International Journal of Human–Computer Interaction

ISSN: 1044-7318 (Print) 1532-7590 (Online) Journal homepage: www.tandfonline.com/journals/hihc20

PromptPilot: Exploring User Experience of Prompting with AI-Enhanced Initiative in LLMs

Soomin Kim, Jinsu Eun, Yoobin Elyson Park, Kwangwon Lee, Gyuho Lee & Joonhwan Lee

To cite this article: Soomin Kim, Jinsu Eun, Yoobin Elyson Park, Kwangwon Lee, Gyuho Lee & Joonhwan Lee (22 May 2025): PromptPilot: Exploring User Experience of Prompting with AI-Enhanced Initiative in LLMs, International Journal of Human–Computer Interaction, DOI: 10.1080/10447318.2025.2489030

To link to this article: https://doi.org/10.1080/10447318.2025.2489030

INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION
https://doi.org/10.1080/10447318.2025.2489030

PromptPilot: Exploring User Experience of Prompting with AI-Enhanced Initiative in LLMs

Soomin Kim[a]*, Jinsu Eun[b]*, Yoobin Elyson Park[a], Kwangwon Lee[c], Gyuho Lee[a], and Joonhwan Lee[a,c]

[a]Department of Communication, Seoul National University, Seoul, Korea; [b]Institute of Convergence Science and Technology, Seoul National University, Seoul, Korea; [c]Interdisciplinary Program in Artificial Intelligence, Seoul National University, Seoul, Korea

ABSTRACT

Large language models (LLMs) enhance productivity and creativity, but many users struggle to formulate appropriate prompts, discouraging consistent usage. We introduce PromptPilot that assists users by recommending context-appropriate prompts based on task types and the user input. We evaluated PromptPilot through an online experiment using a 3×3 mixed factorial design. The study involved 273 participants and examined three initiative conditions (AI-initiative, mixed-initiative, user-initiative) as a between-subjects variable, across three distinct task types (browsing, daily ideation, brainstorming) as a within-subjects variable. We found that the AI-initiative and mixed-initiative systems yielded superior performance results compared to the user-initiative system. Notably, participants in the mixed initiative generated prompts using fewer words compared to those in the AI and user-initiative. The proportion of AI-generated prompts in the AI-initiative was 2.3 times that of the mixed-initiative. We discuss implications for user interaction where AI can support users' prompting process.

KEYWORDS

Generative AI; human-AI interaction; large language model; prompt engineering

1. Introduction

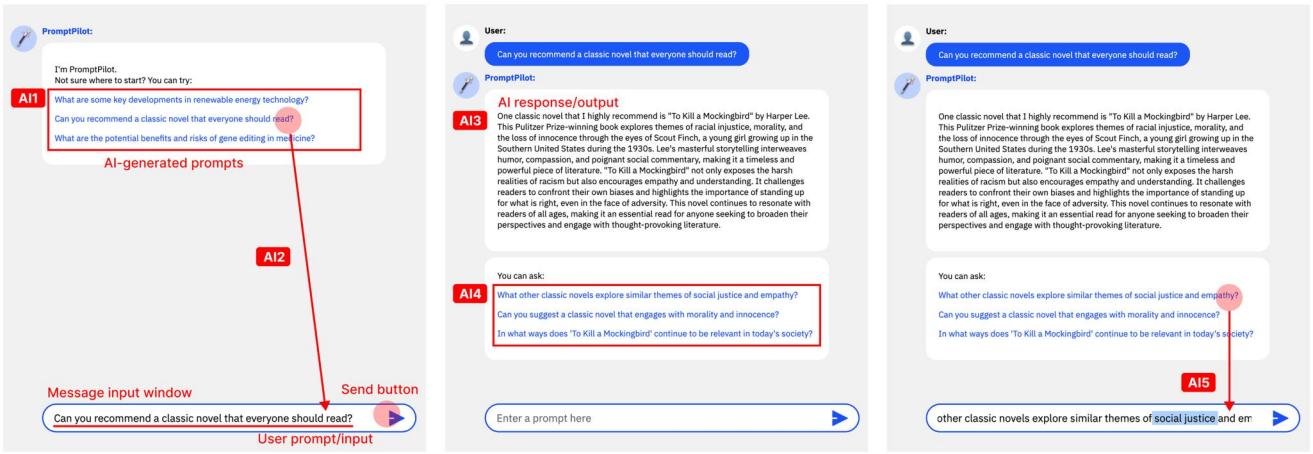
AI is permeating various fields of human life, and the rapid development of agents that apply large language models (LLMs) like ChatGPT and Gemini is fostering this proliferation. As these models transition from specialized applications to widespread public adoption, their influence continues to expand. LLMs are applied in a variety of areas, from enhancing productivity (Nijkamp et al., 2022; Petridis et al., 2023; Ross et al., 2023) to facilitating creativity (Shakeri et al., 2021; Yuan et al., 2022). As users and LLMs engage in increasingly novel interactions, understanding the user experience with LLMs has emerged as a critical issue within HCI community.

As the adoption of LLMs increases, the challenges of designing effective interactions between AI and users become more pronounced. One representative user challenge is the art of "prompting"—constructing queries or statements to derive appropriate and intended responses from the models (Liu et al., 2023). In particular, individuals without specialized AI expertise could face greater difficulties in prompt creation than experts (Arawjo et al., 2024; Mishra et al., 2023; Zamfirescu-Pereira et al., 2023). This challenge is not driven by the models' capabilities but rather arises from the high cognitive load imposed by the complexity of formulating effective prompts. Indeed, users often face substantial cognitive effort in identifying and paraphrasing verbal expressions to generate appropriate prompts (Jiang et al., 2022; Mishra & Nouri, 2022). This obstacle can discourage a broader range of users from fully leveraging these models, thereby reducing both user engagement and consistent usage (Reuters, 2023).

Cognitive Load Theory (CLT) provides a theoretical framework for understanding these challenges by emphasizing extraneous load—the additional burden imposed by suboptimal information presentation or interface design (Sweller, 1988). In domains such as web searches and learning, high extraneous load, which occurs when users must sift through irrelevant information, has been shown to impair performance (Sundararajan & Adesope, 2020; Sweller, 2011, 2024). Similarly, when interacting with LLM systems, ambiguous interfaces or unstructured prompts can create unnecessary mental burden that distracts users from their actual task goals (Zamfirescu-Pereira et al., 2023). Research in human-computer interaction has demonstrated that reducing extraneous load through improved system design and clear guidance can enhance learning and task performance (Chen et al., 2017). From this perspective, contextually relevant, AI-generated prompt suggestions could help users focus their cognitive resources more effectively.

Building on the understanding that reducing extraneous cognitive load can significantly enhance performance, it becomes imperative to reconsider the traditional user-initiated framework commonly employed in existing LLM systems (Horvitz, 1999). In the majority of current LLM systems, users take initiative in the prompt creation, with AI merely responding to their requests. Alternative paradigms,

CONTACT Joonhwan Lee joonhwan@snu.ac.kr Seoul National University, Seoul, Korea
*These authors contributed equally to this work.



(A) AI-initiative PromptPilot

(B) Mixed-initiative PromptPilot

Figure 1. The workflow for generating prompts and responses of PromptPilot. In the (A) AI-initiative condition, prompts are created based on task characteristics (AI1) with suggested follow-ups (AI4). In the (B) mixed-initiative condition, PromptPilot uses user input for prompt generation (MX1-3) and allows continuous Magic Wand use (MX6).

such as *AI-initiative* or *mixed-initiative* approaches where the AI either leads the interaction or collaboratively engages with the user, offer potential solutions (Horvitz, 2007). Under the AI-enhanced initiative paradigm, LLMs can take a proactive role by providing guidance and suggestions in creating prompts. These approaches can mitigate the cognitive demands on users, provoke inspiration, and enhance interactivity, enabling more effective use of AI capabilities. Nonetheless, it is important to note that the effectiveness of these approaches may vary depending on the task's nature and complexity; calibration of initiative between the user and AI could be necessary in optimizing the interaction (Oh et al., 2018).

In response to these challenges and opportunities, we introduce "PromptPilot," an LLM-based agent designed to support users' prompt generation by fostering the initiative of AI. Rather than leaving users to explore the complexities of prompt creation on their own, PromptPilot proactively suggests appropriate prompts. These suggestions are contextually curated based on the task characteristics and users' specific inputs (keywords and/or phrases). While LLMs have diverse applications, we focus specifically on supporting typical users with minimal prompt engineering expertise in their daily interactions with these systems. Our goal is to explore how AI-assisted prompting can support individuals who may lack advanced experience in crafting effective prompts. Figure 1 presents the AI-enhanced prompt generation feature of PromptPilot. This feature was implemented through two distinct approaches. In the AI-initiative system, PromptPilot suggests task-appropriate prompts and also formulates follow-up prompts that users might consider based on the AI's response (Figure 1(A)). Meanwhile, the mixed-initiative system generates prompts based on both the task's characteristics and the keywords or phrases provided by the user. Throughout their interaction, users engage with PromptPilot via the "magic wand" feature (Figure 1(B)). PromptPilot uses the GPT-3.5-turbo model for prompt and response generation (OpenAI, 2023b).

To evaluate the effectiveness of PromptPilot, we conducted an online experiment with 273 participants, examining three types of initiatives (user-initiative, AI-initiative, mixed-initiative) and three distinct task types (browsing, daily ideation, brainstorming). These tasks were selected as they represent frequent scenarios where users might benefit from prompting assistance in their daily planning and ideation activities. Both quantitative and qualitative methods

were utilized for analysis. In the user-initiative condition, participants crafted their own prompts, with PromptPilot merely providing responses. Under the AI-initiative, PromptPilot generated task-appropriate prompts and then delivered responses based on either the AI-generated prompts chosen by users or those manually provided by them. The mixed-initiative had both the user and AI collaboratively creating the prompt, with PromptPilot generating prompts based on keywords and phrases entered by the participants. We measured performance (output quality), user behavior (the number of words and unique terms used in prompts, acceptance and trial rates of AI-prompt generation feature), and user perceptions (satisfaction, usefulness). We also conducted a qualitative analysis by analyzing the open-ended responses. The results indicate the following:

- Even with no difference in perceived satisfaction and usefulness, users in both the AI-initiative and mixed-initiative conditions generated better outputs than those in the user-initiative condition.
- Users in the mixed-initiative condition created more concise prompts. The feature of creating prompts based on user input enhanced users' self-articulation process.
- The AI-initiative condition had an acceptance rate 2.3 times greater than the mixed-initiative condition.

To summarize, our study contributes:

- We designed and developed a system leveraging LLM to aid users in the prompt creation process. This system is specifically tailored to support general users with minimal or no prior experience with LLMs, rather than expert users familiar with prompt engineering techniques.
- Using both quantitative and qualitative methods, we present findings regarding user perception and behavior. We identified a significant influence of the degree of initiative between the AI and the user during prompt generation.
- We discuss the implications of interactions wherein AI can support and collaborate with users in generating prompts.

2. Related work

Improving users' prompting has been a focus of HCI research since the inception of generative AI and LLMs. This research on supporting users' prompt creation can be reviewed from the perspective of (1) LLM applications, (2) prompting challenges and solutions and (3) the initiative and leading role among users and AI.

2.1. Applications of LLM

LLMs have gained significant popularity due to their ability to understand, generate, and manipulate natural language. With the maturation of these models, their application in diverse domains has been profound, particularly in information seeking and ideation processes.

The introduction of LLMs has transformed traditional information seeking by enhancing traditional information retrieval systems and introducing novel generative paradigms (Ai et al., 2023; Zhai, 2024; Zhu et al., 2023). Recent studies underscore several major advances. First, LLMs have been applied to refine core retrieval elements, such as query expansion, re-ranking, and document retrieval. For example, Query2Doc leverages LLMs to generate pseudo-documents for improving query clarity, boosting retrieval performance by up to 15% on benchmark datasets (Wang et al., 2023). Similarly, REPLUG introduces retrieval-augmented frameworks in which LLMs guide retrieval models to improve prediction accuracy (Shi et al., 2024). Second, LLMs have enabled a shift from passive retrieval to proactive generation in recommendation systems (Dai et al., 2023; Zhang et al., 2024). Agent4Rec simulates user interactions with recommender models, capturing preferences while exploring causal relationships (Zhang et al., 2024). Furthermore, LLM-based tools like KAR improve personalized content discovery and address cold-start issues in recommendation systems, leading to increased accuracy (Xi et al., 2024). While LLMs enhance information seeking process, researchers emphasize that LLMs should complement rather than replace traditional search engines (Zhu et al., 2023). Future LLM-based retrieval systems should combine generative capabilities with traditional search functionalities, enabling balanced interactions between precise lookup and content generation (Zhai, 2024).

In addition to information retrieval, LLMs have demonstrated significant capabilities in ideation and content generation. Research has shown that LLM-powered systems can effectively support various creative tasks, from multi-user writing collaboration to specialized content development. For example, studies of SAGA (Shakeri et al., 2021) and ABScribe (Reza et al., 2024) found that asynchronous collaboration through LLMs enabled users to effectively alternate between creation and review roles, while systems like AngleKindling (Petridis et al., 2023) and CharacterMeet (Qin et al., 2024) demonstrated how LLMs can successfully assist with specialized creative tasks such as journalistic ideation and character development. These examples illustrate how LLMs can be effectively integrated into creative workflows while maintaining human agency in the creative process.

Expanding beyond these collaborative workflows, research has quantified LLMs' broader capabilities in ideation tasks. Studies have demonstrated that LLMs can match or surpass human performance in divergent thinking (Bellemare-Pepin et al., 2024) and provide significant advantages in idea generation, with AI-generated ideas being seven times more likely to rank among the top 10% in product development contexts (Girotra et al., 2023). Building on these capabilities, researchers have developed frameworks to optimize human-LLM collaboration, from various interaction approaches (Lim & Perrault, 2024) to structured support through a three-stage process of Ideation, Illumination, and

Implementation (Wan et al., 2024). These frameworks aim to maximize LLMs' creative potential while maintaining individual users' style and authenticity in the creative process (Wasi et al., 2024).

