promote further investigation.

Regarding *RQ3*, which tested if predictability explained the relationship between type of interaction and trust, an increase in the explained variance was achieved when integrating predictability into the models. This supports the theory that predictability has a key role in trust dynamics (Christov-Moore et al., 2024; Schoeller et al., 2021). On the mediation paths analyses, the LMM showed that type of interaction with ChatGPT significantly predicted both predictability and trust, and also that predictability strongly predicted trust. The increase observed in both predictability and trust specifically for reflexive interactions after only one exposure, and independently of order, suggests that participants experienced changes likely due to being presented with novel content. As trust depends on trial-and-error experience (Schoeller et al., 2021) but might increase after an initial direct encounter (Glikson & Woolley, 2020). Considering that participants were not novices in the use of ChatGPT nor AI, according to their responses on AI prior knowledge, it is plausible that they were already familiar with the capabilities of task-oriented conversations and therefore did not exhibit changes in trust and predictability before and after that interaction as notable as in the reflexive interactions.

When testing separately the first and second interactions in a simplified full mediation model[10], the results were not significant for the first interaction. Even though the mediation model for the first interaction showed that predictability had a significant effect on trust, the data did not support a full mediation. The significant statistical effects of path b, where predictability has a strong and positive effect on trust, are consistent with active inference framework where theory proposes (Christov-Moore et al., 2024) that we tend to trust other agents as long as we are able to integrate them in our predictive models. As our prediction scales accounted for value similarity, understanding,

[10] The simplified full mediation model collapsed predictability and trust into mathematical difference scores without considering repeated measures, nor random effects, as explained in the previous sections.

and expectation besides direct predictability, the composite variable assessed different aspects that, according to literature (Christov-Moore et al., 2024; Hommel & Colzato, 2015; Körber, 2019; Schoeller et al., 2021; Yokoi et al., 2021; Yokoi & Nakayachi, 2021), are key to increasing the predictability upon someone's action such as superposition of the own's model and the other agent's model. In juxtaposition, a full mediation hypothesis was supported considering the data of the second interaction. This difference in results might be explained due to previous significant results from the LMM with moment of interaction being significantly different. Results should be interpreted carefully as this discrepancy might also reflect the loss of variance and robustness from simplifying the design and testing the mediation in separate models for first and second interactions, as well as the limited sample size.

For the moderation research question on prior knowledge (*RQ4*), frequency of use did not reach significant effects on predictability, whereas AI literacy showed significant main effects. AI familiarity achieved significant main effects as well as moderation effects on predictability, in line with the theoretical suggestion of Körber (2019) of its moderation role. On AI literacy factors, practical application was a significant predictor and moderator of predictability, while technical understanding only showed moderation results. In a research from Cotter and Reisdorf (2020), their empirical findings show that understanding a system might not be solely explained by frequency of use but requires a breadth of varied experiences. The mere act of engaging with an AI system seems to moderate the refinement of users' mental models. As having used a system for different reasons might increase our rendering of its behaviour as more predictable (Cotter & Reisdorf, 2020; Schoeller et al., 2021), this might explain why familiarity achieved significant results. The result of AI literacy as a significant predictor of predictability is in line with the theoretical approach of active inference as knowledge might drive understanding and the modelling of a technology (Laupichler et al., 2023a; Schoeller et al., 2021). As this was an exploratory question, empirical examinations need to be further investigated. Still, our results show that different forms of prior knowledge might play a role to promote the refinement of users' mental models.

On *RQ5*, our results on attitudes towards AI as a predictor of AI trust is in line with prior research (Daly et al., 2025). However, results did not reach significance when testing for our exploratory question on moderation. Also, as ChatGPT or other AI systems are opaque and complex to understand, attitudes might not just be towards the system itself but the corporations who rule them (Schepman & Rodway, 2023) and further exploration in this matter is required. Exploratory *RQ4* and *RQ5* might serve to identify initial patterns through our empirical results that can inform subsequent hypotheses and investigation.

For our tests on emotional responses based on accuracy of prediction (*RQ6 & RQ7*), significant values were reached solely for reflexive interactions, highlighting once again that these types of

interactions with ChatGPT should be further explored. As we calculated prediction accuracy by subtraction of prediction verification minus prediction in absolute numbers, we should carefully interpret results as bigger numbers imply farther points between prediction and prediction verification, and thus, less accuracy. Prediction accuracy showed a significant, negative and weak correlation with boredom in reflexive interactions. This means that the narrower the difference between prediction and prediction verification, hence more accuracy of prediction, the more intensively boredom was reported by participants. This comes in line with theoretical frameworks which state that boredom might be experienced in exploitative situations rather than explorative ones (Darling, 2023; Yu et al., 2019). On the contrary, excitement showed a positive correlation with prediction accuracy: the more inaccurate the prediction, hence wider differences between prediction and prediction verification, the more intense the excitement reported. This suggests that participants encountered unexpected stimuli, and in line with previous research on adaptive responses (Ketonen et al., 2023), their mental models were probably attempting to adjust to the situation while showing opposite directions of boredom and excitement. As excitement implies high arousal and high pleasure (Russell, 1980), participants might have been alert, engaged in a more explorative situation, similar to what happens in play scenarios. In such scenarios, individuals deliberately seek to gravitate around the "sweet spot" and create situations of moderate complexity, aiming to resolve uncertainty (Andersen et al., 2023). Complementarily, a significant positive correlation between positive valence and accuracy of prediction in reflexive interactions suggest that as prediction is less accurate, participants experienced higher intensity of positive valence; counterintuitively with theory (e.g. Kiverstein & Miller, 2015; Schoeller et al., 2021) that proposes we tend to feel good when our prediction is accurate. This might be explained as participants being positively surprised due to ChatGPT's human-like responses, interactiveness, or performance (Skjuve et al., 2023). This novel experience potentially allowed for an update of participants' mental model, again resembling play scenarios, where a positive affect is elicited precisely because prediction errors are reduced faster than anticipated, aligning with the pleasurable quality associated with exploration and discovery (Andersen et al., 2023). Even though statistically significant, the magnitudes seen in emotional responses from reflexive interactions showed weak correlations. On the contrary, in the task-oriented interactions, prediction accuracy did not reach any significant correlations with any emotional response, probably because the experience did not elicit an emotional experience intensive enough to be captured by the assessment. Also, a ceiling effect might have happened since participants began with relatively neutral or positive predictability scores, when compared to reflexive ones, leaving little variance in predictability of expectation (prediction) vs. actual experience (prediction verification).

In sum, these findings are especially noteworthy since they potentially wrap up the dynamics

that participants might have experienced on reflexive emotions. We interpret that, even though familiar with ChatGPT, they were presented with relatively new content of interaction (reflexive topics), to which they perceived as less predictable when compared to task-oriented before the conversation took place. Despite the 69% of participants reporting that their use of ChatGPT includes personal reasons, these personal topics may include categories such as small talk, entertainment, personal-assistance, or even simple-reflexive topics, as all these might fit in the *personal reasons* for using it. Nonetheless, our reflexive prompts included cues of deeper introspection, were ranked in *Study One* as intermediate/somewhat difficult, and might not have been perceived as predictable. After the conversation, participants integrated this new experience into their mental model, and, as a plausible consequence, an increase in predictability was reported, explaining the increase in trust shown in the mixed model results, where emotional responses might have served as conscious cues of this process.

3.1.1. Limitations

To better mimic natural usage patterns, this study's design was intentionally open-ended, although this enhances the ecological validity of our findings, some of its innovations also present limitations. As users' accessed directly to their own ChatGPT account (either new or existent), the precise ChatGPT version used was not controlled for. Additionally, filter bubbles (Sohail et al., 2023), that is, when the algorithm selectively implies and shares the most relevant information for a specific user based on their past behaviours, might have contributed to possible different outcomes by participant's archetype and hinder the generalisation of experiences and effects. Nevertheless, this is not exclusively a concern for our study but for the research on GenAI as a whole in the era of hyperpersonalisation. To address this, conversations in the present study were guided using recommended prompting techniques (Schulhoff, 2025). Participants were given a prompt template including the objective (task-oriented/reflexive), output format (length, response style) and details (preferences, context) where they embedded their selected cue to promote their active involvement while seeking to keep comparable outcomes as much as possible. As participants could interact freely with ChatGPT, we could not verify nor control the prompts that were specifically sent, as well as for the time of interaction. While this could have been limited or monitored by developing an interface and connecting ChatGPT through an API, we were trying to measure the interactions with this system through the native interface, as the lay population would normally do to mirror real world interactions.

In addition, the researchers from this study had no access to the conversations, which is especially relevant for reflexive interactions about participant's personal life. Whilst people interact

more freely with chatbots as they do not feel judged (Skjuve et al., 2021), having no access to the conversations limits the possibility of deeper qualitative analysis, more nuanced results, or better control and transparency. This should be further explored in future research with ethics, care, and transparency. Informed consent ought to be clear and anonymization methods should be strictly applied. Even so, participants might interact differently knowing that a person will review their conversation and this should also be accounted for and monitored parallelly with quantitative tools.

Although participants were instructed to use and personalise one of the provided prompts, 15 of them did not comply with this protocol reporting *"other, not listed*" as their interaction prompt. As this 13,63% of the sample showed high quality responses, confirmed their interaction type, and finally were not outliers in main variables such as predictability or trust scales, we decided to retain these participants to maintain statistical power. We have, however, accounted for this as a limitation, given that the prompts employed by these individuals could not be verified. Parallely, 40.65% of participants interacted using the example prompts, even though these were not intended for use, as they represented slight outliers in terms of complexity and as our aim was for participants to engage more actively by personalising their template from the provided cues rather than copying examples. Nevertheless, we retained these cases in the analyses to preserve statistical power.

Another limitation was the intended method to validate the interaction type. The question's phrasing (*Did you interact with ChatGPT in a reflexive/task-oriented conversation?*) may have prompted a qualitative judgement of the interaction rather than a simple categorical confirmation. Their exclusion was required for this reason alone, highlighting the impact of the flawed validation method. Although this issue only affected three participants, this limitation is noteworthy as these individuals had submitted responses of excellent quality. Future studies could find better ways of confirming that users actually conversed in the indicated type of interaction to avoid losing valuable data. While this could be addressed by using an API integration, it means that researchers would have access to participants' conversations in order to verify their interaction types, which at the same time could potentially impact participants' self-disclosure and general responses by the sheer fact of knowing their conversations would be monitored.

Our assessment of trust was exclusively self-reported, meaning we did not include any observed behavioural trust measurement. This contrasts with other studies that monitor whether participants act on recommendations from an AI agent in collaborative settings. Behavioural trust is a key component of the trusting process, as it accounts for the perceived risks associated with an action. We did not monitor if the responses obtained from ChatGPT were subsequently integrated and used by the participants. Nevertheless, we attempted to address this behavioural aspect by allowing users to select from different prompts they felt would be more relevant to their current realities. Future studies could then monitor longitudinal interactions from participants with ChatGPT or other LLMs to

follow trust and predictability changes over time as well as further applications in the participants' life (e.g. application of recommendations given by the LLM, monitorisation rate of information received from the LLM, etc).

On the mediation data analyses, a major pitfall was the mismatch between the available tools and our study design. The 2 x 2 full design with repeated measures and random effects could not be tested integrally on the whole with the tools at hand. This limitation constrained our findings and even led to divergent results, since we had to run tests separately: using the LMM without mediation but with full moment to moment data, versus the mediation model that did not consider the study's design entirely, but separate mediations for interaction one and two.

3.1.2. Implications for Future Research

Participants came from a variety of different contexts, fulfilling our objective of recruiting a diverse sample, spread across a range of different ages, nationalities, and education levels; even though groups were homogeneous in background pertaining to their field of study, with social sciences representing half of the total sample, they were quite heterogeneous regarding occupation, with representation of students and workers from diverse areas. These demographic characteristics and their baseline measures in prior knowledge makes us consider that results can be generalised to the lay population. Even so, future research could test if results remain similar for other populations. Moreover, future studies could look for nuanced sociodemographic differences, as reports on AI adoption (Appel et al., 2025) show that usage, intent patterns, and adoption vary among such sociodemographic contexts.

Even though participants were neither tech experts nor from a technology context, they were mostly familiar with ChatGPT and similar systems. Also, they reported high scores in critical appraisal while scoring lower for technical understanding, in line with their non expert background. Although participants comprised different nationalities, countries of residence, academic, and professional backgrounds, it is possible that, because of the type of study, recruiting method, and time commitment in a digital task, participants were in general tech savvy. The latter could have influenced the results obtained for prior knowledge variables; also, participants' score in attitudes towards AI was high (Schepman & Rodway, 2023; Yilmaz et al., 2023). This possible trend of interest in tandem with generally positive attitudes towards AI needs to be considered when interpreting the results; future research could examine if results remain in a population that might be not as tech savvy as participants in this study. Moreover, as technological prior knowledge may be unequally distributed (Cotter & Reisdorf, 2020), future studies could explore in detail relationships between age, AI prior knowledge and attitudes towards AI, while ensuring that people on different levels of

adoption, varying degrees of interest in technology, and sociodemographic backgrounds participate in the study.

Regarding our quality assurance process, three participants passed attention checks and the bogus item but failed to confirm if they had had the specified type of interaction with ChatGPT. The wording of "*did you interact in a (task-oriented / reflexive) conversation?*" might have been confusing for participants, as evaluating if, for example, the conversation had really been the reflexion they were expecting vs. simply confirming that they engaged in the practical exercise of interacting with ChatGPT following the instructions given. Although three participants might not be a representative number, they were participants from excellent data quality that could not be considered in the analyses for this given reason.

As predictability seems to theoretically and empirically be a key element to explain trust dynamics; future research could continue to explore trust in AI interactions by considering predictability as the mediator that explains the effects of interaction in trust. In future studies a full multi level mediation will indeed be valuable to account for the particularities of the study design and ground on the theoretical framework of active inference. Even so, predictability as a self-assessed measure might be a simple approximation of the complex mechanisms of the predictive mental models of participants. Besides psychological aspects, future research integrating neuroscientific approaches would add further understanding of these dynamics.

