Proceedings of the 11th World Congress on Electrical Engineering and Computer Systems and Sciences (EECSS'25)
Paris, France - August, 2025
Paper No. MHCI 122
DOI: 10.11159/mhci25.122

Towards a Typology of Prompts for Human-AI Interaction: Mapping Intent and Complexity with Lay Users

Marijose Páez Velázquez, Elzbieta Bobrowicz-Campos, Patrícia Arriaga
ISCTE-Instituto Universitário de Lisboa
Lisbon, Portugal
mpvzz@iscte-iul.pt; elzbieta.campos@iscte-iul.pt; patricia.arriaga@iscte-iul.pt

***Abstract* -** Large Language Models enable broader access to AI for end users with diverse backgrounds and varying levels of expertise. A significant and growing proportion of the population is interacting with highly sophisticated technology. This opens questions related to the nature, dynamics, and effects of such interactions, specially among lay users. To understand these dynamics, we must first identify the types of interactions that may take place according to the user's intent and their complexity, since LLMs allow unprecedented freedom to be used in different contexts. To date, no categorisation of prompts with comparable complexity levels has been developed from the user's perspective, avoiding confounding variables when studying human-AI interactions. To address this gap in prompts, we applied three sequential methodological approaches. First, we *explored* prompt categories and complexity levels through iterative queries with ChatGPT. The prompts were written by GPT itself. Second, we analysed these textual data using a *thematic qualitative approach* and curated a pre-set of 34 prompts with comparable complexity. Prompts were classified into two main categories: "task-oriented" and "reflexive", with two additional controls: "both" and "none". Third, we conducted a *validation study* with 28 lay users from different countries through an online survey. "Task-oriented" prompts achieved a mean category confirmation rate of 62% (Max = 82%), and "reflexive" prompts reached 52% (Max = 71%). Complexity levels averaged near the central point of the scale (M =4.10). A smaller set of 12 prompts with at least 60% of category agreement was obtained. This study lays an empirical foundation for investigating complexity-comparable types of interaction between lay users and LLM-powered conversational agents. Ultimately, this study contributes to advancing research in human-AI interaction, offering a validated set of prompts suitable for investigating trust dynamics, emotional responses, and other key constructs in HCI.

***Keywords*:** Large-Language-Model, LLM, artificial intelligence, human-AI interaction, prompt-based interactions, ChatGPT, Human Computer Interaction, mixed methods

1. Introduction

Human interactions with Artificial Intelligence (AI) will continue to grow, transforming our daily lives in unprecedented ways as Large Language Models (LLMs) broaden access to AI for users from diverse backgrounds and with many levels of expertise [1], [2]. Yet, research literature on the effects of LLMs is still in its infancy [3]. This paper aims to provide a foundation for classifying types of interactions between lay users and LLMs. It addresses the need for a shared understanding of prompt categorisation based on interaction type and complexity level. Developing a structured categorisation of prompts can help future studies exert more control over the types of interactions being elicited and analysed to systematically compare emotional responses, trust dynamics, epistemic engagement, and other constructs, thereby improving the comparability of findings in Human-Computer Interaction (HCI).

2.1. Large Language Models

Machine Learning algorithms have evolved into highly sophisticated systems, such as Deep Learning [4], [5]. 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 [5], [6]. Namely, autoregressive LLMs are models trained on a massive amount of data to process, understand and generate human-like language. They 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 [6],[7]. This has led LLMs to an unprecedented freedom of applications in many different contexts (for more information, see [5]). While the history of language models and other forms of AI is not new (see [4] for a review of this evolution), recent developments have created a paradigm shift [8]: Besides the rapid rhythm of performance improvements in AI tools [9], LLMs and other forms of generative models enable a non-technical audience to actively create many types of content—as text and graphics— through simple conversational inputs (prompts). This ability to create novel content has come to public discourse as the relatively young term of "Generative Artificial Intelligence" (GenAI) [10]. AI systems have been integrated into entertainment, healthcare, or customer service [2], [4], [11] for years. However, the development of LLMs introduces new dimensions for the lay population, making it fundamental to understanding the types of interactions explored in the following sections.

2.2. ChatGPT

The first LLM with a relevant impact on public perception was OpenAI's ChatGPT [4]. Among several LLMs available in both free and paid versions, ChatGPT stands out not only as the most commonly used [12], but also as the AI system with the fastest adoption growth in the history of technology [4]. In April 2025, Sam Altman reported that 10% of the global population uses OpenAI systems, suggesting that ChatGPT had reached around 800 million users [13].