While research investigates LLMs' capabilities in both information seeking and ideation tasks, users often struggle to effectively access these capabilities through the standardized interfaces like those found in ChatGPT, Gemini, and Claude. We investigate how AI can support users' prompting processes in everyday information search and ideation tasks, aiming to improve the current input-output interaction format. By enhancing these interactions, we seek to create more intuitive and engaging experiences that foster sustained and effective use of LLM systems.

2.2. Prompting strategies and challenges

Prompting, the act of providing textual instructions to LLMs, serves as the primary interface between users and AI systems (Liu et al., 2023). Current LLM interfaces like ChatGPT and Gemini rely on user-initiative frameworks, where users craft queries independently without system guidance (Brandtzaeg et al., 2024). While this approach allows flexibility, it demands expertise and cognitive effort that often challenges non-expert users (Arawjo et al., 2024; Mishra et al., 2023; Zamfirescu-Pereira et al., 2023).

To address these challenges, researchers and industry leaders have developed various prompting strategies. The fundamental principle, emphasized by both industry guidelines and research, is creating clear, concise prompts that reduce ambiguity and complexity (Anthropic, 2024; Google, 2024; Meta, 2024). Clear prompts reduce ambiguity and complexity, enabling the model to process input effectively (Crispino et al., 2023; Renze & Guven, 2024).

In this study, we define "prompt conciseness" as the property of a prompt that includes all essential task instructions while excluding superfluous elements and presenting a clear structure. In practice, a prompt is considered concise if its word (or token) count is low while still conveying all critical information needed for the task. This definition emphasizes that achieving prompt conciseness is not simply a matter of including as much information as possible; rather, it requires striking an optimal balance between sufficient information and brevity.

Research has shown that overly verbose prompts can introduce unnecessary noise, increase cognitive load, and ultimately impair the reasoning performance of LLMs (Levy et al., 2024). Recent empirical findings further underscore that concise prompts not only enhance the conveyance of user intent but also improve output quality. For example, Renze and Guven (2024) found that eliminating redundant language enables models to focus on the essential task, leading to more coherent and relevant responses. Similarly, research on gist compression demonstrates that stripping away unnecessary verbosity not only reduces token usage but also improves interpretability and performance (Li et al., 2024). These findings naturally extend to prompt compression techniques, which aim to condense prompts while preserving critical information, thereby optimizing input complexity and reinforcing the benefits of conciseness (Li et al., 2023; Wan et al., 2023). Complementing these findings, Joshi et al. (2024) observed that prompt engineers typically favor concise structures. Collectively, these studies suggest that by removing superfluous details, concise prompts facilitate clearer conveyance of user intent and a more focused reasoning process.

Moreover, reducing response length through concise prompt strategies can have significant cost benefits for AI systems engineers, as many third-party LLM APIs charge per token (Anthropic, 2025; OpenAI, 2025). Shorter outputs lead to lower operational costs, reduced energy consumption, and faster response times.

Beyond clarity, several advanced strategies have emerged to enhance LLM performance. Few-shot prompting, for instance, integrates a small number of input-output examples to improve results (Brown et al., 2020). Another strategy is chaining, which structures LLM outputs sequentially. This includes Chain-of-Thought (CoT) prompting for step-by-step reasoning (Wei et al., 2022) and Tree-of-Thoughts (ToT) for exploring structured reasoning paths (Yao et al., 2023). Tools like Prompt Chainer help users design and debug multi-step prompts, increasing transparency and effectiveness (Wu et al., 2022). Other strategies involve iterative refinement, allowing users to improve prompts based on feedback (Mishra et al., 2023), and methods to reduce hallucination by leveraging external knowledge (Lewis et al., 2020; Li et al., 2023).

Despite these advances, users still face significant challenges in prompt creation. Users often struggle to find the right wording or level of specificity (Skjuve et al., 2023), and biases and misconceptions can further complicate the process (Skjuve et al., 2023; Zamfirescu-Pereira et al., 2023). Most critically, two fundamental issues persist: prompt formulation uncertainty and verbosity. Prompt formulation uncertainty arises when users struggle to identify what to input, leading to vague or ineffective prompts that fail to fully utilize the model's capabilities (Zamfirescu-Pereira et al., 2023). Verbosity, on the other hand, diminishes prompt focus and precision, making it harder for the model to process and generate relevant responses (Nayab et al., 2024; Zamfirescu-Pereira et al., 2023). The limited system guidance further exacerbate these difficulties (Jiang et al., 2022).

Although these challenges in the prompting process are widely recognized, few studies have implemented and verified proactive AI assistance in prompt creation through direct system manipulation. Our research addresses this gap by investigating how LLMs can actively assist users in crafting prompts. In addition, we measure prompt conciseness as an indicator of effectiveness, aiming to understand how system interventions can improve both prompt quality and overall user experience.

2.3. Initiative and leading role among users and AI

In the design of interactions between users and AI, the aspect of initiative, or who takes the lead, is crucial. In this

respect, the mixed-initiative paradigm emphasizes an "elegant coupling" between direct user manipulation and automated interface agents (Horvitz, 1999). In support of direct manipulation, researchers argue that it gives users control and predictability over their interfaces. Conversely, those advocating interface agents contend that users should delegate certain tasks to agents. By combining the advantages of both approaches, mixed-initiative systems enable efficient collaboration towards achieving user goals (Birnbaum et al., 1997; Shneiderman & Maes, 1997). Although a number of HCI studies have addressed the issue of taking the initiative between users and AI (Ashktorab et al., 2021; Graesser et al., 2005; Nguyen et al., 2018; Oh et al., 2018), none has investigated initiative in the context of prompting generation when using LLMs.

In conversational systems, initiative indicates who—either the human or AI—takes the lead during interactions (Walker & Whittaker, 1990). Likewise, both the AI and user can take turns in guiding the discourse in LLM systems. When users browse information or perform tasks via these systems, research indicates that AI can take charge by asking clarifying questions. On one hand, researchers utilized a predefined set of questions. For example, research using the Qulac dataset revealed that a single well-crafted question related to a user's original query can amplify performance by 170% (Aliannejadi et al., 2019). Rao and Daumé III subsequently designed a model to prioritize clarification questions using StackExchange data (Rao & Daumé III, 2018). On the other hand, another body of research focuses on generating queries based on user inputs. The sequence-to-sequence model tailored for framing clarification questions has been shown to surpass retrieval-based models in terms of usefulness (Rao & Daumé III, 2019). Additionally, the efficiency of both supervised and reinforcement learning models has been confirmed (Zamani et al., 2020). In recent advancements, fine-tuned GPT-2 has been employed to generate clarifying questions (Sekulíc et al., 2021). Overall, enhancing user prompt creation in conversational systems has embraced an AI-enhanced, mixed-initiative framework, with contemporary methodologies employing LLMs such as GPT.

Drawing from prior research, we adopt a mixed-initiative approach by leveraging the capabilities of LLM to assist in the user's prompting process. Unlike the prevalent *user-initiative* interaction in LLMs, where users independently formulate prompts and AI merely responds, we introduce two alternative interactions: *AI-initiative* and *mixed-initiative*. In the *AI-initiative* approach, the AI suggests potential prompts from which users can choose. Though it's AI-initiative, it still permits a degree of user initiative by presenting them with choices. Meanwhile, the *mixed-initiative* approach allows users to input keywords or phrases, upon which the AI generates a corresponding prompt. In this approach, users take on a more leading role in prompt creation compared to the AI-initiative approach. Here, we note that adjusting the level of initiative between the user and AI is a continual process, rather than one divided into strict segments. We aim to investigate how different levels of initiative between users and AI influence user behavior and perceptions during the prompt creation process. Given this background, we aim to explore the following research questions in our paper:

- RQ1. How does the level of initiative in prompt generation between the AI and the user affect user behavior and perceptions?
- RQ2. How do the effects of initiative levels on user behavior and perceptions differ across varying task types?
- RQ3. What is the rate of acceptance for these AI-generated prompts, and what factors contribute to this acceptance rate?

3. Design of PromptPilot

To better understand the user experience of prompting during interactions with LLMs, we designed a research prototype, PromptPilot. In this section, we outline the overarching structure and the design process of PromptPilot.

3.1. Manipulation of initiative between AI and the user

In the design of PromptPilot, we considered initiative as a main factor and devised three different conditions based on the degree of initiative between the AI and the user (Horvitz, 1999). Based on who takes the lead in the interaction, we classified three conditions: (1) user-initiative, (2) AI-initiative, and (3) mixed-initiative.

3.1.1. User-initiative

In the user-initiative condition, users take the lead in the interaction. When the user enters a prompt in the message window, PromptPilot responds accordingly. Users directly manipulate an interface to invoke the system, and it passively responds to their requests (Horvitz, 1999). In its initial version, OpenAI's ChatGPT uses a direct manipulation interface.

3.1.2. AI-initiative

Here, PromptPilot plays a proactive role, suggesting prompts based on task specifics. PromptPilot suggests prompts to users considering task characteristics, and users can select from the suggestions. In its early iteration, Google's Bard employs this form of AI-driven interaction.[1]

3.1.3. Mixed-initiative

Both AI and the user can take the initiative together under the mixed-initiative condition, coupling direct manipulation and automated agents (Horvitz, 1999). PromptPilot generates and suggests prompt candidates associated with the user's entries in the Magic Wand window.

3.1.4. Chat scenarios

Upon entering the system, the user performs the tasks through a dialogue with PromptPilot. In the *user-initiative*

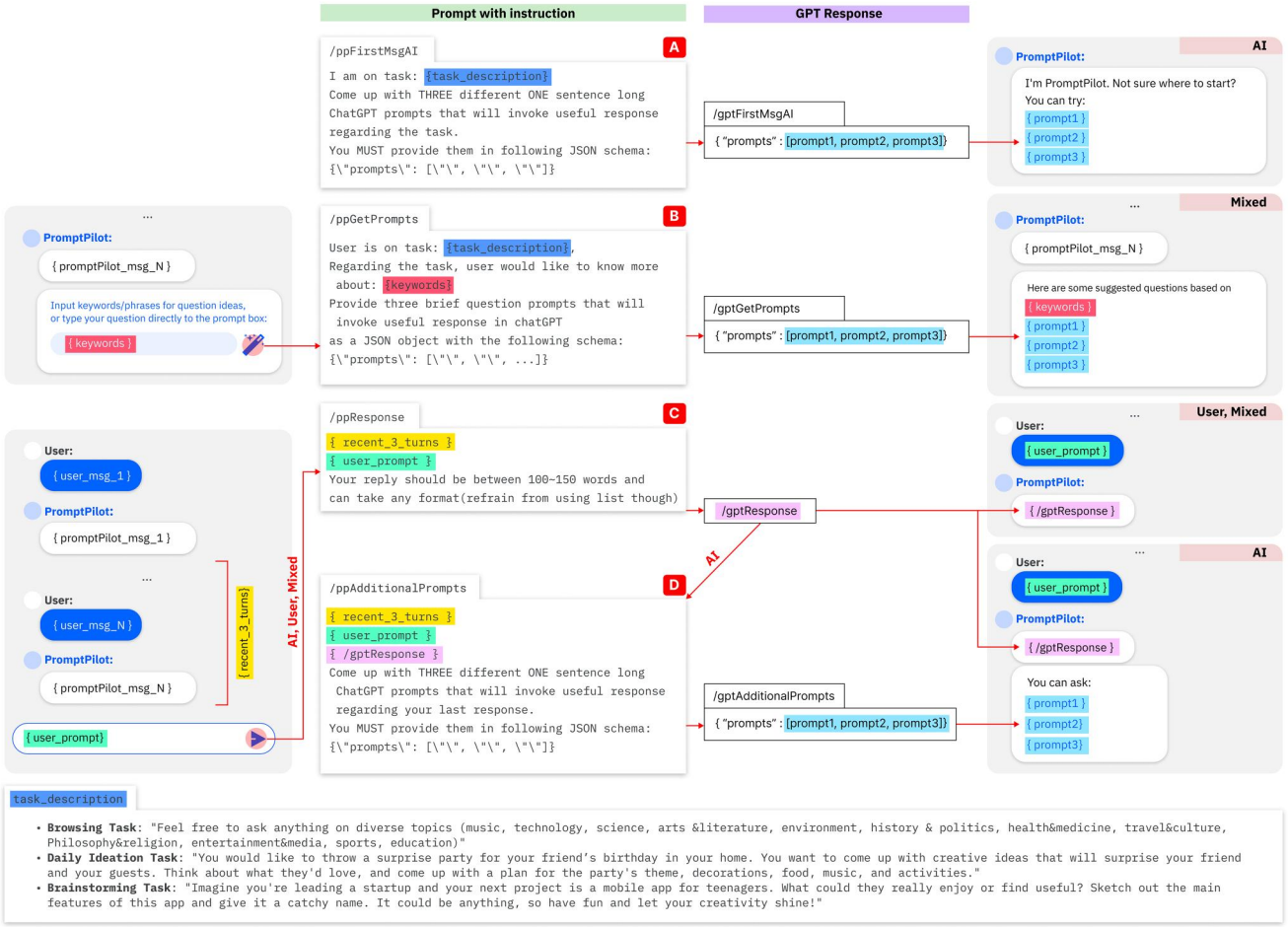


Figure 2. Overall structure for prompt and response generation. In each condition, four instructions were used: three to suggest prompts to the user (A, B, D) and one to respond to the user's input (C). In the AI-initiative condition, the system suggests a task-related prompt (A) at the beginning of the conversation. After each response, it suggests the next prompt (D) that might follow the user's response. In the mixed-initiative condition, the user types keywords, and the system suggests prompts (B) related to the task and those keywords.

system, users directly input their prompts into the message input window and send them by either clicking the "Send" button or pressing the "Enter" key. In the *AI-initiative* system, PromptPilot offers three AI-generated prompts tailored to the user's task. Users can choose one of these or input their own. If a suggested prompt is chosen, it appears in the message input field, ready for editing or sending. Once sent, PromptPilot provides the relevant response. For the *mixed-initiative* system, users can either type their prompts or use the "Magic Wand" feature, which generates based on use input (keywords or phrases). AI-suggested prompts appear in the message input box for potential editing. After submission, PromptPilot produces the corresponding response.