Although structured AI literacy efforts can enhance users' predictability of AI, it appears that the practical application and familiarity through direct interaction with these systems may prove even more influential. Researchers could focus on developing a robust scale that accounts for AI prior knowledge as a whole (knowing, using, time of usage, literacy, appraisal, etc), as this could help future studies, industry, and society for a holistic assessment of individuals' prior AI knowledge. Future studies could expand on our exploratory questions through other scales, analysing the different constructs of prior knowledge all together, or examining if knowing and having used these systems before remain as key for predictability.

Regarding emotional responses correlation with prediction accuracy after interacting with ChatGPT, future studies could research if interactions with AI in participants' mother language systematically elicit more intense emotions when compared to second languages, as many of our participants interacted with ChatGPT in English but were nationals from a non-English speaking country. In addition, such research could explore cultural differences on emotional responses after interacting with AI in their mother language.

Overall, the effects for reflexive interactions were generally the most significant and, as proposed before, users might have experienced novel content in these type of interactions, potentially explaining why overall effects were stronger. While recurring users from Replika and

other types of chatbots might seek companionship or have engaged in introspective conversations, this might not be the case for the participants in this study, who reported to have engaged with ChatGPT before, even in personal conversations, probably because ChatGPT is a more generalistic LLM. To compare results, future studies could explore the dynamics of reflexive interactions with populations that already use chatbots for introspection. In addition, as studying reflexive interactions with generalistic LLMs is still developing, qualitative methods studies might offer valuable insights to better understand such dynamics.

Even so, trust dynamics for broader types of interactions, besides reflexive ones, should not be overestimated. As AI continues to develop connections and applications, such as RAG (recently used to philosophically showcase how these type of technology might extend one's mind; Smart et al., 2025) or agentic AI (with multi-agent collaborative autonomy for complex goals; Sapkota et al., 2025), there is a high possibility that we will keep interacting with AI, specially LLMs, as the user-oriented interactive system, for more and more intents.

4.1. Conclusion *Main Study*

This study aimed to examine the effect of interaction type (task-oriented/reflexive) on user trust in ChatGPT, exploring predictability as the core mediating mechanism to explain this effect and to test the correlation between prediction accuracy and emotional responses. The findings confirm that trust is dynamically constructed and significantly influenced by the nature of the interaction. Whilst task-oriented interactions yielded higher initial trust, consistent with user familiarity, reflexive interactions elicited a more pronounced increase in both trust and predictability post-interaction. This suggests that, when users were exposed to novel and introspective conversations, they actively refined their mental models of the AI's capabilities. The empirical support for predictability as a strong predictor of trust, and its mediation of the relationship between type of interaction and trust (the latter specifically in the second interaction), aligns robustly with the active inference framework. Results suggest that trust evolves as users integrate the system's behaviour into their own predictive models and that this integration takes place specially through interaction.

Users' mental model refinement through novel reflexive conversations was further evidenced by the correlation results on emotional responses with prediction accuracy exclusively within such reflexive interactions. Counter-intuitively, prediction inaccuracy in these conversations correlated with positive valence, accompanied with excitement, suggesting that users experienced a "positive surprise" of the system's performance on reflexive engagements, probably due to human-like responses. This, in turn, facilitated the updating of their mental models.

This research contributes to current knowledge by examining trust through predictability

dynamics, distinguishing effects elicited by interaction type, and exploring emotional responses through the lens of prediction accuracy. Despite its limitations, our *Main Study* underscores the critical role of conversational framing and predictability in the development of human-AI trust; it also opens new research directions for exploring how prior knowledge and beliefs might have an effect in predictability and trust. Future research should further dissect these dynamics and employ multi-level mediation models to validate the proposed mechanisms.

CHAPTER FOUR

What we (*infer to*) know

1.1. Final General Discussion

This thesis' research project sought to systematically explore human interaction with ChatGPT by first establishing a user-centred framework for prompt classification based on type of interaction, complexity, and user fit (*Study One*). Subsequently it tested the effects of these interactions in users' prediction, trust and emotion (*Main Study*). The sample of the *Main Study* shared similar sociodemographic characteristics with *Study One*, which enhances comparability and promotes continuity across studies.

The initial study identified common elements in prompts and constructed the 'task-oriented' and 'reflexive' categories; while users clearly recognised the former, the reflexive category lacked a strong consensus with less levels of confirmation rate in their categorisations as "reflexive" prompts. This might probably be explained due to participants evaluating the systems' intrinsic capacity to help reflect upon a situation. As AI systems lack own narratives and introspection, participants might have judged not the potential conversation that could be elicited through a reflexive prompt but the system's capacity to maintain that type of conversation. Also, *Study One*'s results suggest that reflexive prompts were perceived as intrinsically more complex when compared to task-oriented ones. In the last phase of this study, a smaller set of prompts was selected considering prompts from both types of interactions that shared a comparable level of complexity. The apparent lack of consensus in reflexive prompts might gain explanatory power when juxtaposed with the findings from the *Main Study*, where these same reflexive interactions elicited the most significant effects in user trust, their predictability on ChatGPT, and emotional responses. This suggests that while users may not have a pre-existing mental framework for reflexive interactions with ChatGPT (that are less familiar and therefore less predictable prompts), the novel experience, might have elicited a substantial update to their mental models of the system's capabilities as results led to the largest post-interaction gains in trust and predictability. At the same time, it might reflect participants' perception of a shared world understanding with ChatGPT, given the plausible human-like responses they received, even if such a shared world was, to some extent, *projected*. The sense of mutual understanding is so fundamental to human communication that we often take it for granted; when it is absent, as may occur with AI systems, trust and coordination break down. The fluency with which current generative AIs converse may create an illusion of *shared* world models; however, linguistic fluency does not necessarily entail a genuine common ground (Pezzulo et al., 2025).

Furthermore, while both studies confirmed that lay participants were generally familiar with ChatGPT, results from the *Main Study* showed that AI prior knowledge variables (AI familiarity, frequency of use, and AI literacy), were significant predictors of a participant's perceived predictability of the system's behaviour. Nevertheless, neither frequency of use nor the whole construct of AI literacy showed moderation effects between type of interaction with ChatGPT and predictability. This potentially refines our understanding of practical knowledge's importance, as AI familiarity reached both main and moderation effects, as well as practical application[11]. This means that participants who already knew, had used ChatGPT, and were aware of how AI systems could be applied in everyday life, reported higher predictability rates. By contrast, it is interesting to note that frequency of use did not statistically promote higher levels of predictability; probably due to previous research pointing that understandability of a system might not depend on frequency but on the breadth of experiences (Cotter & Reisdorf, 2020).

Notably, the emotional responses of the interactions were significant only within the reflexive interactions. The finding that prediction inaccuracy for reflexive prompts correlated with lower boredom, higher excitement, and higher positive valence, reinforces the idea that these interactions were more explorative and cognitively engaging for participants (Darling, 2023; Pekrun et al., 2017). These emotional responses seem to serve as conscious indicators of the users' mental model updating process, where unexpected AI behaviour in a personal context is met with heightened attention (Barrett, 2017; Hesp et al., 2021; Kiverstein & Miller, 2015; di Paolo et al., 2024).

Limitations of *Study One* were related to size: a small sample, and a relatively small number of prompts to be validated with lay users. As a consequence, this limited the number of final prompts with category agreement, restricting the breadth of insights and future choice. Limitations of the *Main Study* concern the challenge of balancing experimental control against ecological validity, including the inability to control the precise prompts used or access the conversation logs, as this was a trade-off made to mirror real-world usage. This highlights a critical methodological challenge for future human-AI interaction research. Future work could use LLM APIs to enhance repeatability, transparency, and robustness, sharing all model settings (parameters, weights, prompts, etc) via supplementary materials, adopting strategies proposed in *Study One* to enhance comparability; and considering prompting templates used in the *Main Study*, while also developing ethical protocols for analysing conversation content, as also proposed in the *Main Study*.

Some statistical models used (in the *Main Study*) were not fully able to capture the design's complexity, indicating a need for using more advanced analytical approaches. A limitation where

[11] One of the three subfactors of the AI Literacy construct

both studies converge is that some of the scales used were not validated for the specific population as they are (a) part of recently growing research efforts or (b) custom created for the studies. Although all scales were in English (their original language) and inclusion criteria on both studies included intermediate level, it is plausible that participants' responses may have been influenced by their cultural context. Future research could further develop and validate scales for other languages than English and different cultural contexts.

Despite the previously acknowledged limitations, this body of research presents several strong aspects and makes novel contributions to research, knowledge, and society. The strength of this research is amplified by its mixed-methods approach and sequential development: the framework established in the first study provided the foundation of the pilot-tested prompts used in the *Main Study* for analyses in a counterbalanced design. The primary innovation lies in the development and methodological validation of a novel prompt classification, distinguishing between "task-oriented" and '"reflexive" interactions, and for the first time, introducing perceived prompt complexity as a critical variable from a user's perspective (Velázquez et al., 2025). Also, it showed how reflexive interactions can significantly increase user trust by updating their predictability about the AI agent. This provides robust, theory-driven evidence for the active inference framework within HCI. By intentionally focusing on lay users and a widely accessible tool, the findings propose high potential application and contribute directly to a broader understanding of how humans can engage with AI, suggesting that even a brief exposure to novel, reflexive conversation can increase predictability and trust.

For society at large, this body of research moves the focus of interest beyond the purely transactional capabilities of LLMs to explore their effects on users regarding reflexion. In recent weeks, OpenAI (Chatterji et al., 2025) and Anthropic (Appel et al., 2025) published independent reports on how people use their LLM solutions (ChatGPT and Claude, respectively). Even though intents vary upon sociodemographic characteristics of the user, both studies show an increasing trend for personal advice, life guidance and related topics. OpenAI also reports that even though general usage of ChatGPT in both work-related and non-work-related conversations has been growing, the non-work-related conversations are growing significantly faster. Both corporations also report an increase in so-called *doing* or *directive* conversations, that is, users not only *asking* specific things but delegating more and more autonomous tasks to LLMs. As users increasingly engage with and delegate to LLMs and other AI systems for personal matters, such as seeking advice, facilitating self-reflection, and navigating complex interpersonal problems, critical societal questions arise, necessitating advanced research, a proactive focus on AI literacy, robust ethical guidelines, and transparent privacy controls.

Growth in usage of LLMs for reflexive conversations, and afterwards, the increase of trust in

both the LLM and the conversation itself, might become paradoxical on many different levels:

(1) Trust is created among different living systems that need to coordinate to survive as they face an uncertain world they cannot control. (2) We assume others share our basic core assumptions and understanding about the physical, social and moral world. (3) We might trust more someone who shares our priors (beliefs, values, views, narratives), which is to say, we tend to trust others whose mental model overlaps with ours. (4) Trust is a virtual minimizer of uncertainty: by having high certainty on our assumptions of others (trusting them), our mental model of the world expands through what they know, do and say to us. We integrate the information they give us about the world in the form of beliefs; not only about the world itself but even about ourselves. (5) Eventually, trust leads to empathy (Christov-Moore et al., 2024; Pezzulo et al., 2025). If we are supposed to trust living, predictable, benevolent, compatible agents that drive our empathy and expand our mental model of our shared world, how then might we be trusting non living agents that exhibit erratic behaviour (e.g. hallucinations[12]), answer to profit-driven corporate interests (e.g. ads and content based on our previous interactions; Meta, 2025a), with whom we do not seem to share narratives of our world? Our constant energetic efficiency seeking might be the answer (Christov-Moore et al., 2024; Schoeller et al., 2021). We will plausibly continue to expand our capabilities through technology; what makes AI different from previous technological advances might be that it starts behaving less like a tool and more like an agent. The way GenAI systems interact creates the *illusion* of a shared understanding of the world; yet, this illusion does not mean that our core assumptions are *actually* shared with AI. We might be overestimating an alignment where it might not exist (Pezzulo et al., 2025). Nonetheless, the possibility of expanding our mental model and agency through AI seems to only be accomplished by trusting these systems at *any* given level. How may our society find equilibrium?

It is a short yet highly complex question that we just might fail to answer. Nevertheless, we suggest that, as corporate efforts to improve AI development may not stop any time soon, we as academics, researchers, entrepreneurs, workers, individuals, should seek to increase the understanding of these systems, for ourselves and others, while demanding corporations our right for privacy. In European countries, for example, conversations' content on Meta (2025a; 2025b) will not be used for publicity purposes due to the General Data Protection Regulation of the European Union, contrary to other regions. Continuing current efforts and promoting new advances in AI and privacy regulation prove beneficial for individuals and societies.

Future research could explore the presence of AI in our daily lives; how are these systems being

[12] Hallucination is a generated response by an IA system that is either nonsensical, factually incorrect, or not grounded in reality (di Paolo et al., 2024).

embedded in our environment?; what are the effects of our direct interactions with AI, besides chat conversations?; while exploring its representations in daily common scenarios, including entertainment, how are these AI agents being depicted in movies or series? The characterisation of AI in mainstream media might serve as the primary source of information for the general public, and thus, one's perception of AI might be influenced by its representation in mass media (Nader et al., 2024), which in turn might guide attitudes, general discourse, and future actions (Cave & Dihal, 2020). Trust might be co-produced by social practices and social imaginaries[13], not only by system performance; for example, everyday embedding of AI systems in our daily lives might shape user expectations and the cultural preconditions for trust (Sheridan, 2019; Viaene et al., 2021). Even our emotions and other subjective experiences are shaped by social sharing and public interpretation beside private signals (Pezzulo et al., 2025).