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 [14], problem-solving [8], or rating and recommendation [6]. From the users' perspective, research in educational contexts has studied ChatGPT's acceptance, usage, learning practices and motivation among students [15], [16], [17], [18]. Industry research has explored how GPT-powered tools affect productivity and employee retention [19]. A recent study by OpenAI and MIT [20] 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 [21]. 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.

2.3. Types of interactions

Historically, research on user engagement with technology has been grounded on its instrumental value [22] as AI service chatbots have been widely used in customer service, with ongoing efforts to enhance communication style and empathetic responses to improve business outcomes [22], [23], [24], [25]. Nevertheless, 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 tasks (e.g., [7], [14]) and opened up new forms of interaction that extend beyond traditional brand-consumer transactional exchanges [5]. More recently, emerging research has aimed to understand users' behaviours and meaning-making processes, including usage intent, emotional responses, trust, engagement, and action [1], [20], [22]. For example, recent research on health and AI examines LLM responses to caregivers seeking support [26] and emotional coping [27]. In professional contexts, research shows that one-third of employees think that robots would provide more unbiased feedback than managers [23]. Some users have even described conversational AI agents as a family member or friend [28], [29]. Research on LLM-powered social chatbots—such as Replika and Xiaolce—examines systems designed for companionship and therapeutic conversations [27]. Users report discussing everyday activities—hobbies and sleeping habits—, mental states and philosophical topics, personal worldviews, as well as personal problems—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 [28]. Although these interactions often pivot on companion-seeking behaviour, research shows that users also engage in similar self-disclosing conversations with more generalist LLMs [20]. This potentially broadens the nature of such conversations, diversifying interaction types, increasing the breadth of information exchange, and deepening vulnerability [28]. 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 [20]; though, heavy users tend to include more affective cues compared to casual users [21]. Conversely, research on Replika indicates that some users start with deep conversations and gradually reduce their emotional expression over time

[28], possibly explained by differences in the user's initial motivations [30]. Research on AI voice assistants [1] found that asking questions (20%), entertainment-jokes (12%), and searching for information (11%) are among the top 5 interactions children have with Alexa, followed by playing music (40%). 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 [1], though typically across all interaction types without accounting for differences in complexity. These patterns highlight the need for more nuanced analysis, as the conversation type significantly affects user's emotional and psychological responses [21]. 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.

2.4. 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 [31]. While there has been an increasing interest in prompt-related research for optimised constructing and testing—such as prompt catalogs, classifications, and usage recommendations [31], [32], [33], [34], 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 [35], adapt, and optimize [36] prompts. Recent studies have explored how different prompt types elicit distinct user interactions. OpenAI and MIT [20] 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. Similarly, TUM researchers [37] 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 [38], 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.

2.5. Research question

This study seeks to identify and validate a set of user-oriented prompt categories across comparable levels of complexity. Our aim is to support future HCI research and practical applications by providing a structured resource that enables consistent design and analysis of LLM-based interactions. RQ: Which prompts can be reliably grouped under the same intent category and level of complexity, from a user-centered perspective?

3. 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 validate the selected set of prompts. Given ChatGPT's demonstrated performance in assisting with complex tasks, alongside its perceived human-likeness which may encourage more personal-oriented interactions, this study focused on identifying prompts that specifically elicit goal-oriented or personal conversations, excluding other applications and types of interactions from its scope.

3.1. First Phase Procedure: Exploring prompts with ChatGPT

As prior studies [3], [39] 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 [20], [21], [28]. This raw data was used in the qualitative analysis. Following current Best Practices on LLMs usage in research [7], we provide the list of questions sent to ChatGPT in both iterative sessions as Supplementary Material 1.

3.2. Second Phase Procedure: Refining through qualitative approach

Through a qualitative thematic approach [40], we interpreted and constructed themes by analysing raw responses from both GPT versions. Working with prompts around everyday life topics reduced background knowledge bias and allowed for better control when comparing complexity levels across prompts. For example, an advanced prompt focusing on "emotional regulation strategies" could be simple to a Psychologist but complex to a Chemist. An intermediate "carbon footprint and emission" prompt could be relatively simple for the Chemist, but more complex for the Psychologist. All selected prompts 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 recoding took place as we detected superpositions on the categories and complexities of 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. We selected prompts of comparable complexity, intentionally avoiding overly straightforward ones, to ensure that future research could explore more substantive and meaningful interactions with LLMs. We filtered out the prompts that were still outside of the four broader groups, too specific, too wide, repetitive, or too context-specific. 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.