3.2. Prompt and response generation

In this section, we detail our iterative process of instruction engineering used for generating LLM-based prompts and responses. For prompt and response generation, we use the GPT-3.5-turbo model. We provided instructions (semantic descriptions) for each task in the user prompts to improve the relevance of the generated answers. These instructions were structured as system prompts, based on resources from Microsoft's Semantic Kernel and OpenAI's plugins supplements (Microsoft, 2023c; OpenAI, 2023a).

3.2.1. Response generation (ALL)

Responses are generated based on prior conversations and user's prompt with instruction (Figure 2(C)). The utilization of a more extensive chat history often results in responses that better reflect the context of previous conversations. However, due to the token limit of GPT-Turbo-3 (4096 tokens), we employed the last three turns of the conversation rather than the entire chat history. Additional instructions were incorporated to prevent excessively lengthy replies and to ensure the generation of readable responses.

3.2.2. Prompt generation based on task (AI-initiative)

In the AI-Initiative condition, PromptPilot presents three prompts at the start of each conversational turn before users input their own. Initially, PromptPilot suggests three task-related prompts in the form of questions, designed to guide user engagement and facilitate the task (Figure 2(A)) (Goldberg et al., 2021; Kuang et al., 2024). The instructions provided to PromptPilot focus on generating open-ended

questions that can invoke useful responses from the model, tailored to the task's context.

Users have the option to either type their own prompts or select one from the suggested list. Upon selection or input, the AI generates a response accompanied by three related questions that users might consider for their next interaction (Figure 2(C,D)). These suggested prompts are returned in JSON format, enabling their display as a list within the interface.

3.2.3. Prompt generation based on task and user input (mixed-initiative)

In the mixed-Initiative condition, when a user inputs one or more keywords or phrases, PromptPilot generates and suggests prompts related to those input, tailored to the context of the specified task (Figure 2(B)). Rather than providing direct solutions, PromptPilot primarily generates question-based prompts to encourage users to critically reflect on their tasks and further develop their ideas (Chan et al., 2016b). These generated prompts are returned in JSON format.

3.2.4. Iterative design of prompt generation

In Tables 1 and 2, we provide examples of prompts generated by our final instructions. To ensure appropriate responses from GPT across various conditions, we tailored instructions based on the task and appended them to the user's prompt. This task was undertaken iteratively by three researchers. Given the inherent unpredictability of LLMs, achieving a consistent outcome from an input prompt to GPT can be challenging. The instruction optimization process tends to be labor-intensive and manual. During this iterative fine-tuning, we encountered a series of challenges, which we discuss in detail, along with the solutions implemented.

A primary concern was the model's inconsistent response structures. Initially, we attempted to guide the model to format its outputs using a JSON schema. This schema encompassed both a list of suggested prompts and the response to the user's query. While this approach seemed promising at first, we noticed deviations in which the model would not adhere to standard JSON conventions; for instance, curly braces or double quotes might be unexpectedly omitted or inserted. Rather than combining the user's response with the list of prompt suggestions, we decided to separate the two processes. Furthermore, we simplified the JSON structure to return only the list of suggested prompts. Although this JSON-formatted list was reliable, any user inputs deviating from the expected format triggered the error message: "Something went wrong. Please try the conversation again."

Another challenge arose from LLM's difficulty in processing long conversations. Such conversations, characterized by extended exchanges between the user and the model and further compounded by the model's lengthy responses, often led to missed or misunderstood instructions. While our initial approach was to increase the volume of instructions within the prompts to improve output quality, this strategy often backfired. As we added more instructions, the model began omitting or misinterpreting key directives, leading to inaccurate outputs (Liu et al., 2023). To address this, we reduced the number of instructions and limited the model's response length. This not only ensured the efficient execution of our directives but also minimized the impact of lengthy responses on subsequent interactions, promoting smoother user-model dialogues.

During iterative experimentation with various prompts generated by LLMs, we observed significant challenges with consistency and usability. We experimented with various formats including direct recommendations, examples, and questions. Formats such as direct recommendations and examples often varied in structure and level of detail, resulting in unpredictable outputs. To address this, we adopted a question-based format for generated prompts. This format proved to be both stable and effective, consistently supporting ideation tasks by encouraging user engagement and critical thinking (Goldberg et al., 2021; Kuang et al., 2024).

Lastly, we noticed a trend in prompt generation: the fewer keywords provided by the user, the more irrelevant prompts were generated. To ensure that the prompts

Table 1. Examples of AI-generated prompts (AI-initiative).

Task	Prompt	Follow-up prompts
Browsing	What are some current trends in the music industry?	How have streaming platforms impacted the music industry? Can you give an example of a successful collaboration in the music industry? What are some strategies that artists use to engage with fans on social media?
	What are the health benefits of regular exercise?	How does exercise specifically improve cardiovascular health? What are some strategies for maintaining a healthy weight through exercise? Can you explain how exercise impacts mental health and cognitive function?
Daily	What are some unique themes for a birthday party?	How can one incorporate retro arcade games into the party? What kind of masks and attire would be suitable for a masquerade ball? Can you provide examples of clues and puzzles for a treasure hunt party?
	What are some fun activities or games that would make the party memorable?	What are some other interactive games that can be played at the party? How can I create a personalized scavenger hunt for the birthday party? Are there any unique twists I can add to the photo booth experience?
Brainstorming	Can you suggest a catchy name for a mobile app targeted at teenagers?	What are some specific features that TeensConnect could have? How does TeensConnect ensure the safety of its users? Are there any gamification elements in TeensConnect to engage teenagers?
	How can a mobile app for teenagers be both enjoyable and useful?	What are some creative ways to incorporate education into the app? How can the app promote a sense of community among teenagers? How can the study planner feature be made more engaging for teenagers?

Table 2. Examples of AI-generated prompts (mixed-initiative).

Task	User input	Prompts
Browsing	Beethoven, Symphony, Netflix	What were Beethoven's most famous symphonies? Tell me about the life and works of Ludwig van Beethoven. Can you recommend any must-watch shows or movies on Netflix?
	Existentialism, Sartre, Meaning of life	What is existentialism and how does it relate to philosophy? Who is Jean-Paul Sartre and what are his contributions to existentialism? What does existentialism say about the meaning of life?
Daily	80s Music, Retro, Instagram	What are some popular 80s songs that would be great to play at a surprise party? How can I create a retro atmosphere for the party decorations? What are some creative ways to incorporate Instagram into the surprise party?
	DIY, painting	What are some creative DIY ideas for decorations at a surprise party? How can I incorporate painting into the theme of the surprise party? Are there any painting-related activities that can be included in the party?
Brainstorming	Budgeting, finance for teenagers	What are some effective ways to teach budgeting and finance to teenagers? What are some key financial topics that should be included in a mobile app for teenagers? How can a mobile app for teenagers promote responsible financial behavior?
	Global history, cultural exchange	What are some popular mobile apps for teenagers in different countries? How can a mobile app for teenagers promote cultural exchange? What role does global history play in shaping the interests of teenagers?

generated were more aligned with the user's task, we incorporated a detailed description of each task within the instruction.

3.3. Implementation of PromptPilot

PromptPilot is implemented as a web application using React for the frontend and FastAPI for the backend. The system uses OpenAI's GPT API, specifically the gpt-3.5-turbo model, as its core language model. When a user inputs a prompt on the frontend, a request is sent to the backend. The backend logs the user input, processes it based on the specific task type and initiative condition, formulates an appropriate query for the GPT API, and returns the processed response to the frontend. This architecture supports different initiative conditions and task types while enabling efficient processing and real-time interactions. The system employs asynchronous processing, input debouncing, and security measures to ensure optimal performance and data protection. This architecture design allows for flexibility and integration of strategies to recommend relevant questions based on tasks and the system's initiative level. By dynamically adjusting to task characteristics and user inputs, PromptPilot supports prompt formulation. In AI-initiative mode, it provides automated suggestions, while in mixed-initiative mode, it generates targeted questions based on user input.

4. Method

4.1. Study design

This study used a 3 × 3 mixed factorial design, with the degree of initiative between the AI and the user (user-initiative, AI-initiative, mixed-initiative) as a between-subjects variable and task type (browsing, daily ideation, brainstorming) as a within-subjects variable. This design is particularly appropriate for studying human-AI interaction as it enables analysis of both individual variations in response to AI-initiative levels (between-subjects) while efficiently measuring how users adapt across different tasks (within-subjects). This mixed factorial approach is widely used in HCI research because it allows comparison between independent groups while controlling for individual differences, providing a robust way to examine both user variations and task effects (Kraus et al., 2020; Oh et al., 2018). Participants interacted with one of the three initiative conditions and completed three distinct tasks. We randomly assigned participants to one of these three conditions. Also, the tasks were presented in a random sequence to avoid order and carry-over effects.

4.2. Participants

Participants were recruited via Prolific and participated in an online experiment. Prolific is an online platform that connects researchers with participants for academic studies. We recruited a total of 302 participants for our study ($M_{age} = 39.9$, $SD_{age} = 11.8$; 51% female). Given that expertise or frequency of LLM usage could affect the user experience, we excluded individuals working in the AI domain and those who use LLMs daily. Our target was individuals with minimal or no experience with LLMs. This focus on general users ensures that the findings are relevant to a broad audience, including novice and casual users who are most likely to benefit from AI-augmented prompting systems. Additionally, in order to control the usage environment, participants were required to use either a laptop or desktop for the study.

4.3. Procedure

Figure 3 depicts the research procedure of our study. Upon accessing the online experiment platform, participants were briefed about the study's objectives and system's capabilities. During the briefing, participants were informed about the expected duration of the experiment, the types of tasks they would perform, and the system's capabilities. Specifically, they were informed that the system generates responses based on the last three turns of conversation, reflecting current model capabilities. They also reassured that their responses would remain anonymous and confidential. After

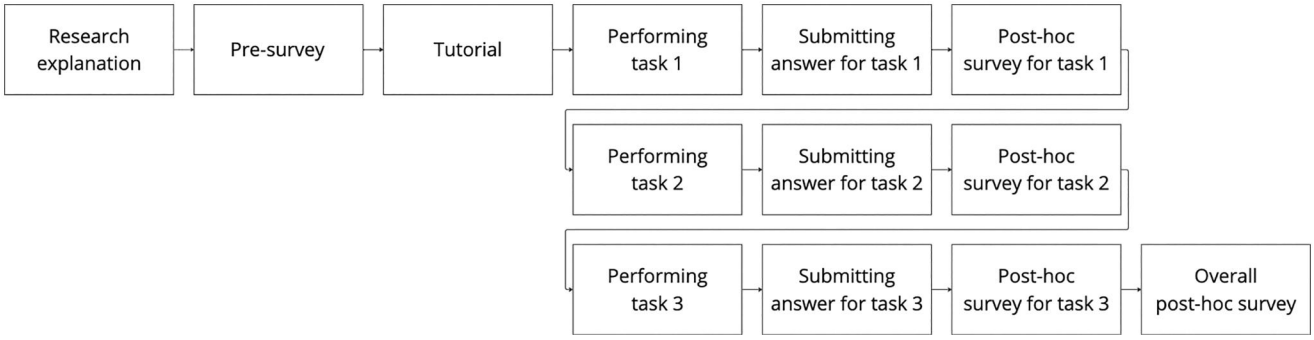


Figure 3. Overall research process. Participants completed three tasks, submitted responses after each, and took a survey. After completing all tasks, they responded to open-ended questionnaires about their overall experience. To counteract any order effects, tasks were presented in a random sequence.

being informed of these details, participants provided their consent.

They first completed a pre-survey that collected demographic information including age, gender, and occupation. They also completed questionnaires about their experience and expertise in LLMS, which were used to filter out unsuitable participants. Following this, they were introduced to a tutorial explaining PromptPilot's operations before proceeding to the three designated tasks.

Participants undertook three tasks. To mitigate potential order effects, the sequence of the three tasks was randomly assigned to participants. Each task began with a brief introduction and specified topics participants needed to address upon task completion. For example, participants were informed that their goal was to design an app for teenagers during the brainstorming task. They were also told that they should provide an app name and its key features at the end of the task. Following these instructions, participants initiated a dialogue with PromptPilot. They could chat back and forth for a minimum of 3 times and a maximum of 20 times. After the third exchange, an "End Chat" button became accessible, allowing participants to either conclude or continue the conversation. When a dialogue reached the 20-exchange threshold, the message input window automatically deactivated, concluding the task.

Once each task was completed, participants were shown their conversation history with PromptPilot. Using this as a reference, they offered answers (task output) relevant to the task's subject matter. Subsequently, they responded to a *post-hoc* survey, comprising six items (two each for initiative, satisfaction, and usefulness).

4.4. Task

We examined how initiative types influenced user experience across three distinct task types commonly supported by LLMs: browsing, daily ideation, and brainstorming (Wu et al., 2022). Each task differed in its primary goal, the degree of personal relevance, and the tangibility of its expected outcome. The tasks were arranged along a continuum, from open-ended information exploration to practical, personalized planning, and finally to complex, high-level conceptualization.