Collaboration is built on trust, alignment and communication rather than blind delegation (Pezzulo et al., 2025). Ultimately, as individuals, we should both question and trust the responses given by AI just as we might do for ideas that "suddenly bubble up from our own biological unconscious" (Clark, 2025, p. 2), acknowledging this critical thinking may not be universal, but rather a function of educational and socioeconomic privilege. Crucially, as these systems become more integrated into users' personal lives, a parallel effort must be made to reinforce the value of human support networks and interpersonal engagement. This is particularly salient, given the potential for problematic usage patterns emerging, as shown by research (Fang et al., 2025).

While industry reports are just recently showing a growing trend towards users employing AI for these more personal, reflexive purposes, rigorous academic investigation is required to examine and interpret these phenomena. Such research would be invaluable for understanding the long-term effects of these systems on individual well-being and social dynamics, ensuring the discourse is led by an empirical and pro-societal approach. Building on this research, the next steps could focus on deeper explorations of the reflexive interactions, developing a holistic scale for AI prior knowledge, and exploring the impact of different modalities and languages on user engagement, trust, and emotional responses, specifically for reflexive interactions. Consequently, research, government, and society could, in what would be an almost utopical effort, coordinate programmes to provide tools and environments for individuals to integrate AI solutions in ways that benefit themselves and our community.

[13] Social imaginaries, term developed by Charles Taylor that refers to how ordinary people imagine their collective social life. Rather than abstract theories of elites, it pertains to the popular understanding, norms and practices that shape daily interactions, maintain and transform societies (O' Neill, 2016).

1.2. General Conclusion

Our world-revealing perception depends heavily on our prediction models, which are based on the very structure of our brain, our past experiences and our imagination. As our lives become increasingly intertwined with technology, we will also be shaped to the bone by these systems since our perception is built with the world. Organic-artificial perception will increasingly be made visible with AI's rapid development and adoption. Plausibly our future will be more and more conversational with these systems including wider topics, from professional non-personal goal oriented conversations, to personal introspective reflexive ones: we must both trust and question all these interactions. As our society and ways of communication are being transformed, studying trust, predictability, emotional responses and their influencers of attitudes, prior knowledge and usage is key.

This research contributes to addressing this challenge by empirically examining how different types of human–AI interactions shape trust and emotional responses among lay users. By analysing predictability as a key mechanism to explaining trust, and exploring the relationship between prediction accuracy and emotion, this work offers insights into how users build and calibrate their reliance on conversational AI. Through the development and assessment of prompts by their characteristics and the interaction types they may elicit, this body of research also provides methodological foundations for future studies aiming to investigate human–AI collaboration in real-world contexts.

Epilogue

Mudam-se os tempos, mudam-se as vontades,
muda-se o ser, muda-se a *confiança*;
todo o mundo é composto de mudança,
tomando sempre novas qualidades.
Continuamente vemos novidades,
diferentes em tudo da esperança;
do mal ficam as mágoas na lembrança,
e do bem — se algum houve —, as saüdades.
O tempo cobre o chão de verde manto,
que já coberto foi de neve fria,
e enfim converte em choro o doce canto.
E, afora este mudar-se cada dia,
outra mudança faz de mór espanto:
que não se muda já como soía.

[Times change, and so do wills,
our very being shifts, trust transforms;
the whole world is made of change,
forever taking on new qualities.
We constantly behold what's new,
all so different from our hopes;
from evil remain sorrows in memory,
and from good — if any there was —, longing.
Time covers the ground with a green mantle,
once wrapped in cold white snow,
and turns, at last, the sweet song into weeping.
And, beyond this daily turning of things,
there's yet another change, most strange of all:
that change itself no longer changes as it used to.]

de Camões (1980, p. 257)

References

Abdurahman, S., Salkhordeh Ziabari, A., Moore, A. K., Bartels, D. M., & Dehghani, M. (2025). A Primer for Evaluating Large Language Models in Social-Science Research. *Advances in Methods and Practices in Psychological Science*, *8*(2). https://doi.org/10.1177/25152459251325174

Ali, S., Akhlaq, F., Imran, A. S., Kastrati, Z., Daudpota, S. M., & Moosa, M. (2023). The enlightening role of explainable artificial intelligence in medical & healthcare domains: A systematic literature review. In *Computers in Biology and Medicine* (Vol. 166). https://doi.org/10.1016/j.compbiomed.2023.107555

Amazon. (2025, February 26). *Introducing Alexa+, the next generation of Alexa* [Press Release]. https://www.aboutamazon.com/news/devices/new-alexa-generative-artificial-intelligence

Andersen, M. M., Kiverstein, J., Miller, M., & Roepstorff, A. (2022). Play in Predictive Minds: A Cognitive Theory of Play. *Psychological Review*, *130*(2). https://doi.org/10.1037/rev0000369

Andries, V., & Robertson, J. (2023). Alexa doesn't have that many feelings: Children's understanding of AI through interactions with smart speakers in their homes. *Computers and Education: Artificial Intelligence*, *5*. https://doi.org/10.1016/j.caeai.2023.100176

Appel, R., McCrory, P., Tamkin, A., McCain, M., Neylon, T., Stern, M. (2025, September 15). *The Anthropic Economic Index report: Uneven geographic and enterprise AI adoption*. Anthropic. https://assets.anthropic.com/m/218c82b858610fac/original/Economic-Index.pdf

Argentzell, E., Bäckström, M., Lund, K., & Eklund, M. (2020). Exploring mediators of the recovery process over time among mental health service users, using a mixed model regression analysis based on cluster RCT data. *BMC Psychiatry*, *20*(1). https://doi.org/10.1186/s12888-020-02924-2

Bagozzi, R. P., Brady, M. K., & Huang, M. H. (2022). AI Service and Emotion. In *Journal of Service Research* (Vol. 25, Issue 4). https://doi.org/10.1177/10946705221118579

Barone, P., Corradi, G., & Gomila, A. (2019). Infants' performance in spontaneous-response false belief tasks: A review and meta-analysis. In *Infant Behavior and Development* (Vol. 57). https://doi.org/10.1016/j.infbeh.2019.101350

Barrett, L. F. (2006). Are Emotions Natural Kinds? *Perspectives on Psychological Science*, *1*(1). https://doi.org/10.1111/j.1745-6916.2006.00003.x

Barrett, L. F. (2017). The theory of constructed emotion: an active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, *12*(1). https://doi.org/10.1093/scan/nsw154

Barrett, L. F., & Russell, J. A. (1999). The structure of current affect: Controversies and emerging consensus. *Current Directions in Psychological Science*, *8*(1). https://doi.org/10.1111/1467-8721.00003

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, *59*. https://doi.org/10.1146/annurev.psych.59.103006.093639

Beni, M. D. (2025). Free will is real—I could not have believed otherwise. *Philosophical Psychology*, 1–25. https://doi.org/10.1080/09515089.2025.2551798

Bodonhelyi, A., Bozkir, E., Yang, S., Kasneci, E., & Kasneci, G. (2024). *User Intent Recognition and Satisfaction with Large Language Models: A User Study with ChatGPT*.

Boyd, R., Richerson, P. J., & Henrich, J. (2011). The cultural niche: Why social learning is essential for human adaptation. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(SUPPL. 2). https://doi.org/10.1073/pnas.1100290108

Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, *11*(4), 589–597. https://doi.org/10.1080/2159676X.2019.1628806

Braun, V., Clarke, V., & Hayfield, N. (2022). 'A starting point for your journey, not a map': Nikki Hayfield in conversation with Virginia Braun and Victoria Clarke about thematic analysis. *Qualitative Research in Psychology*, *19*(2), 424–445. https://doi.org/10.1080/14780887.2019.1670765

Briesemeister, B. B., Kuchinke, L., & Jacobs, A. M. (2012). Emotional valence: A bipolar continuum or two independent dimensions? *SAGE Open*, *2*(4). https://doi.org/10.1177/2158244012466558

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., … Amodei, D. (2020). *Language Models are Few-Shot Learners*. http://arxiv.org/abs/2005.14165

Brown, V. A. (2021). An Introduction to Linear Mixed-Effects Modeling in R. *Advances in Methods and Practices in Psychological Science*, *4*(1). https://doi.org/10.1177/2515245920960351

Bruineberg, J., Kiverstein, J., & Rietveld, E. (2018). The anticipating brain is not a scientist: the free-energy principle from an ecological-enactive perspective. *Synthese*, *195*(6). https://doi.org/10.1007/s11229-016-1239-1

Bryman, A. & Cramer, D. (1993). *Análise de dados em ciências sociais: introdução às técnicas utilizando o SPSS [Data Analyses in social sciences: introduction to techniques using SPSS]* (2nd ed.). Celta.

Brynjolfsson, E., Li, D., & Raymond, L. (2023). Generative Ai at Work. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4426942

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). *Sparks of Artificial General Intelligence: Early experiments with GPT-4*. https://arxiv.org/abs/2303.12712

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., Deane, G., Fleming, S. M., Frith, C., Ji, X., Kanai, R., Klein, C., Lindsay, G., Michel, M., Mudrik, L., Peters, M. A. K., Schwitzgebel, E., Simon, J., & VanRullen, R. (2023). *Consciousness in Artificial Intelligence: Insights from the Science of Consciousness*. http://arxiv.org/abs/2308.08708

de Camões, L. (1980). Lírica Completa II [Complete Lyric II]. Imprensa Nacional, Casa da Moeda.

Campilho, M. (2014). *Jóquei* [*Jockey*] (1st ed.). Tinta-da-china.

Castro, M., Lisboa, L., & Barcaui, A. (2024, August). Beyond the Code: Understanding Professional Users' Perspectives on AI Implementation. *Proceedings of the 10th World Congress on Electrical Engineering and Computer Systems and Science*. https://doi.org/10.11159/mhci24.111

Chatterji, A., Cunningham, T., Deming, D., Hitzig, Z., Ong, C., Shan, C. & Wadman, K. (2025,September 15). *How People Use ChatGPT*. OpenAI. https://www.nber.org/system/files/working_papers/w34255/w34255.pdf

Chandra, S., Shirish, A., & Srivastava, S. C. (2022). To Be or Not to Be …Human? Theorizing the Role of Human-Like Competencies in Conversational Artificial Intelligence Agents. *Journal of Management Information Systems*, *39*(4). https://doi.org/10.1080/07421222.2022.2127441

Christov-Moore, L., Jinich-Diamant, A., Bolis, D., Lehmann, K., Kanske, P., Durinski, T., Lynch, C., Iacoboni, M., Kaplan, J., Hesp, C., Schilbach, L., & Reggente, N. (2024). *Toward A Multiscale Account of Trust*. https://doi.org/10.31234/osf.io/p2gwa_v1

Chu, J., & Schulz, L. E. (2020). Play, Curiosity, and Cognition. *Annual Review of Developmental Psychology*, *2*(1). https://doi.org/10.1146/annurev-devpsych-070120-014806

Ciaunica, A., Seth, A., Limanowski, J., Hesp, C., & Friston, K. J. (2022). I overthink—Therefore I am not: An active inference account of altered sense of self and agency in depersonalisation disorder. *Consciousness and Cognition*, *101*, 103320. https://doi.org/10.1016/j.concog.2022.103320

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *The Behavioral and brain sciences*, *36*(3). https://doi.org/10.1017/S0140525X12000477

Clark, A. (2025). Extending Minds with Generative AI. *Nature Communications*, *16*(1), 4627. https://doi.org/10.1038/s41467-025-59906-9

Clark, A., & Chalmers, D. (1998). The Extended Mind. *Analysis*, *58*(1), 7–19. https://doi.org/10.1093/analys/58.1.7

Cave, S., & Dihal, K. (2020). The Whiteness of AI. *Philosophy and Technology*, *33*(4). https://doi.org/10.1007/s13347-020-00415-6

da Costa, L., Lanillos, P., Sajid, N., Friston, K., & Khan, S. (2022). How Active Inference Could Help Revolutionise Robotics. *Entropy*, *24*(3). https://doi.org/10.3390/e24030361

Cotter, K., & Reisdorf, B. C. (2020). Algorithmic Knowledge Gaps: A New Dimension of (Digital) Inequality. *International Journal of Communication*, *14*.

Daly, S. J., Wiewiora, A., & Hearn, G. (2025). Shifting attitudes and trust in AI: Influences on organizational AI adoption. *Technological Forecasting and Social Change*, *215*, 124108. https://doi.org/10.1016/j.techfore.2025.124108

Damásio, A. (2017). *A estranha ordem das coisas - A vida, os sentimentos e as culturas humanas* [*The strange order of things - life, feelings, and human cultures*] (1st ed.). Temas e Debates - Círculo de Leitores.

Darling, T. (2023). Synthesising boredom: a predictive processing approach. *Synthese*, *202*(5), 157. https://doi.org/10.1007/s11229-023-04380-3

Daronnat, S., Azzopardi, L., Halvey, M., & Dubiel, M. (2021). Inferring Trust From Users' Behaviours; Agents' Predictability Positively Affects Trust, Task Performance and Cognitive Load in Human-Agent Real-Time Collaboration. *Frontiers in Robotics and AI*, *8*. https://doi.org/10.3389/frobt.2021.642201

Duarte, R., Correia, F., Arriaga, P., & Paiva, A. (2023). AI Trust: Can Explainable AI Enhance Warranted Trust? *Human Behavior and Emerging Technologies*, *2023*. https://doi.org/10.1155/2023/4637678

di Paolo, L. D., White, B., Guénin-Carlut, A., Constant, A., & Clark, A. (2024). Active inference goes to school: the importance of active learning in the age of large language models. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *379*(1911). https://doi.org/10.1098/rstb.2023.0148

Dowling, M., & Lucey, B. (2023). ChatGPT for (Finance) research: The Bananarama Conjecture. *Finance Research Letters*, *53*. https://doi.org/10.1016/j.frl.2023.103662

du Toit, C. W. (2013). Dynamically remembered present: Virtual memory as a basis for the stories we live. *HTS Teologiese Studies / Theological Studies*, *69*(1). https://doi.org/10.4102/hts.v69i1.1937

Edelman, G. M., & Gally, J. A. (2001). Degeneracy and complexity in biological systems. *Proceedings of the National Academy of Sciences*, *98*(24), 13763–13768. https://doi.org/10.1073/pnas.231499798

Einstein, A., & Infeld, L. (1939). The Evolution of Physics. In *Cambridge University Press*. Cambridge University Press.