3.3. Third Phase Procedure: Quantitative validation with end users

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

3.3.1 Ethics

The subject research was approved by the Specialised Committee on Ethics in Psychology of ISCTE-IUL (PSI_24/2024, September 2024). Only the individuals who accepted voluntarily to participate and met the inclusion criteria took part in the research; otherwise, they were redirected to the 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.

3.3.2 Participants

Participants were recruited through convenience sampling and included individuals 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), 45 participants did not complete the study, and 3 were excluded due to inattentive responding (identified through control item "please check somewhat disagree for this item" [2]). 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.

3.3.3 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 [41], 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 frequency of use based on Vizcaino, Buman, Desroches, and Wharton [42]. We also assessed AI Literacy using the 31-item scale for non-experts by Laupichler et. al. [2], which uses a 7-point Likert format (1-strongly disagree to 7-strongly agree), and comprises 3 factors: technical understanding (14 items), critical appraisal (10 items), and practical application (7 items). Each factor was then analysed as composite variables. 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 [21]. 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. Results

4.1. First Phase: Exploring prompts with ChatGPT

GPT-3.5 and GPT-4.o proposed distinct prompt categorisations—10 categories with 7 complexity levels and 13 categories with 3 levels, respectively. Categories included: informational, creative, opinion-based, problem-solving, educational, reflective, feedback, personal assistance, persuasive, analytical, conversational, experiential, comparative, etc; while complexities included: easy, intermediate, advanced, specialised, etc. A detailed list can be found in Supplementary Material 2. We obtained 69 preliminary prompts as examples, and selected the following categories: "Informational," "comparative," and "analytical" for their educational relevance, a widely documented use of LLMs [12]; "Personal-assistance" and "feedback and guidance" for their introspective nature, consistent with prior work describing LLMs as companion partners and tools for self-reflecting [28]; "Problem-solving", given its potential applicability to real-world user needs. We excluded "opinion-based" and "experience" prompts, as ChatGPT lacks personal narratives [26]. Similarly, "creative" and "entertainment" prompts were beyond the study's scope. "Persuasive/argumentative," present only in GPT-3.5, was excluded as it pertained more to tone than intent. After reframing into everyday life topics, we gathered 116 prompts.

4.2. Second Phase: Refining through qualitative approach

Prompts may fall into multiple categories simultaneously, for example, “informational/instructional”, “educational/technical”, “comparative/contrastive”, and “analytical” categories largely overlap in purpose. Qualitative analysis revealed that all aim to explain, clarify, or examine a specific topic. Differences often lie in the level of specificity required, or, in the directive verb used—such as Describe/Explore, Explain, Compare/Contrast, or Analyze—rather than in different types of content. This overlap became evident when asking GPT 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. 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—asked users to reflect on experiences, feelings, or behaviors. These were labeled as “personal” in previous research [20], but we decided to name them as “reflexive” since “personal” potentially includes other topics/intents, such as emotional support or casual conversation/small talk. Verbs marking these prompts included Reflect, Share, Examine, and Advice. Previous literature on self-reflection [28], [29], [30] supported this category. Lastly, we identified prompts that combined both intents or belong to a different intent (as “creative” or “opinion-based”). 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. Our decision of three complexity degrees was further supported as ChatGPT consistently provided examples of prompts in different categories within only three levels of complexity. Moreover, many prompts overlapped across multiple categories and complexity levels. 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" illustrates how the “informational” category relates closely to the “analytical" category, by deepening the difficulty of the same task. Similarly, 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", could actually be comparable in complexity with the previous one (“informational”/advanced). The previous examples show how certain categories can be intrinsically more complex than others. After iterative refinement and thematic comparison, we selected 34 prompts from an initial pool of 116 used on the thematic analysis. These prompts were distributed across 4 categories: task-oriented, reflexive, both, none, and 3 complexity levels: basic, intermediate, advanced. This refined set of prompts will help control comparisons and serve as a foundation for future research on user interaction, trust, and emotional engagement with LLMs. List of prompts can be found in Supplementary Material 3.

4.3. Third Phase: Quantitative validation with end users

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). See table on Supplementary Material 4. Prompts originally labelled as “task-oriented” had a 62% confirmation rate—that is, 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%). 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%). A final set of 12 prompts was retained, each with validated complexity levels and a category confirmation rate of at least 60% for “task-oriented” and “reflexive” categories. Detailed response distribution by prompt can be found in Supplementary Material 5.