4.4.1. Browsing task (exploratory, information-seeking)

In the browsing task, participants engaged in open-ended conversations with PromptPilot, exploring topics of their interest (Xu et al., 2023). The main goal was broad exploration and information discovery, with no requirement to produce a specific deliverable. This scenario resembles initial encounters with systems like ChatGPT or Gemini, where users seek understanding and insights without a concrete endpoint. At the end of the conversation, participants were asked to provide insights or takeaways derived from their interaction.

4.4.2. Daily ideation task (personalized, practical creativity)

In the daily ideation task, participants addressed a concrete, everyday challenge within a personal or social context: planning a surprise party for a friend (Chavula et al., 2022). This required creating a practical, personally relevant outcome, integrating personal preferences and real-world details (e.g., selecting activities, arranging materials), alongside creative thinking. This scenario reflects contexts where LLMs support everyday ideation and planning tasks (Google, 2023; Microsoft, 2023a, 2023b). Participants submitted plans for the surprise party after their conversation.

4.4.3. Brainstorming task (complex, high-level planning)

The brainstorming task asked participants to design a mobile app concept for teenagers, thus demanding advanced thinking that balance both conceptual and tangible elements (Chan et al., 2016a). Unlike the daily ideation task, which drew on common experience, this scenario involved integrating advanced knowledge, considering technical feasibility, and addressing potential market needs. Such a context reflects situations in which LLMs contribute to complex problem-solving and innovation, like product development or strategic brainstorming (Google, 2023; Microsoft, 2023a, 2023b). Participants submitted the app names and key features at the end of the conversation.

4.5. Measures

We evaluated PromptPilot based on three key dimensions: (1) performance, (2) user behavior (prompt conciseness and

adoption rate), and (3) user perceptions (Likert-based surveys and open-ended responses). Our data collection comprised four types: chat logs, task responses, quantitative survey data, and qualitative survey data. The chat log data included task IDs, sender IDs, timestamps, and the message content (including user prompts, PromptPilot-generated prompts and PromptPilot's responses). We measured performance by evaluating the task output. Behavioral patterns were analyzed through word and unique term counts. Both the survey data and open-ended responses provided insights into users' perceptions of PromptPilot.

4.5.1. Performance (output quality)

After completing each task, participants were required to respond to predetermined questions. To assess the quality of their performance, we systematically evaluated their task outputs. From each condition, we randomly selected responses from 15 participants, resulting in an evaluation set of 135 answers across three tasks from a total of 45 participants. We recruited three independent human evaluators to score the 135 outputs using a 10-point scale, with a higher score indicating superior response quality. The inter-rater reliability was significant for these ratings (Krippendorff's $\alpha = 0.92$).

4.5.2. Prompt conciseness

4.5.2.1. The number of words (tokens). We evaluate the verbosity or conciseness of user prompts by counting the token (word) count within user prompts. This metric assesses adherence to the fundamental principle of concise prompting, which is theoretically supported by recent studies on input length (Levy et al., 2024; Renze & Guven, 2024). A lower word (token) count often indicates more precise and clear prompt formulation, while higher counts may suggest less focused expressions that could impact prompt effectiveness.

4.5.2.2. The number of unique terms. We also measure the number of unique terms by counting the distinct normalized terms within the prompts to assess the clarity and focus of user prompts. A normalized term is a token that has had stopwords removed and has been lemmatized. We gathered all the prompts produced during the completion of a task and tailed the distinct normalized terms. For this purpose, we employed the word_tokenize, stopwords, and WordNetLemmatizer functions from the nltk library in Python.

4.5.3. Adoption rate

4.5.3.1. Acceptance rate of AI-generated prompts (AI-initiative and mixed-initiative). This refers to the percentage of instances in which users opt for the AI-generated prompts within the AI-initiative and mixed-initiative conditions. A high acceptance rate indicates that users find the AI's prompts relevant and engaging. Conversely, a lower rate suggests a mismatch between AI suggestions and user preferences or intentions.

4.5.3.2. Trial rate of Magic Wand function (mixed-initiative). This refers to the percentage of trials during which users input keywords or phrases in the Magic Wand box in the mixed-initiative condition.

4.5.4. Likert-based surveys

After completing each task, participants answered questionnaires assessing initiative (Oh et al., 2018), satisfaction, and usefulness (Lund, 2001). In order to verify the manipulation of the system, initiative was used; satisfaction and usefulness were used as dependent variables. Each variable was measured using two items. For initiative, they responded to "PromptPilot takes the leading role while conducting a task," and "PromptPilot usually establishes the task direction." Regarding satisfaction, they responded to "I am satisfied with PromptPilot," and "It is pleasant to use PromptPilot." For usefulness, they reflected on "PromptPilot helps me be more effective," and "PromptPilot helps me be more productive." These items were rated on a Likert scale of 1 (strongly disagree) through 7 (strongly agree).

4.5.5. Open-ended responses

We also adopted a qualitative method through open-ended surveys to achieve a more comprehensive insight into the user experience with PromptPilot. Participants were asked about their experience interacting with PromptPilot. For example, participants were asked: "What were the positive aspects or highlights of your experience using PromptPilot?," "What challenges or frustrations did you experience while using the system?," "What specific aspects of PromptPilot would you suggest improving?," "How likely are you to use PromptPilot in the future, and why?," and "What are your thoughts on the AI-assisted prompt generation feature?" These questions were deliberately open-ended to allow participants to express their experiences, concerns, and suggestions without constraint, offering insights into both the benefits and limitations of the system.

4.6. Analysis

We gathered four types of data: task output, dialogue data, quantitative data (Likert-based surveys) and qualitative data (open-ended responses). Quantitative analysis was performed on the task output, dialogue and quantitative data, while qualitative analysis was performed on the qualitative data.

For RQ1 and RQ2, we aimed to investigate both the main effect of the initiative and its interaction effect with task types. A repeated measures ANOVA was employed to assess these effects. Furthermore, to examine the simple effect of the initiative for each task, we conducted a one-way ANOVA (RQ2). Given the diversity of tasks that can be undertaken in LLMs, it is reasonable to examine the effects

of initiative for each task separately. Such nuanced effects might be overlooked when only considering aggregate effects. For our exploratory data analysis and ANOVA, we employed the Pingouin package in Python and the standalone software JASP (Pingouin, 2024). Both tools utilize validated statistical libraries in Python and R and automatically apply adjustments (e.g., the Tukey–Kramer method) to accommodate unequal sample sizes in post-hoc tests. For RQ3, we evaluated the adoption rate of AI-generated prompts within the mixed-initiative and AI-initiative conditions. We also conducted a cross-tabulation analysis to verify statistical significance.

For the qualitative responses, we analyzed the open-ended responses using the grounded theory approach (Glaser & Strauss, 2017), while also drawing on Braun and Clarke's reflexive approach to thematic analysis (Braun & Clarke, 2013). This analysis was conducted in three stages. First, two researchers collaboratively reviewed the organized data, exchanging insights about the main findings from the experiments. This procedure was repeated three times, with each iteration refining their understanding and establishing a shared interpretative framework.

Next, we used Reframer, a software for qualitative research, to perform keyword tagging and identify themes. Original responses were dissected into individual sentences, resulting in 1,229 observations. Following Braun and Clarke's guidance on the organic development of themes (Braun & Clarke, 2019), we engaged in collaborative coding sessions where each meaning unit was annotated with multiple keywords to capture its central ideas. This systematic coding process generated 218 unique keyword tags. Through an iterative process involving three rounds of discussions, we synthesized these tags to develop broader themes. In instances of disagreement, a third researcher was consulted to reach consensus, ensuring the robustness of our interpretations while acknowledging the inherently subjective nature of qualitative analysis (Creswell & Poth, 2016). This process yielded 35 distinct themes.

Finally, we refined, interconnected, and merged these themes into seven main categories. Among these, we focus on five categories that align directly with our research questions. The remaining two categories, which addressed broader implications for AI system design and general attitudes toward AI assistance. Following Braun and Clarke's justification in their reflexive approach (Braun & Clarke, 2013), we did not conduct intercoder reliability testing, as our emphasis was on collaborative interpretation rather than on quantifying interrater agreement. The results of this process provided a deeper understanding of the participants' experience with PromptPilot.

5. Results

In this section, we describe the result of our analysis on quantitative results, followed by the results of qualitative survey analysis.

5.1. Manipulation check for initiative

We defined three conditions based on the degree of initiative between the user and AI: AI-initiative, mixed-initiative, and user-initiative. A one-way ANOVA revealed significant differences in perceived initiative across the three conditions ($F(2,270) = 5.23$, $p = 0.01$). Subsequent *post-hoc* *t*-tests showed that participants perceived the AI's level of initiative as significantly higher in the order of AI-initiative, mixed-initiative, and then user-initiative. This result implies that three different versions of PromptPilot were adeptly designed to manifest distinct initiative levels.

5.2. Descriptive analysis

A total of 273 participant data were statistically analyzed. From the recruited 302 participants recruited, 8 participants who did not complete the experiment were excluded, and 21 participants were deemed to have given insincere responses and were excluded from analysis (user-initiative: 12; mixed-initiative: 6; AI-initiative: 3). Finally, 93 people in the AI-initiative condition, 93 people in the mixed-initiative condition, and 87 people in the user-initiative condition are the subjects of the final analysis.

Regarding time durations, no significant differences were observed between the conditions. On average, participants took 41 min and 47 s ($SD = 5$ min 9 sec) to complete the three tasks. Also, no significant difference was observed in the number of turn-taking ($F(2,42) = 0.257$, $p = 0.77$). The average number of turn-taking is 4.1 for user-initiative, 4.2 for mixed-initiative, and 4.3 for AI-initiative. Participants in all conditions exchanged approximately four pairs of conversations with PromptPilot.

5.3. Main effect of AI-initiative on user experience (RQ1)

Research question 1 focuses on the influence of AI-initiative on user behavior and perceptions. Table 4 demonstrates the main effects for both initiative and task type based on the results from the repeated measures ANOVA. The results indicate that participants in the AI-initiative and mixed-initiative conditions consistently outperformed those in the user-initiative condition in terms of output quality. While perceived satisfaction and usefulness did not differ significantly across conditions, behavioral measures such as word count and unique term count revealed notable differences. Figure 4 presents a graph highlighting the impact of initiative conditions on performance, user behavior and perception.

Regarding *performance*, the repeated measure ANOVA yielded a significant main effect for initiative ($F(2,42) = 105.9$, $p = 0.000$, $\eta^2 = 0.835$) but not for task type ($F(2,42)=1.41$, $p = 0.71$). This effect size implies that 83.5% of output quality variance is explained by the initiative condition. Participants produced task outputs of higher quality in the AI-initiative and mixed-initiative conditions compared to the user-initiative condition (AI: $M = 7.5$, $SD = 1.1$; MX: $M = 7.4$ $SD = 0.7$; US: $M = 4.4$ $SD = 1.3$).

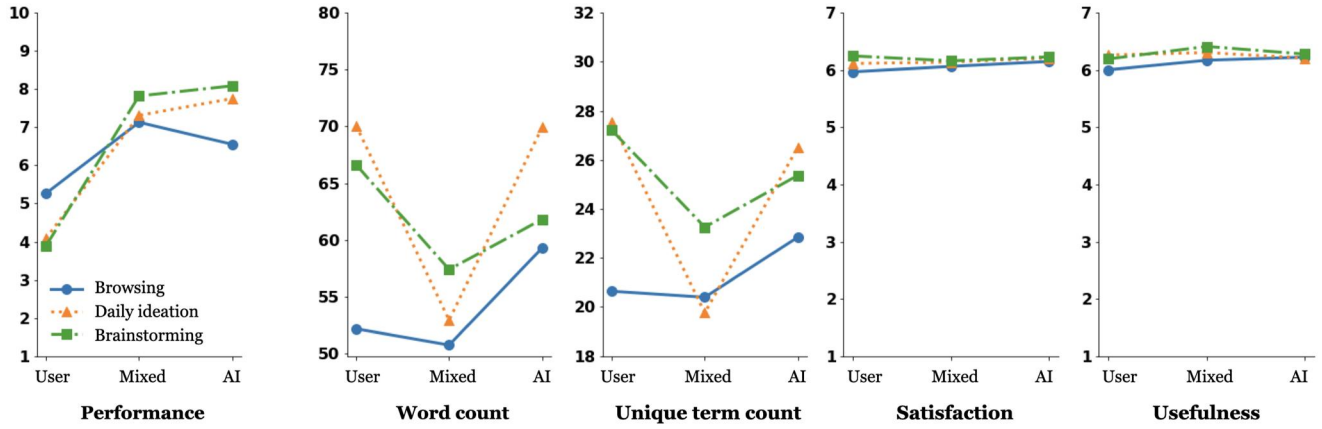


Figure 4. Line graphs representing performance, user behavior, and perceptions based on initiative and task types. Performance analysis was conducted with 15 randomly selected participants per initiative condition (resulting in 125 outputs), while user behavior and perception analysis encompassed all participants (a total of 273 participants).

Table 3. Results of the factorial repeated measures ANOVA.