Elyoseph, Z., Hadar-Shoval, D., Asraf, K., & Lvovsky, M. (2023). ChatGPT outperforms humans in emotional awareness evaluations. *Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1199058

Fan, X., Oh, S., McNeese, M., Yen, J., Cuevas, H., Strater, L., & Endsley, M. R. (2008). The influence of agent reliability on trust in human-agent collaboration. *ACM International Conference Proceeding Series*, *369*. https://doi.org/10.1145/1473018.1473028

Fang, C. M., Liu, A. R., Danry, V., Lee, E., Chan, S. W. T., Pataranutaporn, P., Maes, P., Phang, J., Lampe, M., Ahmad, L., & Agarwal, S. (2025). *How AI and Human Behaviors Shape Psychosocial Effects of Chatbot Use: A Longitudinal Controlled Study*. https://arxiv.org/abs/2503.17473

Feldman, M. J., Siegel, E., Barrett, L. F., Quigley, K. S., & Wormwood, J. B. (2022). Affect and Social Judgment: The Roles of Physiological Reactivity and Interoceptive Sensitivity. *Affective Science*, *3*(2). https://doi.org/10.1007/s42761-022-00114-9

Fiore, S. M., Wiltshire, T. J., Lobato, E. J. C., Jentsch, F. G., Huang, W. H., & Axelrod, B. (2013). Toward understanding social cues and signals in human-robot interaction: Effects of robot gaze and proxemic behavior. *Frontiers in Psychology*. https://doi.org/10.3389/fpsyg.2013.00859

Gallucci, M. (2019). *GAMLj: General analyses for linear models*. [jamovi module]. Retrieved from https://gamlj.github.io/.

Gallucci, M. (2020). *Model goodness of fit in GAMLj*. Retrieved from https://gamlj.github.io/details_goodness.html

Gasper, K., & Clore, G. L. (2002). Attending to the big picture: Mood and global versus local processing of visual information. *Psychological Science*, *13*(1). https://doi.org/10.1111/1467-9280.00406

Geng, S., Liu, S., Fu, Z., Ge, Y., & Zhang, Y. (2022). Recommendation as Language Processing (RLP): A Unified Pretrain, Personalized Prompt & Predict Paradigm. *RecSys 2022 - Proceedings of the 16th ACM Conference on Recommender Systems*. https://doi.org/10.1145/3523227.3546767

Giuste, F., Shi, W., Zhu, Y., Naren, T., Isgut, M., Sha, Y., Tong, L., Gupte, M., & Wang, M. D. (2023). Explainable Artificial Intelligence Methods in Combating Pandemics: A Systematic Review. *IEEE Reviews in Biomedical Engineering*, *16*. https://doi.org/10.1109/RBME.2022.3185953

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2). https://doi.org/10.5465/annals.2018.0057

Google Cloud (2025). *Generative AI Leader* [Online Course]. https://www.skills.google/paths/1951

Grassini, S. (2023). Development and validation of the AI attitude scale (AIAS-4): a brief measure of general attitude toward artificial intelligence. *Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1191628

Guingrich, R. E., & Graziano, M. S. A. (2023). *Chatbots as social companions: How people perceive consciousness, human likeness, and social health benefits in machines*. https://doi.org/10.1093/9780198945215.003.0011

Gulati, S., Sousa, S., & Lamas, D. (2019). Design, development and evaluation of a human-computer trust scale. *Behaviour and Information Technology*, *38*(10). https://doi.org/10.1080/0144929X.2019.1656779

Gulati, S., Sousa, S., & Lamas, D. (2018). Modelling trust in human-like technologies. *ACM International Conference Proceeding Series*. https://doi.org/10.1145/3297121.3297124

Habibi, A., Muhaimin, M., Danibao, B. K., Wibowo, Y. G., Wahyuni, S., & Octavia, A. (2023). ChatGPT in higher education learning: Acceptance and use. *Computers and Education: Artificial Intelligence*, *5*. https://doi.org/10.1016/j.caeai.2023.100190

Hanum Siregar, F., Hasmayni, B., & Lubis, A. H. (2023). The Analysis of Chat GPT Usage Impact on Learning Motivation among Scout Students. *International Journal of Research and Review*, *10*(7). https://doi.org/10.52403/ijrr.20230774

Harari, Y. N. (2025, June 30). *AI and the paradox of trust*. [Video]. Youtube. https://www.youtube.com /watch?v=8GaW36EfidI

Hargrove, T. (Host) & Miller. M. (2022). Mark Miller on Predictive Processing. [Audio podcast episode]. In *The better movement podcast*. https://toddhargrove.substack.com/p/mark-miller-on-predictive-processing#details

Hayes. (2022). Introduction to Mediation, Moderation, and Conditional Process Analysis - Model Numbers. In *the Guilford Press* (Vol. 46, Issue 3).

Hesp, C., Smith, R., Parr, T., Allen, M., Friston, K. J., & Ramstead, M. J. D. (2021). Deeply felt affect: The emergence of valence in deep active inference. In *Neural Computation* (Vol. 33, Issue 2). https://doi.org/10.1162/neco_a_01341

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3). https://doi.org/10.1177/0018720814547570

Hommel, B., & Colzato, L. S. (2015). Interpersonal trust: an event-based account. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01399

Hutchinson, J. B., & Barrett, L. F. (2019). The Power of Predictions: An Emerging Paradigm for Psychological Research. *Current Directions in Psychological Science*, *28*(3), 280–291. https://doi.org/10.1177/0963721419831992

Ibrahim, L., Akbulut, C., Elasmar, R., Rastogi, C., Kahng, M., Morris, M. R., McKee, K. R., Rieser, V., Shanahan, M., & Weidinger, L. (2025). *Multi-turn Evaluation of Anthropomorphic Behaviours in Large Language Models*. http://arxiv.org/abs/2502.07077

Jermutus, E., Kneale, D., Thomas, J., & Michie, S. (2022). Influences on User Trust in Healthcare Artificial Intelligence: A Systematic Review. *Wellcome Open Research*, *7*. https://doi.org/10.12688/wellcomeopenres.17550.1

Johnsen, M. (2025). *Developing AI Applications With Large Language Models*.

Johnston, V. S. (2003). The origin and function of pleasure. In *Cognition and Emotion*, 17(2). https://doi.org/10.1080/02699930302290

Jung, J., Lee, H., Jung, H., & Kim, H. (2023). Essential properties and explanation effectiveness of explainable artificial intelligence in healthcare: A systematic review. In *Heliyon*, 9(5). https://doi.org/10.1016/j.heliyon.2023.e16110

Kalantzis, M., & Cope, B. (2025). Literacy in the Time of Artificial Intelligence. *Reading Research Quarterly*, *60*(1). https://doi.org/10.1002/rrq.591

Kelly, S., Kaye, S. A., & Oviedo-Trespalacios, O. (2023). What factors contribute to the acceptance of artificial intelligence? A systematic review. *Telematics and Informatics*, *77*. https://doi.org/10.1016/j.tele.2022.101925

Ketonen, E. E., Salonen, V., Lonka, K., & Salmela-Aro, K. (2023). Can you feel the excitement? Physiological correlates of students' self-reported emotions. *British Journal of Educational Psychology*, *93*(S1), 113–129. https://doi.org/10.1111/bjep.12534

Kiverstein, J., & Miller, M. (2015). The embodied brain: Towards a radical embodied cognitive neuroscience. *Frontiers in Human Neuroscience*, Sec. Cognitive Neuroscience, *9*. https://doi.org/10.3389/fnhum.2015.00237

Körber, M. (2019). Theoretical considerations and development of a questionnaire to measure trust in automation. *Advances in Intelligent Systems and Computing*, *823*. https://doi.org/10.1007/978-3-319-96074-6_2

Kuntz, D., & Wilson, A. K. (2022). Machine learning, artificial intelligence, and chemistry: How smart algorithms are reshaping simulation and the laboratory. *Pure and Applied Chemistry*, *94*(8), 1019–1054. https://doi.org/10.1515/pac-2022-0202

Kwa, T., West, B., Becker, J., Deng, A., Garcia, K., Hasin, M., Jawhar, S., Kinniment, M., Rush, N., von Arx, S., Bloom, R., Broadley, T., Du, H., Goodrich, B., Jurkovic, N., Miles, L. H., Nix, S., Lin, T., Parikh, N., … Chan, L. (2025). *Measuring AI Ability to Complete Long Tasks*. http://arxiv.org/abs/2503.14499

Laupichler, M. C., Aster, A., Haverkamp, N., & Raupach, T. (2023a). Development of the "Scale for the assessment of non-experts' AI literacy" – An exploratory factor analysis. *Computers in Human Behavior Reports*, *12*. https://doi.org/10.1016/j.chbr.2023.100338

Laupichler, M. C., Aster, A., & Raupach, T. (2023b). Delphi study for the development and preliminary validation of an item set for the assessment of non-experts' AI literacy. *Computers and Education: Artificial Intelligence*, *4*. https://doi.org/10.1016/j.caeai.2023.100126

Lee, J. J., Knox, W. B., Wormwood, J. B., Breazeal, C., & DeSteno, D. (2013). Computationally modeling interpersonal trust. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00893

Leikas, J., Johri, A., Latvanen, M., Wessberg, N., & Hahto, A. (2022). Governing Ethical AI Transformation: A Case Study of AuroraAI. In *Frontiers in Artificial Intelligence*, 5. https://doi.org/10.3389/frai.2022.836557

Liu, J., Liu, C., Zhou, P., Lv, R., Zhou, K., & Zhang, Y. (2023). *Is ChatGPT a Good Recommender? A Preliminary Study*. https://arxiv.org/abs/2304.10149

Long, D., & Magerko, B. (2020). What is AI Literacy? Competencies and Design Considerations | Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3313831.3376727

Lüdecke, Ben-Shachar, Patil & Makowski (2020). *Extracting, computing and exploring the parameters of statistical models using R*. CRAN. Retrieved from https://github.com/easystats/parameters

Malti, T., Averdijk, M., Zuffianò, A., Ribeaud, D., Betts, L. R., Rotenberg, K. J., & Eisner, M. P. (2016). Children's trust and the development of prosocial behavior. *International Journal of Behavioral Development*, *40*(3). https://doi.org/10.1177/0165025415584628

Mazzaglia, P., Verbelen, T., Çatal, O., & Dhoedt, B. (2022). The Free Energy Principle for Perception and Action: A Deep Learning Perspective. *Entropy*, 24(2). https://doi.org/10.3390/e24020301

McEwen, B. S., & Wingfield, J. C. (2010). What is in a name? Integrating homeostasis, allostasis and stress. *Hormones and Behavior*, *57*(2), 105–111. https://doi.org/10.1016/j.yhbeh.2009.09.011

Meta. (2025a, October 1). *Improving your recommendations on our apps with AI at Meta*. [Press Release]. https://about.fb.com/news/2025/10/improving-your-recommendations-apps-ai-meta/

Meta. (2025b, June 16). *Privacy Policy*. https://www.facebook.com/privacy/policy/

Miller, M. (2018). *The Entangled Predictive Brain: Emotion, Prediction and Embodied Cognition*.

Miller, M., & Clark, A. (2018). Happily entangled: prediction, emotion, and the embodied mind. *Synthese*, *195*(6), 2559–2575. https://doi.org/10.1007/s11229-017-1399-7

Mondal, S., Bappon, S. D., & Roy, C. K. (2024). *Enhancing User Interaction in ChatGPT: Characterizing and Consolidating Multiple Prompts for Issue Resolution*. https://arxiv.org/abs/2402.04568

Moore, J. W. (2016). What is the sense of agency and why does it matter? In *Frontiers in Psychology*, 7. https://doi.org/10.3389/fpsyg.2016.01272

Muis, K. R., Pekrun, R., Sinatra, G. M., Azevedo, R., Trevors, G., Meier, E., & Heddy, B. C. (2015). The curious case of climate change: Testing a theoretical model of epistemic beliefs, epistemic emotions, and complex learning. *Learning and Instruction*, *39*. https://doi.org/10.1016/j.learninstruc.2015.06.003

Murray, A., Rhymer, J., & Sirmon, D. G. (2021). Humans and technology: Forms of conjoined agency in organizations. *Academy of Management Review*, *46*(3). https://doi.org/10.5465/amr.2019.0186

Nader, K., Toprac, P., Scott, S., & Baker, S. (2024). Public understanding of artificial intelligence

through entertainment media. *AI and Society*, *39*(2). https://doi.org/10.1007/s00146-022-01427-w

Navarro, A., Luhrmann, T. M., Cassaniti, J., Crane, E. S., Turner, F., Stanford University School of Humanities and Sciences, & Stanford University Department of Anthropology. (2025). *Designing cyborg consciousness : how smartphones transform perception, embodiment, and cognition* [Dissertation]. Stanford University. https://purl.stanford.edu/hx841zf2505

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2023). *A Comprehensive Overview of Large Language Models*. http://arxiv.org/abs/2307.06435

Nazar, M., Alam, M. M., Yafi, E., & Su'Ud, M. M. (2021). A Systematic Review of Human-Computer Interaction and Explainable Artificial Intelligence in Healthcare with Artificial Intelligence Techniques. In *IEEE Access*, 9. https://doi.org/10.1109/ACCESS.2021.3127881

Oliveira, A. (2025). *A Inteligência Artificial Generativa [Generative Artificial Intelligence]*. Fundação Francisco Manuel dos Santos.