5. Discussion

This study aimed to investigate how prompts can be designed and then methodologically grouped under shared intent categories and complexity levels, based on a user-centred approach. 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 [7] as additional factors may contribute to variation (e.g. personalisation algorithms) [3], [10]. 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, the reflexive category did not reach a strong consensus. Generative models tend to be supportive without critical inspection and lack personal narratives [26], 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 [28], 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), 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, it seems that participants perceived prompts containing words as “professional life, career opportunities” not as reflexive. This could be supported by previous research [35] 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. For the first time, perceived complexity levels of prompts were assessed from a user’s perspective. Results confirmed that they were from an intermediate level. 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 and validated 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, validated through a qualitative study with lay users. This mixed methods study does not come without limitations. The sample size and set of prompts were small, resulting in a smaller set of validated prompts, which also limits the 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.

6. 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. 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. 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 prompt engineering strategies 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 by examining how categories change among different populations. By the time this research took place, the free-version ChatGPT outputs were restricted to text. As ChatGPT is now capable of handling multimodal interactions—"read" and generate images, "hear" audio and deliver voice responses—future research could explore the impact of creative tasks performed by a LLM–or other GenAI tools– on user trust, anthropomorphisation, or even their potential role in art-therapy interventions through co-creation with the LLM. Creative prompts involving image, narrative, and music generation could be further explored, as this represents a growing trend in GenAI usage.

Acknowledgements and Replicability Statement

ChatGPT 4.o free version was used to check grammar in some parts of this paper. No ideas, references, nor full written sections were obtained through ChatGPT. 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 the Supplementary Materials.

Author Contributions

CRediT: **Marijose Páez Velázquez**: Conceptualisation, Methodology, Formal Analysis, Investigation, Resources, Data Curation, Writing (Original Draft, Review & Editing), Visualisation, Project Administration. **Elzbieta Bobrowicz-Campos**:

Conceptualisation, Methodology, Writing (Review), Supervision. **Patricia Arriaga**: Conceptualisation, Methodology, Writing (Review & Editing).