Measure	Manipulated variable	*df*	*F*-value	*p* Value
Performance (behavioral)	Initiative Task Initiative: Task	2	105.91	0.000***
		2	1.41	0.710
		4	0.72	0.038*
Word count (behavioral)	Initiative Task Initiative: Task	2	3.14	0.045*
		2	9.81	0.000***
		4	2.63	0.035*
Unique term count (behavioral)	Initiative Task Initiative: Task	2	3.97	0.020*
		2	14.31	0.000***
		4	4.44	0.002**
Satisfaction (perceived)	Initiative Task Initiative: Task	2	0.23	0.797
		2	4.29	0.014*
		4	0.716	0.581
Usefulness (perceived)	Initiative Task Initiative: Task	2	0.56	0.572
		2	5.49	0.004*
		4	1.56	0.184

Main effects of initiative were observed for performance, word count, and unique term count. Both the AI-initiative and mixed-initiative conditions produced higher quality outputs. On the other hand, user prompts in the mixed-initiative were more concise, with fewer words and unique terms, suggesting that participants in this condition generated high-quality output with more succinct prompts. There were also significant interaction effects between initiative and task type for performance, word count, and unique term count. This suggests that the influence of initiative on these variables might vary depending on the specific task type. *Note:* $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

In terms of *word count in user prompts*, a significant main effect was observed for initiative ($F(2,270) = 3.12$, $p = 0.045$, $\eta^2 = 0.617$). This indicates that approximately 61.7% of the variance in user token count is explained by the prompting condition, underscoring a robust effect of the prompting approach on prompt verbosity. Participants in the AI-initiative and user-initiative conditions generated prompts with more words than those in the mixed-initiative condition (AI: $M = 63.7$, $SD = 39.2$; US: $M = 62.9$, $SD = 42.3$; MX: $M = 53.7$, $SD = 32.9$). Similarly, there was a significant main effect of initiative on *the number of unique terms in user prompts* ($F(2,270) = 3.97$, $p = 0.02$, $\eta^2 = 0.647$). This effect size reflects a robust influence of the prompting condition on unique term usage. Prompts from the AI-initiative and user-initiative conditions included a significantly greater number of unique terms compared to the mixed-initiative condition (AI: $M = 24.9$, $SD = 12.6$; US: $M = 25.1$, $SD = 15.3$; MX: $M = 21.1$, $SD = 12.1$) (Table 3).

No significant differences were found across the three conditions in terms of perceived satisfaction and perceived usefulness. The absence of significant differences in user perception, despite clear performance variations, suggests that participants may not readily recognize the advantages of different prompting approaches.

To summarize, while participants did not perceive differences in satisfaction and usefulness, their behavior revealed differences based on the initiative condition. Participants in both the AI and mixed-initiative conditions outperformed those in the user-initiative condition. Notably, participants in the mixed-initiative condition produced high-quality output even with more concise prompts.

5.4. Interaction effect between initiative and task type (RQ2)

In the second research question, we aimed to observe how the effects of initiative on user behavior and perception differ across distinct task types. By analyzing the interaction effect from a repeated measures ANOVA, we could verify if the effect of AI-initiative on user behavior and perception remained consistent across different tasks or exhibited variations. As a result, interaction effects between the initiative and task type were found in output quality ($F(4,270) = 0.72$, $p = 0.038$, $\eta^2 = 0.644$), word count ($F(4,270) = 2.63$,

Table 4. Result of the performance, user behavior, and perception across tasks.

	Browsing				Daily ideation				Brainstorming			
	US	MX	AI	*p*	US	MX	AI	*p*	US	MX	AI	*p*
Performance												
Output quality	5.26	**7.13**	**6.56**	***	4.09	**7.31**	**7.76**	***	3.89	**7.82**	**8.08**	***
Behavior												
# of Turn-taking	4.20	4.36	4.45		4.02	4.02	4.20		4.14	4.32	4.10	
Word count	52.16	50.73	59.28		**70.07**	52.91	**69.93**	**	66.61	57.39	61.81	
# of Unique term	20.64	20.40	22.85		**27.56**	19.76	**26.53**	***	27.22	23.26	25.38	
Perception												
Satisfaction	5.97	6.07	6.15		6.12	6.14	6.21		6.25	6.16	6.23	
Usefulness	6.01	6.17	6.23		6.26	6.31	6.21		6.20	6.41	6.28	

Bolded values indicate the highest mean within each row that was statistically significant based on post-hoc comparisons. Across all tasks, the mixed-initiative and AI-initiative outperformed the user-initiative in producing higher quality output. The main effects of initiative on word count and unique term count were pronounced in the daily ideation task. *Note:* $^{**}p < 0.01$; $^{***}p < 0.001$.

$p = 0.035$, $\eta^2 = 0.348$), and unique term count ($F(4,270) = 4.44$, $p = 0.002$, $\eta^2 = 0.451$) (Table 4).

For output quality, although the main effect of condition accounts for a large portion of the variance (approximately 64.37%), the interaction effect reveals that only about 8.89% of the variance is explained by the variation in how different task types modulate this effect. This relatively smaller yet meaningful interaction indicates that while the prompting approach robustly enhances output quality overall, its impact does vary by task type. Similar interaction effects were observed for word count and unique term count, with effect sizes of 34.8 and 45.1%, respectively, further underscoring the task-dependent nature of the prompting strategy's influence on user behavior and perceptions.

5.5. Simple effect of AI-initiative on user experience by task type (RQ2)

To investigate the differential impacts of initiative types across tasks, one-way ANOVAs were conducted on the variables where interaction effects were observed: performance, word count, and unique term count. The mean values and statistical significance of these variables, as differentiated by condition and task types, are presented in Table 4.

Regarding output quality, used as a metric for measuring performance, significant simple effects of the initiative were found across all tasks (Browsing: $F(2,42) = 17.41$, $p = 0.000$; Daily ideation: $F(2,42) = 87.72$, $p = 0.000$; Brainstorming: $F(2,42) = 68.77$, $p = 0.000$). *Post-hoc* Tukey's HSD tests revealed that both the AI-initiative and mixed-initiative conditions produced significantly higher levels of output quality compared to the user-initiative condition in all three tasks (Browsing: AI vs. US, $t(28) = 3.87$, $p = 0.000$, MX vs. US, $t(28) = 5.83$, $p = 0.000$; Daily ideation: AI vs. US, $t(28 = 10.90$, $p = 0.000$, MX vs. US: $t(28) = 10.39$, $p = 0.000$; Brainstorming: AI vs. US, $t(28) = 9.39$, $p = 0.000$, MX vs. US: $t(28) = 9.34$, $p = 0.000$). However, there were no significant differences in output quality between the AI-initiative and mixed-initiative conditions in any of the tasks (Browsing: $t(28) = -1.82$, $p = 0.187$; Daily ideation: $t(28) = 1.75$, $p = 0.315$; Brainstorming: $t(28) = 0.82$, $p = 0.785$). This indicates that participants in the AI and mixed-initiative conditions consistently outperformed those in the user-initiative condition across all task types. Figure 5 illustrates these task-specific response quality comparisons.

Notably, the effect of initiative on prompt conciseness was task-dependent. For the word count, a significant effect was identified solely in the daily ideation task ($F(2,270) = 5.34$, $p = 0.005$). *Post-hoc* results showed that both user-initiative and AI-initiative conditions generated a significantly higher word count than the mixed-initiative (AI-MX: t $(186) = 2.83$, $p = 0.015$; US-MX: $t(178) = 2.91$, $p = 0.015$). No significant differences based on initiative types were found for the browsing ($F(2,270) = 1.55$, $p = 0.21$) and brainstorming ($F(2,270) = 1.36$, $p = 0.26$) tasks, indicating that initiative type did not significantly influence response length. Consequently, *post-hoc* tests were not performed.

Similarly, significant effects for unique term count were observed only in the daily ideation task (($F(2,270) = 9.72$, $p = 0.000$). Consistent with the word count findings, the user-initiative and AI-initiative conditions resulted in a significantly higher count of unique terms compared to the mixed-initiative condition (AI-MX: t $(186) = 3.55$, $p = 0.000$; US-MX: t $(178) = 4.02$, $p = 0.000$). The browsing and brainstorming tasks showed no significant main effects of initiative on unique term count (Browsing: $F(2,270) = 1.02$, $p = 0.36$; Brainstorming: $F(2,270) = 1.86$, $p = 0.15$); therefore, no *post-hoc* analyses were conducted.

5.6. Acceptance rate of AI-generated prompts (RQ3)

For our third research question, we aimed to investigate the acceptance rate of AI-generated prompts in the AI-initiative and mixed-initiative conditions. Table 5 presents a cross-tabulation table and Figure 6 provides a graphical representation of the acceptance rate of AI-generated prompts in both conditions. For an in-depth qualitative understanding of users' perception of AI-based prompt creation, refer to Section 5.6.4.

In the AI-initiative condition, 656 of 1232 prompts were generated by AI, resulting in an acceptance rate of 53.25%. By task, the acceptance rate for browsing was 49.3% (217/440), for daily ideation 58.0% (229/395), and for brainstorming 52.9% (210/397). Meanwhile, in the mixed-initiative condition, PromptPilot generated 292 prompts based on user input out of a total of 1,250 user prompts, resulting in an

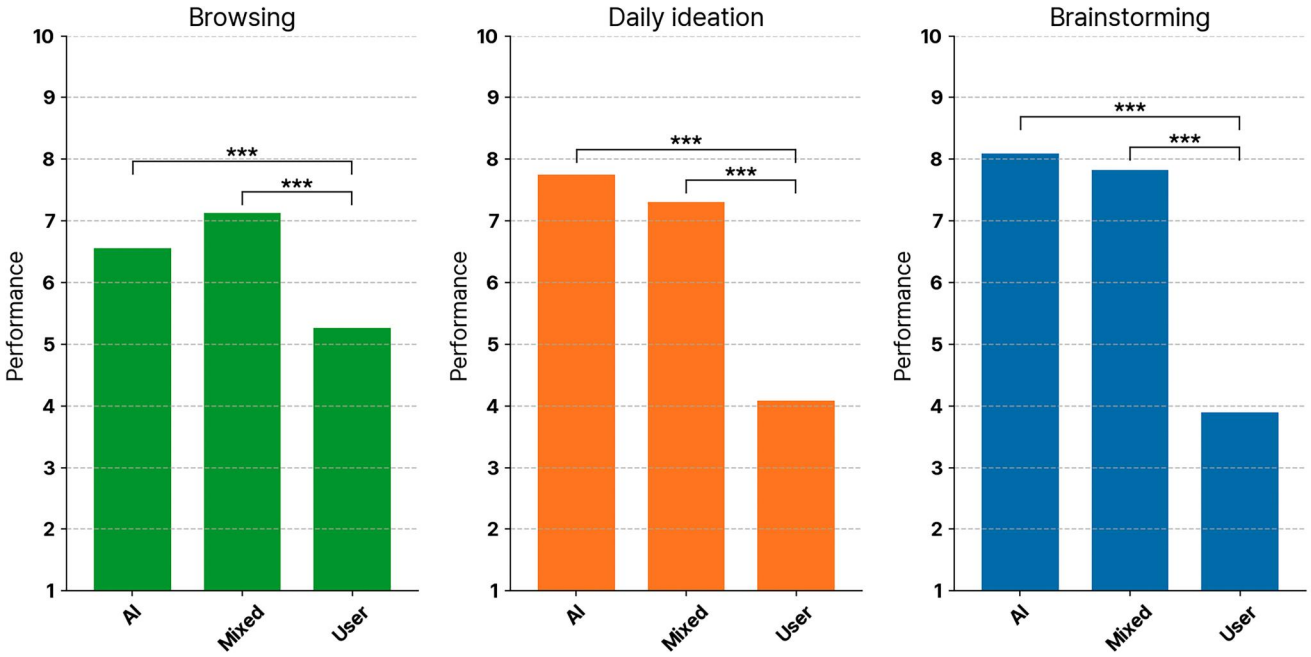


Figure 5. Task-specific response quality across initiative conditions.

Table 5. Cross tabulation table of acceptance rate of AI-generated prompts.

Condition	AI-initiative	Mixed-initiatve	Sum
AI-generated prompt	656 (26.43%)	292 (11.76%)	948 (38.20%)
Human-generated prompt	576 (23.21%)	958 (38.60%)	1534 (61.80%)
Sum	1232 (49.64%)	1250 (50.36%)	2482 (100%)

acceptance rate of 23.3%. It is noteworthy to mention that the Magic Wand feature was activated 324 times. This indicates that upon engaging with the Magic Wand feature, users adopted 90.1% of the AI-generated output. In terms of task-specific rates, the acceptance figures were 23.4% for browsing (103/440), 25.8% for daily ideation (102/395), and 21.9% for brainstorming (87/397).

A chi-square test was conducted to examine the significant relationship between the acceptance rate of AI-generated prompts and the AI-initiative style. The results showed a significant difference in the adoption patterns across the different initiative conditions ($\chi^2 = 233.51$, $df = 1$, $p = 0.000$). Furthermore, the Cochran-Armitage test confirmed the observed gradient in the acceptance rate ($\chi^2 = 234.77$, $df = 1$, $p = 0.000$). This implies that participants in the AI-initiative condition demonstrated a significantly higher propensity to adopt AI-generated prompts than those in the mixed-initiative condition.