O'Neill, J. (2016). Social Imaginaries: An Overview. In *Encyclopedia of Educational Philosophy and Theory*. Springer Singapore. https://doi.org/10.1007/978-981-287-532-7_379-1

OpenAI. (2024). ChatGPT(GPT-3.5) [Large language model]. https://chatgpt.com/

OpenAI. (2024, May 13). *Spring Update. Introducing GPT-4o and making more capabilities available for free in ChatGPT* [Press Release]. https://openai.com/index/spring-update/

OpenAI. (2025). ChatGPT (GPT-4.o) [Large language model]. https://chatgpt.com/

OpenAI. (2025, August 7). *Introducing GPT-5* [Press Release]. https://openai.com/index/introducing-gpt-5/

Orrù, G., Piarulli, A., Conversano, C., & Gemignani, A. (2023). Human-like problem-solving abilities in large language models using ChatGPT. *Frontiers in Artificial Intelligence*, *6*. https://doi.org/10.3389/frai.2023.1199350

Park, J., & Banaji, M. R. (2000). Mood and heuristics: The influence of happy and sad states on sensitivity and bias in stereotyping. *Journal of Personality and Social Psychology*, *78*(6). https://doi.org/10.1037/0022-3514.78.6.1005

Parr, T., Pezzulo, G., & Friston, K. J. (2022). *Active Inference: The Free Energy Principle in Mind, Brain, and Behavior*. The MIT Press. https://doi.org/10.7551/mitpress/12441.001.0001

Parvizi, J. (2009). Corticocentric myopia: old bias in new cognitive sciences. *Trends in Cognitive Sciences*, *13*(8). https://doi.org/10.1016/j.tics.2009.04.008

Paulus, M. P., Feinstein, J. S., & Khalsa, S. S. (2019). An Active Inference Approach to Interoceptive Psychopathology. In *Annual Review of Clinical Psychology*, 15. https://doi.org/10.1146/annurev-clinpsy-050718-095617

Pekrun, R., Vogl, E., Muis, K. R., & Sinatra, G. M. (2017). Measuring emotions during epistemic activities: the Epistemically-Related Emotion Scales. *Cognition and Emotion*, *31*(6). https://doi.org/10.1080/02699931.2016.1204989

Pelau, C., Volkmann, C., Barbul, M., & Bojescu, I. (2023). The Role of Attachment in Improving Consumer-AI Interactions. *Proceedings of the International Conference on Business Excellence*, *17*(1). https://doi.org/10.2478/picbe-2023-0097

Pessoa, F. (1968). *Textos Filosóficos: Vol. II* [Philosophical Texts: Vol II] (A. Pina Coelho ed.). Ática.

Pezzulo, G., Parr, T., Cisek, P., Clark, A., & Friston, K. (2024). Generating meaning: active inference and the scope and limits of passive AI. In *Trends in Cognitive Sciences*, 28,(2). https://doi.org/10.1016/j.tics.2023.10.002

Pezzulo, G., Parr, T., & Friston, K. J. (2025). Shared worlds, shared minds. *EMBO Reports*, *26*(17). https://doi.org/10.1038/s44319-025-00549-8

Pham, K. T., Nabizadeh, A., & Selek, S. (2022). Artificial Intelligence and Chatbots in Psychiatry. In *Psychiatric Quarterly*, 93(1). https://doi.org/10.1007/s11126-022-09973-8

Phang, J., Lampe, M., Ahmad, L., Agarwal, S., Fang, C. M., Liu, A. R., Danry, V., Lee, E., Chan, S. W. T., Pataranutaporn, P., & Maes, P. (2025). *Investigating Affective Use and Emotional Well-being on ChatGPT*. https://arxiv.org/abs/2504.03888

Porsdam Mann, S., Vazirani, A. A., Aboy, M., Earp, B. D., Minssen, T., Cohen, I. G., & Savulescu, J. (2024). Guidelines for ethical use and acknowledgement of large language models in academic writing. *Nature Machine Intelligence*, *6*(11). https://doi.org/10.1038/s42256-024-00922-7

Purington, A., Taft, J. G., Sannon, S., Bazarova, N. N., & Taylor, S. H. (2017). "Alexa is my new BFF": Social roles, user satisfaction, and personification of the Amazon Echo. *Conference on Human Factors in Computing Systems Proceedings*. https://doi.org/10.1145/3027063.3053246

Rainie, L. (2025). *Close encounters of the AI kind: The increasingly human-like way people are engaging with language models*. https://imaginingthedigitalfuture.org/wp-content/uploads/2025/03/ITDF-LLM-User-Report-3-12-25.pdf

R Core Team (2024). *R: A Language and environment for statistical computing*. (Version 4.4) [Computer software]. Retrieved from https://cran.r-project.org. (R packages retrieved from CRAN snapshot 2024-08-07).

Rempel, J. K., Holmes, J. G., & Zanna, M. P. (1985). Trust in Close Relationships. *Journal of Personality and Social Psychology*, *49*(1). https://doi.org/10.1037/0022-3514.49.1.95

Rigotti, M., Barak, O., Warden, M. R., Wang, X.-J., Daw, N. D., Miller, E. K., & Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. *Nature*, *497*(7451). https://doi.org/10.1038/nature12160

Roll, R. (Host) & Harari, Y. N. (2024, October 28). Rage against the machines: Yuval Noah Harari on

surviving AI, the history of information, and the future of humanity. [Audio podcast episode]. In *The Rich Roll Podcast*. https://spotify.link/tsotepv9tXb

Romero, L. M., Dickens, M. J., & Cyr, N. E. (2009). The reactive scope model — A new model integrating homeostasis, allostasis, and stress. *Hormones and Behavior*, *55*(3). https://doi.org/10.1016/j.yhbeh.2008.12.009

Ronge, R., Maier, M., & Rathgeber, B. (2025). Towards a Definition of Generative Artificial Intelligence. *Philosophy & Technology*, *38*(1), 31. https://doi.org/10.1007/s13347-025-00863-y

Rose, S. P. R. . (2003). *The making of memory: from molecules to mind*. Vintage.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6). https://doi.org/10.1037/h0077714

Saha, K., Jain, Y., Liu, C., Kaliappan, S., & Karkar, R. (2025). AI vs. Humans for Online Support: Comparing the Language of Responses from LLMs and Online Communities of Alzheimer's Disease. *ACM Transactions on Computing for Healthcare*. https://doi.org/10.1145/3709366

Samani, C., Atif, A., & Musial-Gabrys, K. (2022). Using Emotional Learning Analytics to Improve Students' Engagement in Online Learning. *ASCILITE Publications*. https://doi.org/10.14742/apubs.2022.129

Sanh, V., Webson, A., Raffel, C., Bach, S. H., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., le Scao, T., Raja, A., Dey, M., Bari, M. S., Xu, C., Thakker, U., Sharma, S., Szczechla, E., Kim, T., Chhablani, G., Nayak, N. v., … Rush, A. M. (2022). Multitask prompted training enables zero-shot task generalization. *ICLR 2022 - 10th International Conference on Learning Representations*. https://arxiv.org/abs/2110.08207

Santu, S. K. K., & Feng, D. (2023). *TELeR: A General Taxonomy of LLM Prompts for Benchmarking Complex Tasks*. https://arxiv.org/abs/2305.11430

Sapkota, R., Roumeliotis, K. I., & Karkee, M. (2025). *AI Agents vs. Agentic AI: A Conceptual Taxonomy, Applications and Challenges*. https://doi.org/10.1016/j.inffus.2025.103599

Schepman, A., & Rodway, P. (2023). The General Attitudes towards Artificial Intelligence Scale (GAAIS): Confirmatory Validation and Associations with Personality, Corporate Distrust, and General Trust. *International Journal of Human-Computer Interaction*, *39*(13). https://doi.org/10.1080/10447318.2022.2085400

Schoeller, F., Miller, M., Salomon, R., & Friston, K. J. (2021). Trust as Extended Control: Human-Machine Interactions as Active Inference. *Frontiers in Systems Neuroscience*, *15*. https://doi.org/10.3389/fnsys.2021.669810

Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., Li, Y., Gupta, A., Han, H., Schulhoff, S., Dulepet, P. S., Vidyadhara, S., Ki, D., Agrawal, S., Pham, C., Kroiz, G., Li, F., Tao, H., Srivastava, A., …

Resnik, P. (2025). The Prompt Report: A Systematic Survey of Prompt Engineering Techniques. https://arxiv.org/abs/2406.06608

Shang, R., Hsieh, G., & Shah, C. (2024). *Trusting Your AI Agent Emotionally and Cognitively: Development and Validation of a Semantic Differential Scale for AI Trust*. https://arxiv.org/abs/2408.05354

Sheridan, T. B. (1988). Trustworthiness of Command and Control Systems. *IFAC Proceedings Volumes*, *21*(5). https://doi.org/10.1016/s1474-6670(17)53945-2

Sheridan, T. B. (2019). Individual differences in attributes of trust in automation: Measurement and application to system design. In *Frontiers in Psychology* (Vol. 10, Issue MAY). https://doi.org/10.3389/fpsyg.2019.01117

Shumanov, M., & Johnson, L. (2021). Making conversations with chatbots more personalized. *Computers in Human Behavior*, *117*. https://doi.org/10.1016/j.chb.2020.106627

Silvia, P. J. (2010). Confusion and interest: The role of knowledge emotions in aesthetic experience. *Psychology of Aesthetics, Creativity, and the Arts*, *4*(2). https://doi.org/10.1037/a0017081

Skjuve, M., Følstad, A., & Brandtzaeg, P. B. (2023). The User Experience of ChatGPT: Findings from a Questionnaire Study of Early Users. *Proceedings of the 5th International Conference on Conversational User Interfaces, CUI 2023*. https://doi.org/10.1145/3571884.3597144

Skjuve, M., Følstad, A., Fostervold, K. I., & Brandtzaeg, P. B. (2021). My Chatbot Companion - a Study of Human-Chatbot Relationships. *International Journal of Human Computer Studies*, *149*. https://doi.org/10.1016/j.ijhcs.2021.102601

Smart, P., Clowes, R., & Clark, A. (2025). ChatGPT, extended: large language models and the extended mind. *Synthese*, *205*(6). https://doi.org/10.1007/s11229-025-05046-y

Sobel, D. M., & Kushnir, T. (2013). Knowledge matters: How children evaluate the reliability of testimony as a process of rational inference. *Psychological Review*, *120*(4). https://doi.org/10.1037/a0034191

Sohail, S. S., Madsen, D. Ø., Himeur, Y., & Ashraf, M. (2023). Using ChatGPT to navigate ambivalent and contradictory research findings on artificial intelligence. *Frontiers in Artificial Intelligence*, *6*. https://doi.org/10.3389/frai.2023.1195797

Steele, J. L. (2023). To GPT or not GPT? Empowering our students to learn with AI. *Computers and Education: Artificial Intelligence*, *5*. https://doi.org/10.1016/j.caeai.2023.100160

Subramonyam, H., Pea, R., Pondoc, C. L., Agrawala, M., & Seifert, C. (2024). Bridging the Gulf of Envisioning: Cognitive Challenges in Prompt Based Interactions with LLMs. *Conference on Human Factors in Computing Systems Proceedings*. https://doi.org/10.1145/3613904.3642754

Szczesniak, M., Colaeo, M., & Rondón, G. (2012). Development of interpersonal trust among children and adolescents. *Polish Psychological Bulletin*, *43*(1). https://doi.org/10.2478/v10059-012-0006-5

Tarchi, C., Braasch, J., Fallaci, A. P., & Guidi, E. (2025). Thinking dispositions and epistemic cognition: Combined influences when reading to learn. *Learning and Individual Differences*, *123*. https://doi.org/10.1016/j.lindif.2025.102783

The jamovi project (2024). *jamovi*. (Version 2.6) [Computer Software]. Retrieved from https://www.jamovi.org.

Vayena, E., Blasimme, A., & Cohen, I. G. (2018). Machine learning in medicine: Addressing ethical challenges. *PLoS Medicine*, *15*(11). https://doi.org/10.1371/journal.pmed.1002689

Velázquez, M. P., Bobrowicz-Campos, E., & Arriaga, P. (2025). *Towards a Typology of Prompts for Human-AI Interaction: Mapping Intent and Complexity with Lay Users*. https://doi.org/10.11159/mhci25.122

Viaene, E., Kuijer, L., & Funk, M. (2021). Learning Systems versus Future Everyday Domestic Life: A Designer's Interpretation of Social Practice Imaginaries. *Frontiers in Artificial Intelligence*, *4*. https://doi.org/10.3389/frai.2021.707562

Vizcaino, M., Buman, M., Desroches, C. T., & Wharton, C. (2019). Reliability of a new measure to assess modern screen time in adults. *BMC Public Health*, *19*(1). https://doi.org/10.1186/s12889-019-7745-6

Vogl, E., Pekrun, R., Murayama, K., Loderer, K., & Schubert, S. (2019). Surprise, Curiosity, and Confusion Promote Knowledge Exploration: Evidence for Robust Effects of Epistemic Emotions. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.02474

Wang, Y. Y., & Wang, Y. S. (2022). Development and validation of an artificial intelligence anxiety scale: an initial application in predicting motivated learning behavior. *Interactive Learning Environments*, *30*(4). https://doi.org/10.1080/10494820.2019.1674887

Web of Stories - Life Stories of Remarkable People. (Uploaded in 2017, recorded in 2005). *Gerald Edelman: The idea of re-entry*. [Video]. Youtube. https://www.youtube.com/watch?v=aTNuZAdzo6k.

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT. *PLoP '23: Proceedings of the 30th Conference on Pattern Languages of Programs.* https://dl.acm.org/doi/10.5555/3721041.3721046

Wormwood, J. B., Siegel, E. H., Kopec, J., Quigley, K. S., & Barrett, L. F. (2019). You are what I feel: A test of the affective realism hypothesis. *Emotion*, *19*(5). https://doi.org/10.1037/emo0000484

Xu, Y., Zhang, J., & Deng, G. (2022). Enhancing customer satisfaction with chatbots: The influence of communication styles and consumer attachment anxiety. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.902782

Yilmaz, H., Maxutov, S., Baitekov, A., & Balta, N. (2023). Student's perception of Chat GPT: a technology acceptance model study. *International Educational Review*, *1*(1).