References

[1] V. Andries and J. Robertson, "Alexa doesn't have that many feelings: Children's understanding of AI through interactions with smart speakers in their homes," *Comp. and Ed.: Art. Int.*, vol. 5, 2023, doi:10.1016/j.caeai.2023.100176
[2] M. C. Laupichler, A. Aster, N. Haverkamp, and T. Raupach, "Development of the "Scale for the assessment of non-experts' AI literacy" – An exploratory factor analysis," *Computers in Human Behavior Reports*, vol. 12, 2023, doi: 10.1016/j.chbr.2023.100338.
[3] S. S. Sohail, D. Ø. Madsen, Y. Himeur, and M. Ashraf, "Using ChatGPT to navigate ambivalent and contradictory research findings on artificial intelligence," *Frontiers in Art. Intell.*, vol. 6, Jul. 2023, doi:10.3389/frai.2023.1195797
[4] A. Oliveira, *A Inteligência Artificial Generativa*. Lisbon: Fundação Francisco Manuel dos Santos, 2025.
[5] M. Johnsen, *Developing AI Applications With Large Language Models*. 2025.
[6] J. Liu, C. Liu, P. Zhou, R. Lv, K. Zhou, and Y. Zhang, "Is ChatGPT a Good Recommender? A Preliminary Study" 2023
[7] S. Abdurahman, A. Salkhordeh Ziabari, A. K. Moore, D. M. Bartels, and M. Dehghani, "A Primer for Evaluating LLMs in Social-Science Research," *Advances in Methods and Practices in Psychological Science*, vol.8, no.2, 2025, doi: 10.1177/25152459251325174.
[8] G. Orrù, A. Piarulli, C. Conversano, and A. Gemignani, "Human-like problem-solving abilities in large language models using ChatGPT," *Frontiers in Artif. Intell.*, vol. 6, 2023, doi: 10.3389/frai.2023.1199350.
[9] T. Kwa, B. West, J. Becker, A. Deng, K. Garcia, M. Hasin, S. Jawhar, M. Kinniment, N. Rush, S. Von Arx, R. Bloom, T. Broadley, H. Du, B. Goodrich, N. Jurkovic, L. Miles, S. Nix, T. Lin, N. Parikh, D. Rein, L. Sato, H. Wijk, D. Ziegler, E. Barnes, L. Chan, "Measuring AI Ability to Complete Long Tasks," Mar. 2025.
[10] R. Ronge, M. Maier, and B. Rathgeber, "Towards a Definition of Generative Artificial Intelligence," *Philosophy & Technology*, vol. 38, no. 1, 2025, doi: 10.1007/s13347-025-00863-y.
[11] M. Castro, L. Lisboa, and A. Barcaui, "Beyond the Code: Understanding Professional Users' Perspectives on AI Implementation," in *Proc. 10th World Congr. on EECSS*, Avestia Publishing, Aug. 2024. doi: 10.11159/mhci24.111
[12] L. Rainie, "Close encounters of the AI kind: The increasingly human-like way people are engaging with language models," 2025. https://imaginingthedigitalfuture.org/wp-content/uploads/2025/03/ITDF-LLM-User-Report-3-12-25.pdf
[13] S. Altman and C. Anderson, "OpenAI's Sam Altman Talks ChatGPT, AI Agents and Superintelligence," YouTube. Accessed: May, 2025 [Online Video]. Available: https://www.youtube.com/watch?v=5MWT_doo68k.
[14] Z. Elyoseph, D. Hadar-Shoval, K. Asraf, and M. Lvovsky, "ChatGPT outperforms humans in emotional awareness evaluations," *Frontiers in Psychology*, vol. 14, 2023, doi: 10.3389/fpsyg.2023.1199058.
[15] J. L. Steele, "To GPT or not GPT? Empowering our students to learn with AI," *Computers and Education: Artif. Intell.*, vol. 5, 2023, doi: 10.1016/j.caeai.2023.100160.
[16] F. Hanum Siregar, B. Hasmayni, and A. H. Lubis, "The Analysis of Chat GPT Usage Impact on Learning Motivation among Scout Students," *Int. Journal of Research and Review*, vol. 10, no. 7, 2023, doi: 10.52403/ijrr.20230774.
[17] A. Habibi, M. Muhaimin, B. K. Danibao, Y. G. Wibowo, S. Wahyuni, and A. Octavia, "ChatGPT in higher education learning: Acceptance and use," *Computers and Education: Artif. Intell.*, vol. 5, 2023, doi:10.1016/j.caeai.2023.100190
[18] Y. Zheng, "ChatGPT for Teaching and Learning: An Experience from Data Science Education," in *SIGITE 2023 - Proc. of the 24th Annual Conf. on Information Technology Education*, 2023. doi: 10.1145/3585059.3611431.
[19] E. Brynjolfsson, D. Li, and L. Raymond, "Generative Ai at Work," *SSRN El. Journal*, 2023, doi:10.2139/ssrn.4426942
[20] J. Phang, M. Lampe, L. Ahmad, S. Agarwal, C. Fang, A. Liu, V. Danry, E. Lee, S. Chan, P. Pataranutaporn, P. Maes, "Investigating Affective Use and Emotional Well-being on ChatGPT," 2025. [Online]. Available: https://cdn.openai.com/papers/15987609-5f71-433c-9972-e91131f399a1/openai-affective-use-study.pdf

[21] C. Fang, A. Liu, V. Danry, E. Lee, S. Chan, P. Pataranutaporn, P. Maes, J. Phang, M. Lampe, L. Ahmad, S. Agarwal, "How AI and Human Behaviors Shape Psychosocial Effects of Chatbot Use: A Longitudinal Controlled Study," 2025.

[22] S. Chandra, A. Shirish, and S. C. Srivastava, "To Be or Not to Be …Human? Theorizing the Role of Human-Like Competencies in Conversational Artificial Intelligence Agents," *Journal of Management Information Systems*, vol. 39, no. 4, 2022, doi: 10.1080/07421222.2022.2127441.

[23] R. P. Bagozzi, M. K. Brady, and M. H. Huang, "AI Service and Emotion," *Journal of Service Research*, vol. 25, no. 4. 2022. doi: 10.1177/10946705221118579.

[24] Y. Xu, J. Zhang, and G. Deng, "Enhancing customer satisfaction with chatbots: The influence of communication styles and consumer attachment anxiety," *Frontiers in Psychology*, vol. 13, 2022, doi: 10.3389/fpsyg.2022.902782.

[25] C. Pelau, C. Volkmann, M. Barbul, and I. Bojescu, "The Role of Attachment in Improving Consumer-AI Interactions," *Proc. of the Int. Conf. on Bus. Excellence*, vol. 17, no. 1, 2023, doi: 10.2478/picbe-2023-0097.