5.7. Qualitative analysis

5.7.1. PromptPilot and users cooperate to accomplish the task

Our findings highlight the cooperative relationship between AI and humans in achieving quality results. Participants expressed joy in collaborating with PromptPilot: "PromptPilot suggested ideas that opened my eyes for even better insights. Together, we honed these ideas to formulate effective plans." (AI69) Several participants noted how PromptPilot facilitated their brainstorming process: "I would love to have it as a companion to foster idea development and learn new things." (MX23); "I appreciate the assistance and insight provided by this AI." (US37) They also perceived that PromptPilot enhanced their capabilities: "Useful, creative, and swift. It feels like a booster for my brain." (AI47)

In particular, PromptPilot offered novel insights and guidance to participants during challenging moments: "It inspired me with ideas when I might otherwise feel stuck." (MX07); "It was handy when I was stuck for ideas and sought assistance." (US72) Furthermore, it provided participants with ideas they would not have otherwise thought of: "PromptPilot provided valuable advice and ideas I probably wouldn't have conceived by myself." (MX90) It also "helped spark their creativity." (MX21)

5.7.2. AI responses actually matter

Various interaction factors in LLMs influence the user experience, including AI response performance, prompting method, humanlikeness, and response generation time. We found that LLM-generated responses directly influenced user satisfaction and usability. Participants across all conditions praised the creativity and novelty of the responses generated by PromptPilot: "It provided many useful recommendations that were relevant to the questions asked." (US55); "The responses were super useful and matched exactly what I needed." (MX19); "PromptPilot did a very good job at providing illustrative examples that addressed the crux of my inquiries." (MX70); "The answers to my questions were comprehensive and well-considered." (AI49); "Lots and lots of details." (AI02) This qualitative evidence supports our findings that user satisfaction and usability were consistent

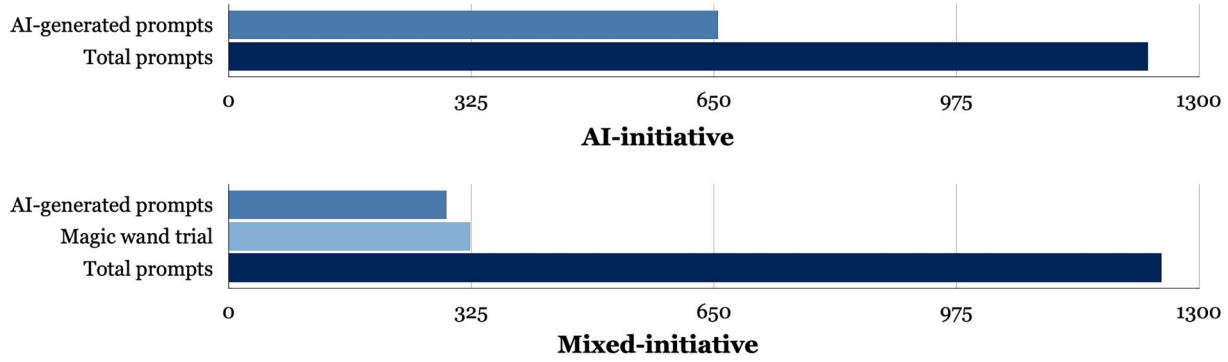


Figure 6. Acceptance rate of AI-generated prompts.

regardless of the prompting method relative to the level of initiative.

5.7.3. Creating and articulating prompts is a major challenge

Writing prompts was revealed as the most prominent challenge of using the LLM system. This issue was reported among participants in the user-initiative condition. Participants described prompting as an art or complex chain reaction: "Just as how conversation is more of an art than a science, so is prompt crafting." (US60); "Making a prompt is like setting the first domino in a chain. It's a tiny piece that catalyzes a complex reaction." (US04) One participant expressed the difficulty of writing a prompt to get the desired results, stating that PromptPilot (user-initiative) did not sufficiently resolve this issue: "It can sometimes be difficult to get PromptPilot to generate the desired results. The prompts need to be carefully crafted in order to get PromptPilot to generate the desired results. This can be time-consuming and frustrating." (US71) Similarly, participants in the mixed-initiative emphasized the importance of clearly articulating their requests to obtain high-quality responses: "If I wasn't very clear on what information I wanted, PromptPilot was going to be broad as well. The user has to know what they want to help the system through follow up questions." (MX28)

Regarding suggestions for system improvement, participants in the user-initiative mentioned the ability to enable prompt clarification or to support prompt creation: "PromptPilot should pose clarifying questions about the prompt." (US09); "I wish PromptPilot would guide me on formulating clearer questions. At times, I'm just not sure how to word things to get the best answers." (US89); "Perhaps there could be a feature where the system suggests prompts based on trending topics or my past queries?" (US28) In addition, a new interaction was also proposed as a possible solution that provides users with the appropriate information by asking the questions: "Rather than me posing questions, it would be intriguing if the system questioned me, building its own analysis to cater to my requirements." (US13)

5.7.4. PromptPilot assist prompt creation

We discovered that both the AI and mixed-initiative systems resolve the challenge of creating prompts. The AI-initiative feature of presenting prompts has enhanced user convenience: "I liked having some prompts to pick from. Way easier than typing everything out." (AI69); "I really liked it. Saved me some time." (AI91) For participants initially uncertain about how to proceed, the AI-generated prompts provided essential guidance: "Those suggestions were on point. Helpful when I was drawing a blank on what to ask for." (AI01); "Facing a prompt like 'plan a party', I'd usually draw a blank. But with PromptPilot, most of the creative work is done for me. It gets me over that initial hurdle." (AI92) Some found the feature to be a source of inspiration when they got aground during the task: "The generated questions helped me generate other ideas when I was stuck." (AI79) Participants also felt it broadened their horizons: "PromptPilot can generate questions on a variety of topics, and the questions are often creative and thought-provoking." (AI18); "It led me down some paths I hadn't even thought about." (AI30)

Participants in the mixed-initiative condition also mentioned positive experiences with input-based prompt generation feature: "Getting a good answer out of AI really depends on how well you phrase the question. The Magic Wand feature solves this issue." (MX12) Another participant mentioned the uniqueness of this feature compared to other LLM systems: "It is a genius idea. I would have never thought of AI doing this. AI is usually only answering your questions. Now it can do both." (MX93) The primary advantage was that it greatly facilitated the articulation process through which users converted nebulous thoughts into coherent, structured words: "Helpful for when I'm having trouble articulating what I want to ask." (MX04); "I only have a general idea of what I want, but it captured the essence of what I wanted." (MX39) Furthermore, this feature was particularly beneficial for novice users: "This could be helpful for new AI users such as myself that don't know how to correctly input a question." (MX61); "At first, I was second-guessing how I worded things. But then, PromptPilot offered me a variety of options, which made things easier." (MX29) Additionally, the keyword- or phrase-based prompt generation aligned well with the search

metaphor, providing a familiar experience: "Usually, I just type in keywords when searching for info. Using a similar approach with the magic wand just felt natural and efficient." (MX74)

On the other hand, while numerous users had a positive user experience with the AI's prompt generation function, there were also reported drawbacks. Notably, three participants noted that the prompt generation function (specifically in the AI-initiative condition) diminished their cognitive engagement: "It was handy, but I worry that it kind of makes me 'switch off' my brain a bit." (AI59); "PromptPilot is great and all, but I feel like it's turning my decisions into no-brainers." (AI17) In the mixed-initiative, Participants who were already adept at formulating their queries did not find the feature as useful: "The question suggestions were okay, but I often found that I'd already formed my own questions in my mind." (MX87) Some users also highlighted issues when the generated prompts weren't precisely what they intended: "The feature is a time-saver, but if there's additional info I want to include that isn't in the suggested question, I need to remember to bring it up later." (MX50)

5.7.5. Users want personalized prompt suggestions and responses

Lastly, participants expressed a preference for the AI to provide more personalized questions and responses. A number of participants expected PromptPilot to generate tailored answers, drawing from task-specific and user data, without necessitating explicit details in the prompt: "(PromptPilot needs) Contextual understanding. If it could pick up on the nuances of what I'm asking, it could give more relevant answers." (US01) Specifically, those in the AI-initiative condition hoped for the generated prompts to adapt to their individual preferences: "I envision that with more use and once it gets to know me better, the questions it suggests would get sharper and more on point." (AI93) Moreover, participants across all conditions were reluctant for the AI to make inaccurate assumptions about them. They hoped PromptPilot to comprehend them more, asking clarifying questions before generating specific prompts and replies: "It shouldn't jump to conclusions. It really should ask for clarification." (MX31); "I feel it should try to zero in on exactly what I want. Maybe it could ask some follow-up questions based on my initial ones to get more specifics." (US79); "It could've asked me a couple more things before giving an answer." (AI57)

6. Discussion

In this section, we discuss the findings of our study and their implications for designing user-AI interactions, specifically focusing on how AI can assist users in generating prompts.

6.1. Enhanced AI initiative and user experience

Our results revealed no significant difference in user perception (satisfaction and usability) across initiative levels. Despite this, user behavior data indicated that AI-initiative and mixed-initiative conditions generated better outputs compared to user-initiative, suggesting that AI assistance helped manage the cognitive demands of prompt creation (Sweller, 2011). Notably, participants in the mixed-initiative condition crafted more concise prompts while maintaining quality, with acceptance rates of 53.3% for AI-initiative and 23.3% for mixed-initiative conditions. This indicates that users were actively leveraging the AI-assisted prompt generation capability. Our qualitative analysis further affirmed that this feature not only elevated PromptPilot's usability but also served as a valuable source of inspiration for users when they were "drawing a blank on what to ask."

From a Cognitive Load Theory perspective, different types of AI assistance may distinctly affect cognitive load management (Sweller, 1988, 2011). The AI-initiative approach reduces extraneous load by providing complete prompt suggestions, while the mixed-initiative condition balances cognitive burden reduction with user engagement via structured input. These effects varied by task type. In daily ideation tasks, which impose high intrinsic load due to personal requirements, the mixed-initiative approach proved particularly effective at reducing cognitive load while maintaining output quality. However, exploratory browsing and complex brainstorming tasks, with their different intrinsic cognitive demands, required different patterns of cognitive support. This aligns with CLT's principle that the effectiveness of cognitive support mechanisms depends on task complexity and specific demands (Lyell et al., 2018).

The mixed-initiative condition's keyword-based approach resembles familiar search processes, making it particularly effective for users struggling with prompt formulation. Our results highlight the efficacy of an enhanced AI-initiative in prompt creation, confirming the positive aspects of mixed-initiative interactions. This framework enables human-machine collaboration (Burstein & McDermott, 1996; Carbonell, 1970; Rodrigues Barbosa et al., 2024), with PromptPilot generating prompts that align with user intentions while reducing risks of inaccurate AI predictions (Horvitz, 2007).

These findings align with the concept of Human-Computer Integration (Rodrigues Barbosa et al., 2024; Mueller et al., 2020), with different initiative conditions representing varying degrees of human-AI partnership. The AI-initiative condition demonstrates technology-led control while maintaining user agency through prompt modification options, whereas the mixed-initiative condition achieves more balanced control between humans and technology (Kim et al., 2024; Shneiderman, 2020). In both cases, PromptPilot maintains user engagement by incorporating user choices and generating prompts that align with user intentions, thereby reducing risks associated with inaccurate predictions (Horvitz, 2007).

- Design Implication (D1): Implement a balanced AI-initiative system that maintains user agency while providing automated assistance. Allow users to modify AI-generated prompts while benefiting from the system's suggestions to optimize cognitive load and task efficiency.

6.2. Mixed-initiative interaction, self-articulation, and prompt conciseness

The Magic Wand function, integral to the mixed-initiative system, helps users craft precise and concise prompts by articulating their thoughts. By requiring users to input specific keywords or phrases, this feature facilitates the transformation of vague ideas into clear, focused concepts (Schaekermann et al., 2018). As one participant noted, "The data and ideas produced were as helpful as the specificity of my questions." (MX38) This structured approach helped users maintain focus on their primary objectives while preventing deviation from the main task (Farnham et al., 2000): "I was able to access the details I sought quickly without going off on tangents." (MX72); "Even when I felt I might stray, Magic Wand kept me directly on point, focused and relevant. I never veered off the main topic." (MX19)

Analysis of user interaction data reinforces these qualitative results. The Magic Wand function allowed users to encapsulate their essential requirements through concise keywords or phrases, especially benefiting tasks requiring personalization. For instance, participant MX29 typed in "cheapest decoration" and then selected the AI-generated prompt, "What are some budget-friendly decoration ideas for a surprise party?" Likewise, other participants provided specific phrases for the daily ideation task, such as "surprise party for a beer-loving friend," "boat, travel, river," and "pet-friendly spaces for parents and kids." This highlights how structured, user-guided input leads to more relevant and targeted AI-generated prompts.

The effectiveness of this approach can be attributed to cognitive scaffolding and LLM optimization. Distilling thoughts into keywords serves as a form of external cognition, helping users organize and clarify their ideas before engaging with the AI (Clark, 1998; Hollan et al., 2000). By integrating user input with AI assistance, mixed-initiative systems balance cognitive load, encouraging users to actively engage while easing the burden of formulating complete prompts (Riche et al., 2010). Moreover, concise keyword-based inputs also align well with best practices in prompt engineering for LLMs, potentially leading to more effective AI-generated prompts (Renze & Guven, 2024). In sum, structured input methods can enhance both user articulation and AI prompt generation, particularly for personalized tasks.

- Design Implication (D2): Incorporate a structured keyword-based input system that helps users distill their thoughts into clear concepts before engaging with the AI. This could include a guided interface for entering key terms that the system then expands into full prompts.

6.3. Mixed-initiative prompting in personalized ideation tasks

Mixed-initiative prompting produced more concise yet high-quality output in the daily ideation task. The daily ideation task's focus on personalized, practical planning benefited from the mixed-initiative's keyword-based interaction. When planning a surprise party, participants needed to articulate specific preferences and constraints (e.g., "outdoor activities," "budget-friendly decorations"). Unlike the user-initiative condition where users had to construct complete prompts, or the AI-initiative condition where suggestions might not capture personal context, the mixed-initiative allowed users to quickly focus the AI's assistance on their specific personalized requirements through keywords. This structured input approach reflects principles of cognitive scaffolding, where systems help users transform vague ideas into clear expressions (Clark, 1998). By focusing on essential elements, participants could maintain control over the creative direction while still benefiting from the system's ability to structure and elaborate their ideas efficiently.

In contrast, the browsing task's exploratory nature and the brainstorming task's complex conceptual demands may have required different interaction patterns. During browsing, users may have required more descriptive, exploratory prompts to guide their information discovery, consistent with research on information-seeking behaviors with LLMs (Zhai, 2024). Similarly, brainstorming a mobile app likely required more detailed and context-rich prompts to convey nuanced requirements. In these cases, the advantages of concise, keyword-based prompting were less pronounced since broad information and complex concepts require more detailed guidance.