Yokoi, R., Eguchi, Y., Fujita, T., & Nakayachi, K. (2021). Artificial Intelligence Is Trusted Less than a Doctor in Medical Treatment Decisions: Influence of Perceived Care and Value Similarity. *International Journal of Human-Computer Interaction*, *37*(10). https://doi.org/10.1080/10447318.2020.1861763

Yokoi, R., & Nakayachi, K. (2021). The Effect of Value Similarity on Trust in the Automation Systems: A Case of Transportation and Medical Care. *International Journal of Human-Computer Interaction*, *37*(13). https://doi.org/10.1080/10447318.2021.1876360

Yu, Y., Chang, A. Y. C., & Kanai, R. (2019). Boredom-driven curious learning by homeo-heterostatic value gradients. *Frontiers in Neurorobotics*, *13*. https://doi.org/10.3389/fnbot.2018.00088

Yu, Z., Guindani, M., Grieco, S. F., Chen, L., Holmes, T. C., & Xu, X. (2022). Beyond t test and ANOVA: applications of mixed-effects models for more rigorous statistical analysis in neuroscience research. In *Neuron*. 110(1). https://doi.org/10.1016/j.neuron.2021.10.030

Zheng, Y. (2023). ChatGPT for Teaching and Learning: An Experience from Data Science Education. *SIGITE 2023 - Proceedings of the 24th Annual Conference on Information Technology Education*. https://doi.org/10.1145/3585059.3611431

Appendix A

List of prompts sent to GPT

Prompts sent to GPT-3.5 free version (May 3, 2024) and GPT4.o free version (July 9, 2024)

Initial exploration and examples

1. What are your categorizations on prompts?
2. Explain the main characteristics of each prompt category
3. Share the key differences between each
4. What are the different levels of prompt difficulty?
5. Give me an example of a prompt that has a context dependent level of difficulty (Just for GPT-3.5)
6. Use the topic “common life” and share an example for each category

Re-framing topic for prompts to be used in qualitative analysis

7. Create prompts of all difficulties (basic, intermediate, advanced) for each category framed into wellbeing and personal human development
8. Now, we’ll create new prompts with a new framing topic. Create 3 prompts for each level of difficulty for each of categories framed in day-to-day activities of lay people.
9. Create new prompts for all difficulty levels and categories
10. Create new prompts for all difficulty levels, framed in day-to-day activities

Appendix B

List of prompts from phase two and assessed during phase three of Study One

(Task-oriented)

1. "Analyse the impact of stress on physical health, discussing common stress-related symptoms such as headaches and digestive issues."
2. "Design a weekly meal plan for a busy individual, incorporating grocery shopping lists"
3. "Propose a customised strategy to improve sleep quality, contributing to overall well-being"
4. "Explain the importance of regular breaks throughout the day for maintaining focus and productivity."
5. "Create a personalized self-care ritual that helps unwind and recharge after a long day"
6. "Propose strategies for reducing sedentary behaviour during a typical workday."
7. “Explore practical ways to reduce plastic waste and carbon footprint in my daily life”
8. "Explore the benefits of establishing morning routines for enhancing productivity and well-being, discussing strategies for creating and maintaining effective rituals."
9. “Propose strategies to reduce my household expenses”
10. "Analyse how food traditions of different cultures shape social interactions in common life"
11. “Develop a personalised stress management toolkit for dealing with common stressors in daily life”
12. "Design a personalised time management system to optimize daily productivity and minimize stress."
13. “Create a detailed itinerary for a two-week vacation in Europe”

(Reflective)

1. “Advice me on how could I create a family tradition that makes us feel connected to each other”
2. “Walk me through the process of finding inspiration and motivation in my life to pursue a goal.”
3. “Help me find inspiration to pursue a new hobby / interest”
4. “Help me reflect on the effectiveness of my current behavior and habbits for my personal development”
5. “Reflect with me on how to create a self-care ritual that helps me unwind and recharge after a long day”
6. “How could I broaden my career opportunities and personal growth?”
7. “Advice me on how to foster authentic connections and emotional resilience”
8. "(Thinking about a challenge in my life) give me your feedback on how I handled it.”
9. “Advice me on how could I improve my mental health while dealing with chronic stress”

10. “How should I approach resolving a long-standing conflict with a family member?”
11. “How should I convince my boss to implement a four-day workweek as a means of improving my own work-life balance?”
12. “Advice me on how I can overcome the imposter syndrome in my professional life”
13. "Help me improve my self-care practices and overall well-being"

(Both)

1. "Explain to me how I could apply conflict resolution and negotiation principles to resolve disputes in common life."
2. “Share how behaviour change techniques could promote healthy habits and sustaining long-term lifestyle changes in my life"
3. "Brainstorm ideas for fostering intergenerational connections in the neighborhood"

(None)

1. “Imagine you can travel through time—where and when would you go, and why?”
2. "Do you believe that gratitude practices can improve mental well-being? Share your perspective."
3. "Reflect on a recent meal that made you feel energized and satisfied, and consider the factors that contributed to its healthfulness."
4. “Imagine you could invent a device to solve any global issue—what would it be and how would it work?”
5. “Should freedom of speech be limited to prevent hate speech and misinformation on social media platforms?”

Note. Texts between (...) are just for organisational purposes and did not appear to participants. Prompts appeared in random order for each participant

Appendix C

Final assent from Specialised Committee on Ethics in Psychology of ISCTE-IUL

COMISSÃO ESPECIALIZADA DE ÉTICA DE PSICOLOGIA

PARECER [Final] **24/2024**

I – Identificação

Projeto nº: 24*/2024*

Identificação do(a) proponente: Marijose Páez Velázquez

Curso: Mestrado em Ciências das Emoções

Título do Projeto: End users' interactions with chat GPT, its impact on trust and emotion. Data de submissão do pedido:

Data do parecer: 23 de setembro de 2024

II – Análise

A informação disponibilizada no *Formulário de Submissão para Apreciação da Comissão de Ética Especializada de Psicologia* e respetivos anexos, **satisfaz os requisitos éticos exigíveis neste tipo de projetos de investigação**, contemplando, nomeadamente: o O problema de investigação e relevância do estudo;

o O(s) objetivo(s) e perguntas de investigação;

o O método, incluindo a caracterização dos participantes e o procedimento de recrutamento;

o Identificação de populações vulneráveis, caso se aplique, e riscos associados à participação e correspondentes medidas de mitigação;

o Os elementos do consentimento informado e *debriefing*;

o A entrega do consentimento informado, do protocolo de investigação (guião de entrevista, questionários, etc.) e do *debriefing*.

o A Declaração de Responsabilidade e de Conduta Ética devidamente preenchida.

No entanto, emite as seguintes CEEP recomendações:

1. No consentimento informado (Appendix A) é mencionado "**The personal data you share is anonymous**, and will only be used for statistical, educational, or scientific purposes." A utilização do termo "personal data" afirmação parece-nos desnecessária dado que, de acordo com o mencionado no formulário de submissão, não serão recolhidos dados pessoais.

2. Menção ao período de conservação dos dados (e.g., Os dados anónimos serão armazenados de forma segura por um período de pelo menos cinco anos, desde o final da dissertação ou, se os investigadores planearem reportar os resultados em publicações científicas, desde a data da publicação original).
3. A CE recomenda que os investigadores se certifiquem de que a anonimização é ativada no Qualtrics, de modo a não incluir informação de localização e endereço de IP.

III – Parecer

Em suma, assegurados que se encontram a natureza voluntária da participação, o consentimento livre e informado e o *debriefing*, o adequado tratamento dos dados pessoais, entende a *Comissão de Ética Especializada de Psicologia* emitir **parecer final favorável à realização da investigação**.

A relatora

……

A relatora

……

Appendix D

Informed consent for Study One

The present study arises within the scope of a thesis project for a Masters' of Science of Emotions, taking place at ISCTE - Instituto Universitário de Lisboa. It is conducted by Marijose Páez Velázquez (majoo.pv@gmail.com), whom you may contact if you have any questions, comments, or want to exercise your rights regarding the processing of your data.

The study aims to categorise and evaluate prompts for starting conversations based on their intent and complexity level, respectively. Your participation in the study, which will be highly valued, will contribute to the selection of prompts that can be used in the human-AI interactions, and will be part of a subsequent research. There are no foreseeable risks, damages nor costs associated with participating in the study. Your participation will involve reading a group of sentences (prompts) and evaluating characteristics of each one. **It will take around 30 minutes**. Eventually, participating could benefit you by providing examples and ideas for future human-AI interactions or human-to-human conversations you might have.

You may participate no matter your professional or academic background in technology, but you need to be 18 years or older and have at least an intermediate English reading level. If you don't meet these criteria, unfortunately your responses won't be taken into account.

All the answers that you provide are collected and processed exclusively for the study's purposes. **None of your answers will be evaluated individually. There are no correct or incorrect answers**. ISCTE is responsible for processing your personal data, based on your consent [Article 6(1)(a) and/or Article 9(2)(a) of the General Data Protection Regulation]. T**he data you share is anonymous**, and will only be used for statistical, educational, or scientific purposes. **Your participation in this study is confidential**. Your data will always be handled by authorised personnel bound by confidentiality obligations. ISCTE ensures the use of appropriate techniques, organisational measures, and security to protect information. All researchers are required to maintain the confidentiality of data.

Participation in the study is strictly voluntary—you can freely choose to participate or not. If you decide to participate, you can interrupt your participation at any time without further explanation or consequences to yourself. **You can choose** to not respond partially or totally to the questions in this online form.

ISCTE has a Data Protection Officer, reachable via email at dpo@iscte-iul.pt. If necessary, you also

have the right to lodge a complaint with the competent supervisory authority—the National Data Protection Commission.

I declare that I have understood all the information presented above.
I accept participating in the study and consent to the use of my data according to the provided information.

Appendix E

Debriefing for Study One

Thank you very much for participating in this study.

Your participation will help shape the selection of prompts and support a subsequent research on human-AI interactions, adding an emotional perspective and human focus. We look forward to creating common ground for ethical and responsible interactions with conversational AIs that contribute to our society.

As mentioned at the beginning of your participation, the study aims to categorise and evaluate prompts for starting conversations based on their intent and complexity level, respectively.

We reiterate the contact information you can use if you have any questions, wish to share comments, or express your interest in receiving information about the study's main results and conclusions:
Marijose Páez Velázquez (majoo.pv@gmail.com).

Once again, thank you for your participation.

Appendix F

Detailed response distribution for all prompts assessed in Study One

Category/Prompt	*M* Confirmation Rate (%)	*M* Complexity Level	*M* Interestingness	*M* General Domain
Task-oriented				
"Analyse the impact of stress on physical health, discussing common stress-related symptoms such as headaches and digestive issues"	60.71	4.36	5.57	3.68
"Design a weekly meal plan for a busy individual, incorporating grocery shopping lists"	82.14	3.07	4.93	04.07
"Propose a customised strategy to improve sleep quality, contributing to overall well-being"	71.43	3.79	5.39	3.86
"Explain the importance of regular breaks throughout the day for maintaining focus and productivity"	50.00	2.54	5.11	4.64
"Create a personalised self-care ritual that helps unwind and recharge after a long day"	35.71	3.32	4.57	3.82
"Propose strategies for reducing sedentary behaviour during a typical workday"	64.29	2.89	4.86	4.29
"Explore practical ways to reduce plastic waste and carbon footprint in my daily life"	75.00	3.29	5.43	4.25
"Explore the benefits of establishing morning routines for enhancing productivity and well-being, discussing strategies for creating and maintaining effective rituals"	46.43	3.64	4.82	3.86
"Propose strategies to reduce my household expenses"	71.43	3.43	5.29	4.21
"Analyse how food traditions of different cultures shape social interactions in common life"	50.00	4.29	4.64	3.39
"Develop a personalised stress management toolkit for dealing with common stressors in daily life"	57.14	4.39	4.82	3.36
"Design a personalised time management system to optimize daily productivity and minimise stress"	57.14	4.14	4.68	3.57
"Create a detailed itinerary for a two-week vacation in Europe"	82.14	3.29	4.00	3.32
Reflexive				
"Advice me on how could I create a family tradition that makes us feel connected to each other"	60.71	3.89	3.82	3.89
"Walk me through the process of finding inspiration and motivation in my life to pursue a goal"	57.14	4.46	3.93	3.75
"Help me find inspiration to pursue a new hobby / interest"	71.43	3.68	4.04	4.32
"Help me reflect on the effectiveness of my current behavior and habits for my personal development"	64.29	5.04	4.61	4.36
"Reflect with me on how to create a self-care ritual that helps me unwind and recharge after a long day"	53.57	3.89	4.32	4.61
"How could I broaden my career opportunities and personal growth?"	35.71	4.86	5.07	4.54
"Advice me on how to foster authentic connections and emotional resilience"	46.43	4.79	3.82	3.79
"(Thinking about a challenge in my life) give me your feedback on how I handled it"	67.86	4.64	3.93	5.50

Appendix F (continued)

Category/Prompt	*M* Confirmation Rate (%)	*M* Complexity Level	*M* Interestingness	*M* General Domain
"Advice me on how could I improve my mental health while dealing with chronic stress"	42.86	5.04	5.07	3.93
"How should I approach resolving a long-standing conflict with a family member?"	60.71	5.25	3.75	4.39
"How should I convince my boss to implement a four-day workweek as a means of improving my own work-life balance?"	35.71	4.71	5.21	4.86
"Advice me on how I can overcome the imposter syndrome in my professional life"	32.14	5.43	4.64	4.18
"Help me improve my self-care practices and overall well-being"	42.86	4.07	4.79	4.43
Both				
"Explain to me how I could apply conflict resolution and negotiation principles to resolve disputes in common life"	35.71	4.54	4.64	3.89
"Share how behaviour change techniques could promote healthy habits and sustaining long-term lifestyle changes in my life"	25.00	4.43	4.96	3.89
"Brainstorm ideas for fostering intergenerational connections in the neighborhood"	25.00	3.93	4.14	4.57
None				
"Imagine you can travel through time—where and when would you go, and why?"	28.57	2.96	5.21	5.57
"Do you believe that gratitude practices can improve mental well-being? Share your perspective"	14.29	3.32	4.46	4.89
"Reflect on a recent meal that made you feel energized and satisfied, and consider the factors that contributed to its healthfulness"	7.14	3.75	5.00	4.54
"Imagine you could invent a device to solve any global issue—what would it be and how would it work?"	14.29	5.25	5.07	4.43
"Should freedom of speech be limited to prevent hate speech and misinformation on social media platforms?"	14.29	5.21	4.32	4.36

Appendix G

Informed consent for *Main Study*

The present study arises within the scope of a thesis project for a Masters' of Science of Emotions, taking place at ISCTE - Instituto Universitario de Lisboa. It is conducted by Marijose Páez Velázquez (majoo.pv@gmail.com), whom you may contact if you have any questions, comments, or want to exercise your rights regarding the processing of your data.