[26] K. Saha, Y. Jain, C. Liu, S. Kaliappan, and R. Karkar, "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*, Jan. 2025, doi: 10.1145/3709366.

[27] K. T. Pham, A. Nabizadeh, and S. Selek, "Artificial Intelligence and Chatbots in Psychiatry," *Psychiatric Quarterly*, vol. 93, no. 1. 2022. doi: 10.1007/s11126-022-09973-8.

[28] M. Skjuve, A. Følstad, K. I. Fostervold, and P. B. Brandtzaeg, "My Chatbot Companion - a Study of Human-Chatbot Relationships," *Int. Journal of Human Computer Studies*, vol. 149, 2021, doi: 10.1016/j.ijhcs.2021.102601.

[29] A. Purington, J. G. Taft, S. Sannon, N. N. Bazarova, and S. H. Taylor, "Alexa is my new BFF: Social roles, user satisfaction, and personification of the Amazon Echo," *Conf. on Hum Facts in Comp Sys*, 2017: 10.1145/3027063.3053246

[30] R. E. Guingrich and M. S. A. Graziano, "Chatbots as social companions: How people perceive consciousness, human likeness, and social health benefits in machines", 2023, doi: 10.1093/9780198945215.003.0011.

[31] V. Sanh, A. Webson, C. Raffel, S. Bach, L. Sutawika, Z. Alyafeai, A. Chaffin, A. Stiegler, T. Le Scao, A. Raja, M. Dey, M. Bari, C. Xu, U. Thakker, S. Sharma, E. Szczechla, T. Kim, G. Chhablani, N. Nayak, D. Datta, J. Chang, M. Jiang, H. Wang, M. Manica, S. Shen, Z. Yong, H. Pandey, M. McKenna, R. Bawden, T. Wang, T. Neeraj, J. Rozen, A. Sharma, A. Santilli, T. Fevry, J. Fries, R. Teehan, T. Bers, S. Biderman, L. Gao, T. Wolf, A. Rush A., "Multitask Prompted Training Enables Zero-Shot Task Generalization," in *ICLR 2022 - 10th Int. Conf. on Learning Representations*, 2022.

[32] J. WhitE, Q. Fu, S. Hays, M. Sandborn, C. Olea, H. Gilbert, A. Elnashar, J. Spencer-Smith, D. Schmidt, "A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT," Feb. 2023.

[33] S. Geng, S. Liu, Z. Fu, Y. Ge, and Y. Zhang, "Recommendation as Language Processing: A Unified Pretrain, Personalized Prompt & Predict Paradigm" *Proc. 16th ACM Conf. on Reco Systems*, 2022. doi: 10.1145/3523227.3546767

[34] S. K. Santu and D. Feng, "TELeR: A General Taxonomy of LLM Prompts for Benchmarking Complex Tasks" 2023.

[35] H. Subramonyam, R. Pea, C. Pondoc, M. Agrawala, and C. Seifert, "Bridging the Gulf of Envisioning: Cognitive Challenges in Prompt Based Interactions with LLMs," *Conf. on Human Factors in Computing Systems Proc.*, 2024 doi:10.1145/3613904.3642754

[36] S. Mondal, S. D. Bappon, and C. K. Roy, "Enhancing User Interaction in ChatGPT: Characterizing and Consolidating Multiple Prompts for Issue Resolution," 2024.

[37] A. Bodonhelyi, E. Bozkir, S. Yang, E. Kasneci, and G. Kasneci, "User Intent Recognition and Satisfaction with Large Language Models: A User Study with ChatGPT," 2024.

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

[39] M. Dowling and B. Lucey, "ChatGPT for (Finance) research: The Bananarama Conjecture," *Finance Research Letters*, vol. 53, 2023, doi: 10.1016/j.frl.2023.103662.

[40] V. Braun and V. Clarke, "Reflecting on reflexive thematic analysis," *Qualitative Research in Sport, Exercise and Health*, vol. 11, no. 4, pp. 589–597, Aug. 2019, doi: 10.1080/2159676X.2019.1628806.

[41] M. Körber, “Theoretical considerations and development of a questionnaire to measure trust in automation,” in *Adv. in Intell. Systems and Computing*, 2019. doi: 10.1007/978-3-319-96074-6_2.
[42] M. Vizcaino, M. Buman, C. T. Desroches, and C. Wharton, “Reliability of a new measure to assess modern screen time in adults,” *BMC Public Health*, vol. 19, no. 1, 2019, doi: 10.1186/s12889-019-7745-6.