- Design Implication (D3): Design task-specific interfaces that adapt the level of AI assistance based on the task type. For personalized tasks, emphasize keyword-based interactions, while providing more detailed prompting support for complex conceptual tasks.

6.4. Designing seamless mixed-initiative interaction

In terms of the acceptance rate of AI-generated prompts, the AI-initiative condition resulted in acceptance rates approximately 2.3 times higher than those of the mixed-initiative condition. However, a critical finding emerges upon closer examination of user behavior in the mixed-initiative condition. When participants in this condition opted to use the Magic Wand feature, they adopted the AI-generated prompt 90.1% of the time. This high adoption rate suggests that users who initially engage with the AI-assisted feature are likely to consistently leverage its benefits.

This finding underscores that there is opportunity to enhance the accessibility and integration of the Magic Wand functionality within PromptPilot's mixed-initiative condition. Exploring design alternatives to create a more seamless interface could potentially increase the initial and ongoing use of mixed-initiative interactions (Case, 2015). By reducing the barrier to entry and improving the fluidity of AI assistance, we may encourage more users to engage with and consistently benefit from the AI-generated suggestions, thereby optimizing the synergy between user input and AI capabilities.

Specifically, the current system supports step-by-step and segmented interactions which can impose dual cognitive loads on users. Potential improvements include real-time

prompt suggestions based on ongoing user input and predictive typing for completions or full sentences. These design opportunities suggest a shift along the continuum from explicit to implicit AI involvement (Park et al., 2021). Moving towards more implicit assistance could increase the use of mixed-initiative features while maintaining user agency. Future research could explore adaptive interfaces that learn from user interactions, optimizing the balance between seamless AI assistance and user control to enhance human-AI collaboration in prompt engineering tasks.

- Design Implication (D4): Create a more fluid interaction model with real-time prompt suggestions and predictive typing features that reduce the cognitive burden of switching between user input and AI assistance modes.

6.5. Human-like interaction and user engagement

Our findings indicate that users perceive PromptPilot in a manner akin to human interaction. This observation is consistent with the Computers as Social Actors (CASA) paradigm, which posits that individuals apply similar social conventions when interacting with both computers and humans (Nass et al., 1994). Our results also align with prior studies indicating that users emulate human-to-human interactions when providing instructions to LLMs (Rastogi et al., 2023; Zamfirescu-Pereira et al., 2023). A number of participants attributed human-like qualities to PromptPilot, noting: "It feels like having a second person to brainstorm with." (AI02); "PromptPilot felt like conversing with a highly knowledgeable friend. The ideas it provided were very realistic and useful." (US84).

While retrieving pertinent information is essential, it is also important to design LLM interactions that mirror human-like interpersonal exchanges. Participants in the user-initiative condition highlighted PromptPilot's reactive interaction style, which responded only to user prompts. These participants expressed a preference for more proactive and human-like interactions. As US18 stated, "It should mimic the natural flow of human conversation, not just react to direct queries." In a similar vein, US47 added, "Incorporating follow-up questions would make the experience more conversational. AI inquiries such as 'Do you need more information?' or 'Is that what you were looking for?' would add a human touch."

Interestingly, while objective performance measures showed clear benefits of AI-assisted prompting, users reported similar levels of satisfaction and usefulness across all conditions. This disconnect between performance and perception suggests that users may not fully recognize the advantages of different prompting approaches, possibly due to their limited experience with LLM interactions. Future iterations of PromptPilot could bridge this gap by incorporating more explicit feedback mechanisms that help users understand output quality and prompting effectiveness (Benharrak et al., 2024). Moreover, longitudinal studies could reveal whether users' awareness of these benefits develops with increased system familiarity, potentially leading to more aligned subjective and objective measures of system effectiveness.

- Design Implication (D5): Incorporate conversational elements and proactive follow-up questions to make the interaction more natural and engaging.

6.6. Practical integrations and real-world applications

While this study demonstrates the effectiveness of PromptPilot in controlled settings, exploring its integration with common LLMs like ChatGPT could further validate its real-world applicability. Implementing the mixed-initiative feature as an enhanced input interface, where users enter keywords and receive real-time prompt suggestions, would be particularly beneficial for novice users. By providing contextually relevant suggestions, mixed-initiative prompting reduces the cognitive load of prompt formulation, minimizing frustration and increasing the likelihood of desired outcomes (Sweller, 2020, 2011; Zamani et al., 2020). This also allows novices to learn prompt engineering best practices, building their skills and confidence in using AI tools independently (Park & Ahn, 2024).

The AI-initiative feature could be integrated as a "suggested prompts" panel in existing LLM interfaces, proactively suggesting follow-up prompts based on the conversation context and task type. The high acceptance rate (53.25%) of AI-generated prompts in our study suggests the effectiveness of this approach. Unlike mixed-initiative, which requires user input, AI-initiative can anticipate user needs and suggest prompts even when users are uncertain about the next steps. This proactive guidance is especially useful for users unfamiliar with task requirements. Additionally, real-time suggestions based on ongoing input and predictive typing can further reduce cognitive load, supporting novice prompt formulation and enhancing interaction fluidity.

- Design Implication (D6): Implement an enhanced input area designed to guide novice users in prompt formulation by combining user-driven input with proactive, context-aware AI suggestions.

6.7. Ethical considerations

While PromptPilot improves AI-assisted prompting, it raises ethical concerns regarding user overreliance and prompt manipulation. AI-initiative systems, where users predominantly rely on AI-generated prompts, may reduce independent critical thinking. Users often default to AI suggestions without critically evaluating alternatives, leading to dependency on automated assistance for problem-solving (Marco et al., 2024). This is particularly relevant in domains requiring critical thinking, creativity, or complex reasoning. To mitigate this, prompting systems should encourage active engagement by requiring users to modify AI-generated prompts or providing justifications that stimulate critical assessment.

Another concern is the risk of prompt manipulation, where AI-generated suggestions could unintentionally shape user inquiries in biased or misleading directions (Jain & Jain, 2024; Li et al., 2024). Research has shown that

opinionated language models can subtly influence human writers' perspectives and beliefs (Jakesch et al., 2023). This issue is particularly relevant in tasks that involve information retrieval, decision-making, or content generation, where subtle biases in AI-generated prompts may influence users' perspectives or limit their exploration of alternative viewpoints. To minimize potential bias, mixed-initiative prompting systems should incorporate transparency mechanisms, such as explaining how AI-generated suggestions are formulated or offering users control over the prompt refinement process.

6.8. Limitation and future work

We outline the limitations of our study and propose directions for future research. First, while our study focused on typical users' interactions with LLMs, this broad user group approach may have overlooked important differences in how varying levels of expertise affect prompt creation needs. Future research should examine how different user profiles - from novices to domain experts - might benefit from different types of prompting support. Second, although we examined three common task types (browsing, ideation, and brainstorming), this wide range may have prevented deeper insights into task-specific prompting requirements. A more focused study design examining fewer tasks in greater depth could better reveal how initiative patterns should be tailored to specific task demands. Third, our study was conducted in an online experimental context rather than a natural context, which might constrain our understanding of user behaviors over extended periods. For subsequent research, we aim to deploy the AI system that assists in prompt generation in real-world settings, free from spatial or temporal limitations. Forth, while we operationalized levels of initiative through automated prompt recommendations and user input-based prompt generation, many other methods could heighten the AI's initiative within the interface. Future studies could explore various ways to integrate mixed-initiative interactions during prompt formulation. It's worth noting that we validated our operationalization of initiative through a manipulation check. Fifth, while our study formulated prompts considering task characteristics and user input, various prompting strategies can be integrated within LLMs. Such strategies may encompass the use of examples for input and output (Brown et al., 2020; White et al., 2023), urging users for more explicit prompts (White et al., 2023), and refining prompt formats (Bach et al., 2022). Lastly, while participants in our study conducted designated tasks, LLM systems can support a diverse range of activities, including collaborative writing and coding. Future investigations could investigate the efficacy of AI-driven prompt guidance across a wider range of tasks.

7. Conclusion

This study investigated the user experience concerning prompt creation when interacting with LLMs, with a specific emphasis on initiative and task characteristics. We introduced "PromptPilot," a research probe aimed at assisting users in crafting prompts. Employing both quantitative and qualitative approaches, we evaluated the efficacy of PromptPilot across varying initiatives and task types. Notably, our results demonstrated the superior output quality of AI-initiative and mixed-initiative over user-initiative. Additionally, we observed intricate user behavioral patterns, such as more concise prompts in the mixed-initiative condition. Drawing from these findings, we suggested design implications for user–AI interactions during prompt creation. We hope that this work will serve as a step toward a deeper and more inclusive understanding of interfaces in which users can leverage the capability of AI when interacting with LLMs.

Note

1. Google's Bard language model was rebranded as Gemini in February 2024.

Acknowledgements

This work was supported by the SNU-Global Excellence Research Center establishment project and was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [NO. RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)].

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Soomin Kim http://orcid.org/0000-0003-2523-7808
Joonhwan Lee http://orcid.org/0000-0002-3115-4024

References

Ai, Q., Bai, T., Cao, Z., Chang, Y., Chen, J., Chen, Z., Cheng, Z., Dong, S., Dou, Z., Feng, F., Gao, S., Guo, J., He, X., Lan, Y., Li, C., Liu, Y., Lyu, Z., Ma, W., Ma, J., … Zhu, X. (2023). Information retrieval meets large language models: A strategic report from Chinese IR community. *AI Open*, *4*, 80–90. https://doi.org/10.1016/j.aiopen.2023.08.001

Aliannejadi, M., Zamani, H., Crestani, F., & Croft, W. B. (2019). Asking clarifying questions in open-domain information-seeking conversations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 475–484). Association for Computing Machinery. https://doi.org/10.1145/3331184.3331265

Anthropic. (2024). Be clear, direct, and detailed – anthropic. https://docs.anthropic.com/en/docs/build-with-claude/prompt-engineering/be-clear-and-direct

Anthropic. (2025). Pricing anthropic. https://www.anthropic.com/pricing#anthropic-api.

Arawjo, I., Swoopes, C., Vaithilingam, P., Wattenberg, M., & Glassman, E. L. (2024). Chain-forge: A visual toolkit for prompt engineering and LLM hypothesis testing. In *Proceedings of the Chi Conference on Human Factors in Computing Systems* (pp. 1–18). Association for Computing Machinery.

Ashktorab, Z., Dugan, C., Johnson, J., Pan, Q., Zhang, W., Kumaravel, S., & Campbell, M. (2021). Effects of communication directionality and AI agent differences in human-AI interaction. In *Proceedings of the 2021 Chi Conference on Human Factors in Computing Systems* (pp. 1–15). Association for Computing Machinery.

Bach, S. H., Sanh, V., Yong, Z.-X., Webson, A., Raffel, C., & Nayak, N. V., Others. (2022). Promptsource: An integrated development environment and repository for natural language prompts. In V. Basile, Z. Kozareva, & S. Stajner (Eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 93–104). Association for Computational Linguistics.

Bellemare-Pepin, A., Lespinasse, F., Thölke, P., Harel, Y., Mathewson, K., Olson, J. A., & Jerbi, K. (2024). Divergent creativity in humans and large language models. arXiv preprint arXiv:2405.13012. https://doi.org/10.48550/arXiv.2405.13012

Benharrak, K., Zindulka, T., Lehmann, F., Heuer, H., & Buschek, D. (2024). Writer-defined AI personas for on-demand feedback generation. In *Proceedings of the Chi Conference on Human Factors in Computing Systems* (pp. 1–18). Association for Computing Machinery.

Birnbaum, L., Horvitz, E., Kurlander, D., Lieberman, H., Marks, J., & Roth, S. (1997). Compelling intelligent user interfaceshow much AI? In *Proceedings of the 2nd International Conference on Intelligent User Interfaces* (pp. 173–175). Association for Computing Machinery.

Brandtzaeg, P. B., Skjuve, M., & Følstad, A. (2024). *Understanding model power in social AI* (pp. 1–11). AI & SOCIETY.

Braun, V., & Clarke, V. (2013). *Successful qualitative research: A practical guide for beginners*. SAGE Publications.

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

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., & Dhariwal, P., Others. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901. https://doi.org/10.5555/3495724.3495883

Burstein, M. H., & McDermott, D. V. (1996). Issues in the development of human-computer mixed-initiative planning. In *Advances in psychology*. (Vol. 113, pp. 285–303). Elsevier.

Carbonell, J. R. (1970). *Mixed-initiative man-computer instructional dialogues (Final Report No. BBN-1971, Job No. 11399)*. Bolt Beranek & Newman, Inc.

Case, A. (2015). *Calm technology: Principles and patterns for non-intrusive design*. O'Reilly Media, Inc.

Chan, J., Dang, S., & Dow, S. P. (2016a). Comparing different sensemaking approaches for large-scale ideation. In *Proceedings of the 2016 Chi Conference on Human Factors in Computing Systems* (pp. 2717–2728). Association for Computing Machinery.

Chan, J., Dang, S., & Dow, S. P. (2016b). Improving crowd innovation with expert facilitation. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (pp. 1223–1235). Association for Computing Machinery.

Chavula, C., Choi, Y., & Rieh, S. Y. (2022). Understanding creative thinking processes in searching for new ideas. In *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval* (pp. 321–326). Association for Computing Machinery.

Chen, O., Kalyuga, S., & Sweller, J. (2017). The expertise reversal effect is a variant of the more general element interactivity effect. *Educational Psychology Review*, *29*(2), 393–405. https://doi.org/10.1007/s10648-016-9359-1

Clark, A. (1998). Magic words: How language augments human computation. In P. Carruthers & J. Boucher (Eds.), *Language and thought: Interdisciplinary themes* (pp. 162–183). Cambridge University Press. https://doi.org/10.1017/CBO9780511597909.011

Creswell, J. W., & Poth, C. N. (2016). *Qualitative inquiry and research design: Choosing among five approaches*. Sage publications.