The study aims to elicit end users' interactions with ChatGPT, and explore their impact on trust and emotion. Your participation in the study, which will be highly valued, will contribute to the human-AI interaction research. There are no foreseeable risks, damages nor costs associated with participating in the study.

Your participation will entail only one session and **requires creating a ChatGPT account**. In this session, you will **interact twice with this conversational technology**, answer different surveys on your feelings and thoughts, and will take **about 35 minutes**. **We highly recommend using a computer for this study.**

You may participate no matter your professional or academic background in technology, but you need to be 18 years or older and have at least an intermediate English reading level. If you don't meet these criteria, unfortunately your responses won't be taken into account.

All the answers that you provide within this online form are collected and processed exclusively for the study's purposes. **None of your answers will be evaluated individually. There are no correct or incorrect answers.** ISCTE is responsible for processing your data, based on your consent [Article 6(1)(a) and/or Article 9(2)(a) of the General Data Protection Regulation, as applicable]. **The data you share in this form is anonymous**, and will only be used for statistical, educational, or scientific purposes. It will be stored securely and eventually deleted in accordance with the Guidelines for Researchers on the Protection of Personal Data in Scientific Research Activities of ISCTE.

Your participation in this study is confidential. Your data will always be handled by authorised personnel bound by confidentiality obligations. ISCTE ensures the use of appropriate techniques, organisational measures, and security to protect information. All researchers are required to maintain the confidentiality of data.

Please consider that **all the information you share in the ChatGPT page** (such as sign up info) and within your conversations there, will be handled by OpenAI according to their own privacy policy (https://openai.com/policies/privacy-policy) and terms of use (https://openai.com/policies/terms-of-use/). Neither the researcher nor ISCTE will have access to your email address, your conversations with ChatGPT or any other information you share in the ChatGPT page.

Participation in the study is strictly voluntary—you can freely choose to participate or not. If you decide to participate, you can interrupt your participation at any time without further explanation or consequences to yourself. **You can choose** to not respond partially or totally to the questions in this online form.

ISCTE has a Data Protection Officer, reachable via email at dpo@iscte-iul.pt. If necessary, you also have the right to lodge a complaint with the competent supervisory authority—the National Data Protection Commission.

I declare that I have understood all the information presented above.
I accept participating in the study and consent to the use of my data according to the provided information.

Appendix H

Debriefing for *Main Study*

Thank you very much for participating in this study.

Your participation will contribute to the human-AI interaction research, integrating an emotional perspective and human focus. We look forward to creating common ground for ethical and responsible interactions with conversational AIs that contribute to our society.

Specific details were initially withheld as part of the study design to ensure unbiased results, but rest assured that none of this information had any harm or risk potential. Now we can provide more information as follows:

As mentioned at the beginning of your participation, the study focuses on end users' interactions with ChatGPT, and their impact on trust and emotion. More specifically, to analyse how predictability explains the effect of type of interaction in trust when engaging with ChatGPT and if accuracy of prediction has an effect on emotional responses. Also, to explore if previous knowledge moderates the relationship between type of interaction and predictability, and if attitudes towards AI moderate the relationship between type of interaction and trust.

We reiterate the contact information you can use if you have any questions, wish to share comments, or express your interest in receiving information about the study's main results and conclusions: Marijose Páez Velázquez (majoo.pv@gmail.com).

Once again, thank you for your participation.

Appendix I

Invitations to prospective participants (examples)

Example on personalised request for dissemination of the study.

Subject: Request for Authorization in sharing Research Study with your community

Distinguished Professor [Name] and [Role e.g. Director] of [Department/Organization],

I hope this message finds you well. My name is Marijose Páez, I am conducting a research study titled **"End users' interactions with ChatGPT, its impact on trust and emotion"**, framed on a masters' thesis project from ISCTE-IUL and supervised by professor Elzbieta Bobrowicz-Campos. I am reaching out to request your authorization and support in sharing this study with the [research/academic] community (students and/or colleagues) of the institution that you [take part, direct, manage, etc]
The general objective is to understand trust and emotion dynamics when interacting with ChatGPT, with users from different levels of expertise in Artificial Intelligence (AI) and a varied set of contexts.

Why it matters

Interactions with AI agents will continue to grow, transforming our daily lives in unprecedented ways. Large Language Models, such as ChatGPT, enable broader access for end users with diverse backgrounds and many levels of expertise. Systems will likely become more and more complex, so will the users' mental models. To expand our capabilities through technology, trust and emotion are key as they will allow safer and more productive interactions.
Currently no trust-related studies include interactions with end users and ChatGPT as part of the experimental design. This study will contribute to the human-AI interaction research.

What we are looking for

Participants of 18 years and older with intermediate or higher English level. No specific technical background is required, as the study is intentionally designed to include end users with diverse levels of AI knowledge.

This is a one-time intervention. Participants will be asked to complete an online questionnaire, which takes approximately 60 minutes and must be done on a computer.

How your valued contribution will impact our research

Your distinguished efforts play a crucial role in reaching a diverse audience for the study, enriching its

scope and impact.

We kindly request your authorization and support to share the invitation to participate in this study with your institution's community, in the terms that work best for you and your organization. We aim to reach an audience consisting of students, professors, and other research colleagues.

In the attached document, you will find the information targeted to potential participants (invitation, general description, and link to participate), which can be shared upon your approval.

We sincerely thank you for your invaluable support in disseminating the invitation to participate in our study within the academic community of your [institution/organization].

Information on ethics

The study is approved by the Ethics Committee of the ISCTE-IUL Psychology Department (PSI_24/2024). Voluntariness, confidentiality and data protection will be strictly maintained according to ISCTE guidelines and General Data Protection Regulation.

There are no foreseeable risks, damages nor costs associated with participating in the study.

Additionally, we will be happy to provide further details or address any questions you may have regarding the research.

We deeply appreciate your invaluable collaboration in advancing this initiative, and we eagerly look forward to your favourable response.

Figure 1. Graphic example for dissemination the study

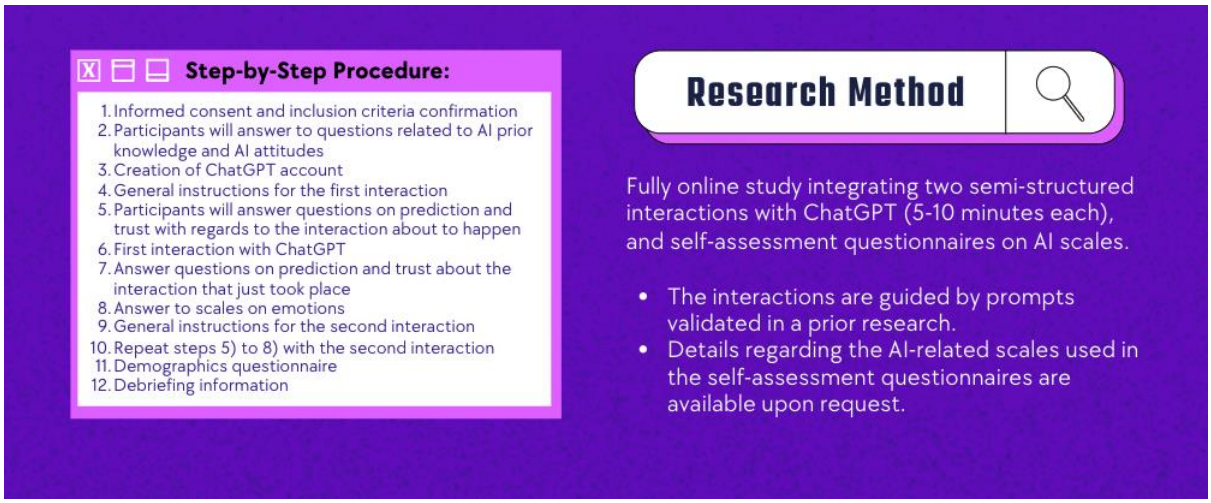

Figure 2. Graphic example for dissemination the study

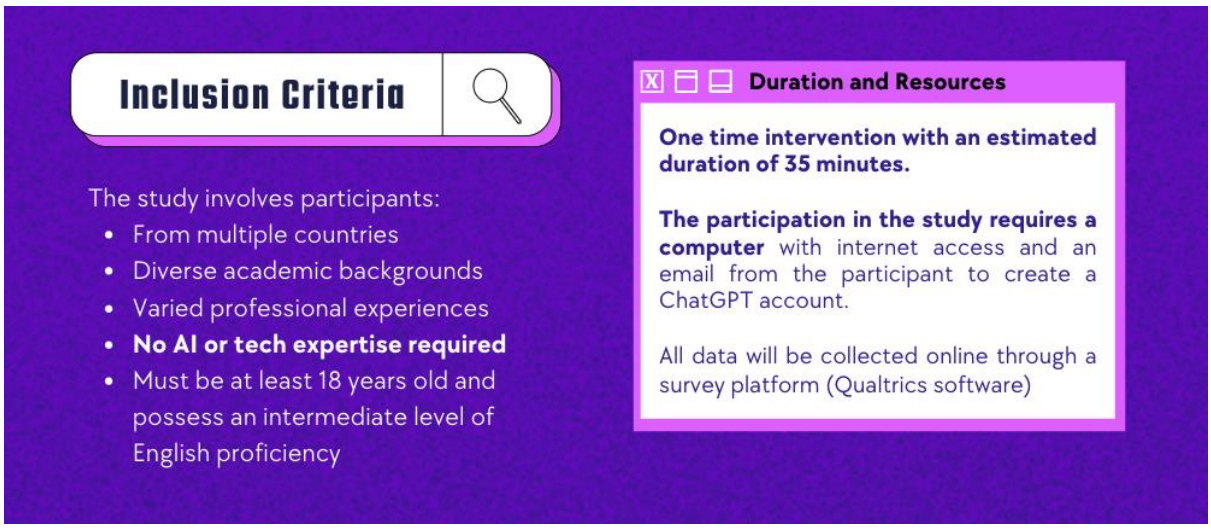

Figure 3. Graphic example for dissemination the study

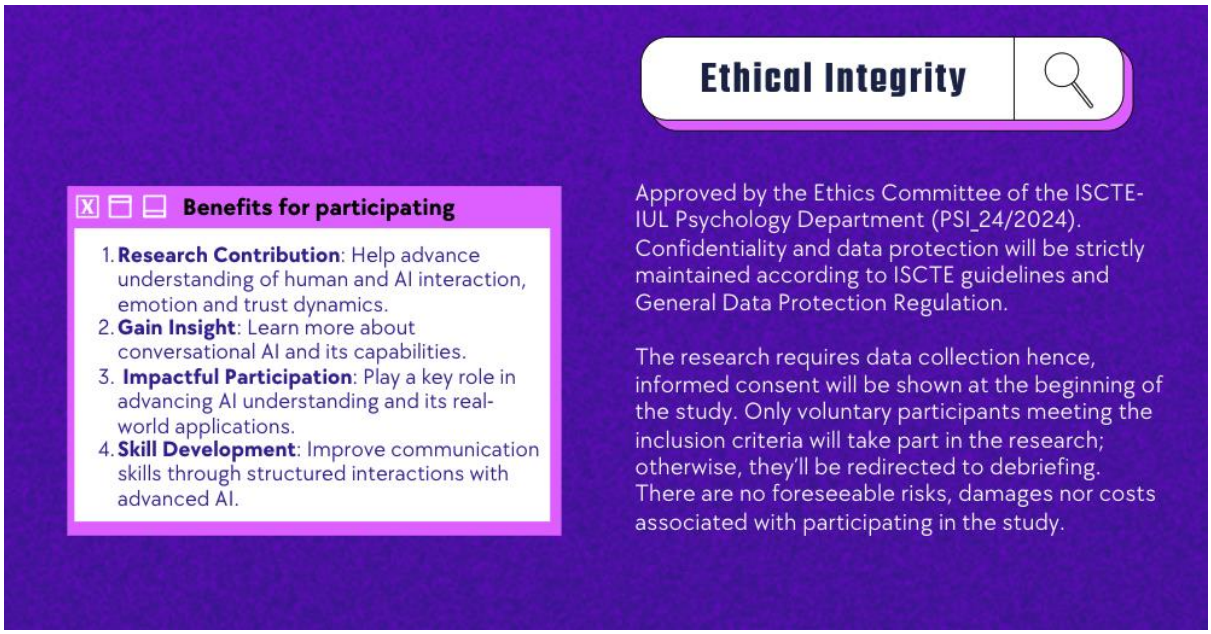

Figure 4. Graphic example for dissemination the study

Figure 5. Graphic example for dissemination the study

Appendix J

Attitudes towards AI

AI attitude scale (AIAS-4) from Grassini (2023).

Please select the option that best describes yourself:

from 1 (not at all) to 10 (completely agree)

1. I believe that AI will improve my life
2. I believe that AI will improve my work
3. I think I will use AI technology in the future
4. I think AI technology is positive for humanity

Appendix K

AI prior knowledge

Familiarity, adapted from Körber (2019), interaction nature (custom-created), and frequency of use, adapted from Viszcaino et al., (2019).

Please select the option that best describes yourself:

1. I already know Chat GPT or similar systems as Co-pilot, Dall-E or other intelligent chatbots (1 strongly disagree - 5 strongly agree)
2. I have already used Chat GPT or similar systems as Co-pilot, Dall-E or other intelligent chatbots (1 strongly disagree - 5 strongly agree)

3. When using GPT or similar systems of AI chatbots, I use it for (Personal reasons/Professional reasons/Both reasons/None)

Please write the number that best describes yourself:

4. Thinking of an average, <u>how many hours</u> do you typically interact with a Large Language Model (LLM) or Generative Artificial Intelligence (GenAi) such as ChatGPT, Co-pilot, or Dall-E on a <u>typical week</u>?
Write the number that best describes yourself. If zero, please type “0” in the box.

AI literacy scale (Laupichler et al., 2023a)

Please select the option that best describes yourself:
from 1 (strongly disagree) to 7 (strongly agree)

<u>(Technical understanding)</u>

I can…

1. describe how machine learning models are trained, validated, and tested.
2. explain how deep learning relates to machine learning.
3. explain how rule-based systems differ from machine learning systems
4. explain how AI applications make decisions.

5. explain how ’reinforcement learning’ works on a basic level (in the context of machine learning).
6. explain the difference between general (or strong) and narrow (or weak) artificial intelligence
7. explain how sensors are used by computers to collect data that can be used for AI purposes.
8. explain what the term ’artificial neural network’ means.
9. explain how machine learning works at a general level.
10. explain the difference between ’supervised learning’ and ’unsupervised learning’ (in the context of machine learning).
11. describe the concept of explainable AI.
12. describe how some artificial intelligence systems can act in their environment and react to their environment.
13. describe the concept of big data.
14. evaluate whether media representations of AI (e.g., in movies or video games) go beyond the current capabilities of AI technologies.

(Critical appraisal)

I can…

15. explain why data privacy must be considered when developing and using artificial intelligence applications.
16. explain why data security must be considered when developing and using artificial intelligence applications
17. identify ethical issues surrounding artificial intelligence.
18. describe risks that may arise when using artificial intelligence systems.
19. name weaknesses of artificial intelligence.
20. describe potential legal problems that may arise when using artificial intelligence.
21. critically reflect on the potential impact of artificial intelligence on individuals and society
22. describe why humans play an important role in the development of artificial intelligence systems.
23. explain why data plays an important role in the development and application of artificial intelligence.
24. describe what artificial intelligence is.

(Practical application)

I can…

25. give examples from my daily life (personal or professional) where I might be in contact with artificial intelligence.
26. name examples of technical applications that are supported by artificial intelligence.

27. tell if the technologies I use are supported by artificial intelligence
28. assess if a problem in my field can and should be solved with artificial intelligence methods.
29. name applications in which AI-assisted natural language processing/ understanding is used
30. explain why AI has recently become increasingly important.
31. critically evaluate the implications of artificial intelligence applications in at least one subject area

Note. Titles of factors are just for organisational purposes and did not appear to participants.

Appendix L

Predictability

Predictability is formed by prediction (before interaction) and prediction verification (after interaction) with ChatGPT. Custom-created scale adapted from Körber (2019), Yokoi et al. (2021), and Yokoi & Nakayachi (2021)

Prediction scale (before interaction):

The next questions are about the interaction that will take place with ChatGPT about (task-oriented/reflective) topics. Please select the option that best describes your belief about the future interaction.

1. ChatGPT's thoughts on topics concerning (task-oriented/reflective) interactions will match my thinking
2. ChatGPT will have similar beliefs on topics concerning (task-oriented/reflective) interactions as I do
3. ChatGPT will answer my (task-oriented/reflective) questions in a way that I desire
4. Chat GPT will respond unpredictably [R]
5. I will be able to understand why things happened
6. I find it difficult to predict what Chat GPT will do next [R]
7. I am uncertain about the outcome of this (task-oriented/reflective) interaction[R.]
8. I can imagine the responses that ChatGPT will deliver on (task-oriented/reflective) topics
9. I have a clear expectation on how ChatGPT will perform in this task

Prediction Verification scale (after interaction):

The next questions are about the interaction that just took place with ChatGPT on (task-oriented/reflective) topics. Please select the option that best describes your perception about the past interaction.

1. The way ChatGPT thinks about (task-oriented/reflective) topics matches my thinking
2. ChatGPT has similar beliefs on topics concerning (task-oriented/reflective) interactions as I do
3. ChatGPT answers my (task-oriented/reflective) questions in a way that I desire

4. ChatGPT responded unpredictably [R]
5. I was able to understand why things happened
6. It's difficult to identify what ChatGPT will do next [R]

7. I am still uncertain about the outcome of this interaction
8. My predictions about ChatGPT responses on (task-oriented/reflective) interactions were accurate
9. My expectation on how ChatGPT would perform this task was met

Note. R Questions with reversed answer to validate congruence

Appendix M

Trust

Trust scale assessed before and after interactions (Gulati et al., 2019)

Please select the option that best describes yourself
from 1 (strongly disagree) to 5 (strongly agree)

1. I believe that there could be negative consequences when using ChatGPT
2. I feel I must be cautious when using ChatGPT
3. It is risky to interact with ChatGPT
4. I believe that ChatGPT will act in my best interest
5. I believe that ChatGPT will do its best to help me if I need help
6. I believe that Chat GPT is interested in understanding my needs and preferences
7. I think that ChatGPT is competent and effective in answering (task-oriented / reflective) questions
8. I think that ChatGPT performs its role as conversational AI very well
9. I believe that ChatGPT has all the functionalities I would expect to answer my (task-oriented/reflective) questions
10. If I use ChatGPT, I think I would be able to depend on it completely
11. I can always rely on Chat GPT for answering my (task-oriented/reflective) questions
12. I can trust the information presented to me by ChatGPT

Appendix N

Emotional responses

Scales on emotional valence (Briesemeister et al., 2012) and epistemic emotions (Pekrun et al., 2017).

We are interested in the emotions you experienced after interacting with ChatGPT in a (reflexive/task-oriented) conversation. Please rate the intensity of positive and negative emotions that you feel.

From 1 (low) to 7 (high)

1. Positive emotion
2. Negative emotion

We are interested in the emotions you experienced when interacting with ChatGPT in a (task-oriented/reflective) conversation.
For each emotion, please indicate the strength of that emotion by selecting the number (1 - not at all to 5 - very strong) that best describes the intensity of your emotional response during interacting with ChatGPT.

1. Surprised
2. Curious
3. Excited
4. Confused
5. Anxious
6. Frustrated
7. Bored

Appendix O

Instructions for ChatGPT account creation

You have already completed more than 30% of the study!
Your complete fulfilment of this survey will help us understand the dynamics of human-AI interactions. Please keep going.

You will now create an OpenAI account and interact with ChatGPT. If you already have an account, you can use your existing one.

Access https://chatgpt.com/auth/login and create your account. **Please avoid sending any messages or start an interaction at this moment** and come back to this page when you finish.

When you have a created account, click “Next”.

Appendix P

General instructions for the interaction with ChatGPT

For task-oriented interactions, the next instructions appeared to participants:

Next, **you will interact with ChatGPT in a task-oriented conversation** that involves asking for analysis, explanations, or customised requests. An example of this type of cue is: “Create a detailed itinerary for a week vacation in Europe”

You can interact freely and will be able to customise your questions within these framing scenarios.

Before interacting, please respond to the next questions…

For reflexive interactions, the next instructions appeared to participants:

Next, **you will interact with ChatGPT in a reflective conversation** that involves asking for advice, guidance, or personal development assistance. An example of this type of cue is: “Help me find inspiration to pursue a new hobby / interest”

You are free to interact and will be able to customise your questions within these framing scenarios.

Before interacting, please respond to the next questions…

Appendix Q

Detailed instructions for the interaction with ChatGPT

For task-oriented interactions, the next detailed instructions appeared to participants:

Now, you will actually interact with ChatGPT in a task-oriented conversation that involves asking for analysis, explanations, or customised task-assistance requests. Example: "Create a detailed itinerary for a week vacation in Europe".

This interaction will take around 5-10 minutes.

You can interact freely and honestly, but please keep your interaction within the scope of a task-oriented conversation. Avoid asking unrelated questions.

Please read and follow the instructions carefully:

Choose one of the following cues to interact with ChatGPT:

a) Propose strategies to reduce my household expenses
b) Propose a customised strategy to improve sleep quality, contributing to overall well-being
If you fail to use one of these cues, your answers will not be considered.

Follow this format adjusting it to one of the cues above (a OR b):

> **Greeting**: Say hi
> Example: Hello Chat!
> **Objective** of interaction: task-oriented
> Example: I want to have a task-oriented conversation with you.
> **Main question, chosen from the previous cues (a OR b).**
> Example: Create a detailed itinerary for a week vacation in Europe.
> **Additional Details:** Share why you want to know this, or add personal preferences for customisation.
> Example: I'm travelling with my sister.
> **Desired Response:** Emphasise information you are interested in.
> Example: Include valuable information on the best spots for taking pictures
> **Open-Ended Mandatory Format**: This will control the answer you get by ChatGPT
> "One step at a time, between 200 to 300 words-length".
> Example: Send one day at a time. Each answer should be between 200 to 300 words.

Ending: Polite closure

Example: Thanks for your insights!

Example:

Hello Chat!

… (ChatGPT's response) …

I want to have a task-oriented conversation with you.

… (ChatGPT's response) …

Create a detailed itinerary for a week vacation in Europe. I'm travelling with my sister. Include valuable information on the best spots for taking pictures. Send one day at a time. Each answer should be between 200 to 300 words. Thanks for your insights!

… (ChatGPT's response) …

(Interaction continues)

Recommendations:

Avoid using the same cue as in the example, use one of the list of cues and personalise it as explained before

Be clear & specific

Keep respectful and polite conversations

Ask for the mandatory format ("send one step or message at a time, between 200-300 words-length")

Share details for customisation

Avoid overwhelming requirements, mixing varied topics.

Please come back to this page when you finish the interaction.

I finished the interaction.

For reflexive interactions, the next detailed instructions appeared to participants:

Now, you will actually interact with ChatGPT in a reflexive conversation that involves asking for advice, guidance, or personal development assistance. Example: "Help me reflect on the

effectiveness of my current behaviour and habits for my personal development."

This interaction will take around 5-10 minutes.

You can interact freely and honestly, but please keep your interaction within the scope of a reflexive conversation. Avoid asking unrelated questions.

Please read and follow the instructions carefully:

Choose one of the following cues to interact with ChatGPT:

a) Help me find inspiration to pursue a new hobby / interest

b) Thinking about a challenge in my life, give me your feedback on how I handled it

If you fail to use one of these cues, your answers will not be considered.

Follow this format adjusting it to one of the cues above (a OR b):

> **Greeting**: Say hi
>
> Example: Hello Chat!
>
> **Objective** of interaction: reflexive
>
> Example: I want to have a reflexive conversation with you.
>
> **Main question, chosen from the previous cues (a OR b).**
>
> Example: Help me reflect on the effectiveness of my current behaviour and habits for my personal development
>
> **Additional Details**: Share why you want to know this, or add personal preferences for customisation.
>
> Example: I've struggled to create a routine that helps me achieve my goals.
>
> **Desired Response:** Emphasise information you are interested in.
>
> Example: I'm interested in your advice.
>
> **Open-Ended Mandatory Format:** This will control the answer you get by ChatGPT
>
> "One step at a time, between 200 to 300 words-length".
>
> Example: Send one step at a time. Each answer should be between 200 to 300 words.
>
> Ending: Polite closure
>
> Example: Your perspective will be much appreciated!

Example:

Hello Chat!

… (ChatGPT's response) …

I want to have a reflexive conversation with you.

… (ChatGPT's response) …

Help me reflect on the effectiveness of my current behaviour and habits for my personal development. I've struggled to create a routine that helps me achieve my goals. I'm interested in your advice. Send one step at a time. Each answer should be between 200 to 300 words. Your perspective will be much appreciated!

… (ChatGPT's response) …

(Interaction continues)

Recommendations:

Avoid using the same cue as in the example, use one of the list of cues and personalise it as explained before

Be clear & specific

Keep respectful and polite conversations

Ask for the mandatory format ("send one message at a time, between 200-300 words-length")

Share details for customisation

Avoid overwhelming requirements, mixing varied topics.

Please come back to this page when you finish the interaction.

I finished the interaction.

Appendix R

Prompt used for interacting with ChatGPT

1. Did you interact with ChatGPT in a (reflexive/task-oriented) conversation?
   (Yes/no)
2. Please select the prompt that you used:
   (list of four options: three prompts pre-selected from Study 1, including the example and "other not listed" option)

You've already completed more than 50% of the study!
If a notice from ChatGPT regarding a limit on the free plan appears, please ignore it and continue interacting with the free plan (you won't be charged).

Appendix S

Questions for the pilot study

Level of agreement for:

The questions of the questionnaire were clear

The instructions for interacting with ChatGPT were clear

The navigation through the questionnaire was straightforward

The response options were always visible and easy to select

How often did you…

Needed to go back to check previous instructions

Went back to change some of your answers

Changed your answers from a previous page

How long did it take you to complete the questionnaire/study?

Please share your feedback.

General comments, specific unclear questions, etc.