Crispino, N., Montgomery, K., Zeng, F., Song, D., & Wang, C. (2023). Agent instructs large language models to be general zero-shot reasoners. arXiv Preprint, arXiv:2310.03710. https://doi.org/10.48550/arXiv.2310.03710

Dai, S., Shao, N., Zhao, H., Yu, W., Si, Z., Xu, C., & Xu, J. (2023). Uncovering Chatgpts capabilities in recommender systems. In *Proceedings of the 17th ACM Conference on Recommender Systems* (pp. 1126–1132). Association for Computing Machinery.

Farnham, S., Chesley, H. R., McGhee, D. E., Kawal, R., & Landau, J. (2000). Structured online interactions: Improving the decision-making of small discussion groups. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work* (pp. 299–308). Association for Computing Machinery.

Girotra, K., Meincke, L., Terwiesch, C., & Ulrich, K. T. (2023). Ideas are dimes a dozen: Large language models for idea generation in innovation. *The Wharton School Research Paper (Forthcoming)*. SSRN. https://doi.org/10.2139/ssrn.4526071

Glaser, B., & Strauss, A. (2017). *Discovery of grounded theory: Strategies for qualitative research*. Routledge.

Goldberg, S., Belyaev, S., & Sluchak, V. (2021). Dr. Watson type artificial intellect (AI) systems. arXiv Preprint, arXiv:2106.13322. https://doi.org/10.48550/arXiv.2106.13322

Google. (2023). An important next step on our AI journey. https://blog.google/technology/ai/bard-google-ai-search-updates/

Google (2024). A quick-start handbook for effective prompts. https://services.google.com/fh/files/misc/gemini-for-google-workspace-prompting-guide-101.pdf.

Graesser, A. C., Chipman, P., Haynes, B. C., & Olney, A. (2005). Autotutor: An intelligent tutoring system with mixed-initiative dialogue. *IEEE Transactions on Education*, *48*(4), 612–618. https://doi.org/10.1109/TE.2005.856149

Hollan, J., Hutchins, E., & Kirsh, D. (2000). Distributed cognition: Toward a new foundation for human-computer interaction research. *ACM Transactions on Computer-Human Interaction*, *7*(2), 174–196. https://doi.org/10.1145/353485.353487

Horvitz, E. (1999). Principles of mixed-initiative user interfaces. In *Proceedings of the Sigchi Conference on Human Factors in Computing Systems* (pp. 159–166). Association for Computing Machinery.

Horvitz, E. J. (2007). Reflections on challenges and promises of mixed-initiative interaction. *AI Magazine*, *28*(2), 3–3. https://doi.org/10.1609/aimag.v28i2.2036

Jain, R., & Jain, A. (2024). Generative AI in writing research papers: A new type of algorithmic bias and uncertainty in scholarly work. In *Intelligent Systems Conference* (pp. 656–669). Springer Nature.

Jakesch, M., Bhat, A., Buschek, D., Zalmanson, L., & Naaman, M. (2023). Co-writing with opinionated language models affects users' views. In *Proceedings of the 2023 Chi Conference on Human Factors in Computing Systems* (pp. 1–15). Association for Computing Machinery.

Jiang, E., Olson, K., Toh, E., Molina, A., Donsbach, A., Terry, M., & Cai, C. J. (2022, April). Promptmaker: Prompt-based prototyping with large language models. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts* (pp. 1–8). Association for Computing Machinery.

Joshi, I., Shahid, S., Venneti, S., Vasu, M., Zheng, Y., Li, Y., & Chan, G. Y.-Y. (2024). Coprompter: User-centric evaluation of LLM instruction alignment for improved prompt engineering. arXiv preprint arXiv:2411.06099. https://doi.org/10.48550/arXiv.2411.06099

Kim, S., Eun, J., Oh, C., & Lee, J. (2024). Journey of finding the best query: Understanding the user experience of AI image generation system. *International Journal of Human–Computer Interaction*, *41*(2), 951–969. https://doi.org/10.1080/10447318.2024.2307670

Kraus, M., Wagner, N., & Minker, W. (2020). Effects of proactive dialogue strategies on human-computer trust. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization* (pp. 107–116). Association for Computing Machinery.

Kuang, E., Li, M., Fan, M., & Shinohara, K. (2024). Enhancing UX evaluation through collaboration with conversational AI assistants: Effects of proactive dialogue and timing. In *Proceedings of the Chi*

Conference on Human Factors in Computing Systems (pp. 1–16). Association for Computing Machinery.

Levy, M., Jacoby, A., & Goldberg, Y. (2024). Same task, more tokens: The impact of input length on the reasoning performance of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 15339–15353). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.818

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., & Goyal, N., Others. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, *33*, 9459–9474.

Li, X., Liu, Z., Xiong, C., Yu, S., Yan, Y., Wang, S., & Yu, G. (2024). Say more with less: Understanding prompt learning behaviors through gist compression. arXiv preprint arXiv:2402.16058. https://doi.org/10.48550/arXiv.2402.16058

Li, X., Zhao, R., Chia, Y. K., Ding, B., Joty, S., Poria, S., & Bing, L. (2023). Chain-of-knowledge: Grounding large language models via dynamic knowledge adapting over heterogeneous sources. arXiv preprint arXiv:2305.13269. https://doi.org/10.48550/arXiv.2305.13269

Li, Y., Dong, B., Guerin, F., & Lin, C. (2023, December). Compressing context to enhance inference efficiency of large language models. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 6342–6353). Association for Computational Linguistics.

Li, Z., Liang, C., Peng, J., & Yin, M. (2024). The value, benefits, and concerns of generative AI-powered assistance in writing. In *Proceedings of the Chi Conference on Human Factors in Computing Systems* (pp. 1–25). Association for Computing Machinery.

Lim, G., & Perrault, S. T. (2024). Rapid aideation: Generating ideas with the self and in collaboration with large language models. arXiv Preprint, arXiv:2403.12928. https://doi.org/10.48550/arXiv.2403.12928

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., & Liang, P. (2023). Lost in the middle: How language models use long contexts. arXiv Preprint, arXiv:2307.03172. https://doi.org/10.48550/arXiv.2307.03172

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, *55*(9), 1–35. https://doi.org/10.1145/3560815

Lund, A. M. (2001). Measuring usability with the use questionnaire12. *Usability Interface*, *8*(2), 3–6.

Lyell, D., Magrabi, F., & Coiera, E. (2018). The effect of cognitive load and task complexity on automation bias in electronic prescribing. *Human Factors*, *60*(7), 1008–1021. https://doi.org/10.1177/0018720818781224

Marco, G., Gonzalo, J., del Castillo, R., & Girona, M. T. M. (2024). Pron vs. prompt: Can large language models already challenge a world-class fiction author at creative text writing? arXiv Preprint, arXiv:2407.01119. https://doi.org/10.48550/arXiv.2407.01119

Meta. (2024). Prompting—how-to guides. https://www.llama.com/docs/how-to-guides/prompting.

Microsoft. (2023a). How to use bing *AI* in *Microsoft* edge to get inspired and boost creativity. https://youtu.be/wKYqA1MLrXQ?si = FMP9KAgQ3VibLY7O

Microsoft. (2023b). Reinventing search with a new *AI*-powered *Microsoft* bing and edge, your copilot for the web. https://blogs.microsoft.com/blog/2023/02/07/reinventing-search-with-a-new-ai-powered-microsoft-bing-and-edge-your-copilot-for-the-web/.

Microsoft. (2023c). Understanding AI plugins in semantic kernel. https://learn.microsoft.com/en-us/semantic-kernel/ai-orchestration/plugins/?tabs = Csharp.

Mishra, A., Soni, U., Arunkumar, A., Huang, J., Kwon, B. C., & Bryan, C. (2023). Promptaid: Prompt exploration, perturbation, testing and iteration using visual analytics for large language models. arXiv Preprint, arXiv:2304.01964. https://doi.org/10.48550/arXiv.2304.01964

Mishra, S., & Nouri, E. (2022). Help me think: A simple prompting strategy for non-experts to create customized content with models. arXiv Preprint, arXiv:2208.08232. https://doi.org/10.48550/arXiv.2208.08232

Mueller, F. F., Lopes, P., Strohmeier, P., Ju, W., Seim, C., & Weigel, M., Others. (2020). Next steps for human-computer integration. In *Proceedings of the 2020 Chi Conference on Human Factors in Computing Systems* (pp. 1–15). Association for Computing Machinery.

Nass, C., Steuer, J., & Tauber, E. R. (1994). Computers are social actors. In *Proceedings of the Sigchi Conference on Human Factors in Computing Systems* (pp. 72–78). Association for Computing Machinery.

Nayab, S., Rossolini, G., Buttazzo, G., Manes, N., & Giacomelli, F. (2024). Concise thoughts: Impact of output length on LLM reasoning and cost. arXiv Preprint, arXiv:2407.19825. https://doi.org/10.48550/arXiv.2407.19825

Nguyen, A. T., Kharosekar, A., Krishnan, S., Krishnan, S., Tate, E., Wallace, B. C., & Lease, M. (2018). Believe it or not: Designing a human-AI partnership for mixed-initiative fact-checking. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology* (pp. 189–199). Association for Computing Machinery.

Nijkamp, E., Pang, B., Hayashi, H., Tu, L., Wang, H., Zhou, Y., & Xiong, C. (2022). Codegen: An open large language model for code with multi-turn program synthesis. arXiv preprint arXiv:2203.13474. https://doi.org/10.48550/arXiv.2203.13474

Oh, C., Song, J., Choi, J., Kim, S., Lee, S., & Suh, B. (2018). I lead, you help but only with enough details: Understanding user experience of co-creation with artificial intelligence. In *Proceedings of the 2018 Chi Conference on Human Factors in Computing Systems* (pp. 1–13). Association for Computing Machinery.

OpenAI. (2023a). *OpenAI API*. https://platform.openai.com/docs/plugins/getting-started/plugin-manifest.

OpenAI. (2023b). *OpenAI platform*. https://platform.openai.com/docs/guides/gpt.

OpenAI. (2025). *Pricing—openAI*. https://openai.com/api/pricing/

Park, H., & Ahn, D. (2024). The promise and peril of chatgpt in higher education: Opportunities, challenges, and design implications. Proceedings of the Chi Conference on Human Factors in Computing Systems (pp. 1–21). Association for Computing Machinery.

Park, S., Li, H., Patel, A., Mudgal, S., Lee, S., Kim, Y.-B., & Sarikaya, R. (2021). A scalable framework for learning from implicit user feedback to improve natural language understanding in large-scale conversational AI systems. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 6054–6063). Association for Computational Linguistics.

Petridis, S., Diakopoulos, N., Crowston, K., Hansen, M., Henderson, K., Jastrzebski, S., & Chilton, L. B. (2023). Anglekindling: Supporting journalistic angle ideation with large language models. In *Proceedings of the 2023 Chi Conference on Human Factors in Computing Systems* (pp. 1–16). Association for Computing Machinery.

Pingouin. (2024). *Pingouin.pairwise tukey*. https://pingouin-stats.org/build/html/generated/pingouin.pairwise tukey.html.

Qin, H. X., Jin, S., Gao, Z., Fan, M., & Hui, P. (2024). Charactermeet: Supporting creative writers' entire story character construction processes through conversation with LLM-powered chatbot avatars. In *Proceedings of the Chi Conference on Human Factors in Computing Systems* (pp. 1–19). Association for Computing Machinery.

Rao, S., & Daumé III, H., (2018, July). Learning to ask good questions: Ranking clarification questions using neural expected value of perfect information. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (vol. 1: Long papers, pp. 2737–2746). Association for Computational Linguistics.

Rao, S., & Daumé III, H., (2019). Answer-based adversarial training for generating clarification questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 143–155). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1013

Rastogi, C., Tulio Ribeiro, M., King, N., Nori, H., & Amershi, S. (2023). Supporting human-AI collaboration in auditing LLMS with LLMS. In *Proceedings of the 2023 AAAI/ACM Conference on AI,*

Ethics, and Society (pp. 913–926). Association for Computing Machinery.

Renze, M., & Guven, E. (2024). The benefits of a concise chain of thought on problem-solving in large language models. In *Proceedings of the 2nd International Conference on Foundation and Large Language Models (FLLM 2024)* (pp. 476–483). IEEE. https://doi.org/10.1109/FLLM63129.2024.10852493

Reuters. (2023). *Chatgpt's explosive growth shows first decline in traffic since launch*. https://www.reuters.com/technology/booming-traffic-openais-chatgpt-posts-first-ever-monthly-d.

Reza, M., Laundry, N. M., Musabirov, I., Dushniku, P., Yu, Z. Y. M., Mittal, K., & Williams, J. J. (2024). Abscribe: Rapid exploration & organization of multiple writing variations in human-AI co-writing tasks using large language models. In *Proceedings of the Chi Conference on Human Factors in Computing Systems* (pp. 1–18). Association for Computing Machinery.

Riche, Y., Henry Riche, N., Isenberg, P., & Bezerianos, A. (2010). Hard-to-use interfaces considered beneficial (some of the time). In *CHI'10 Extended Abstracts on Human Factors in Computing Systems* (pp. 2705–2714). Association for Computing Machinery.

Rodrigues Barbosa, G. A., da Silva Fernandes, U., Sales Santos, N., & Oliveira Prates, R. (2024). Human-computer integration as an extension of interaction: Understanding its state-of-the-art and the next challenges. *International Journal of Human–Computer Interaction*, *40*(11), 2761–2780. https://doi.org/10.1080/10447318.2023.2177797

Ross, S. I., Martinez, F., Houde, S., Muller, M., & Weisz, J. D. (2023). The programmers assistant: Conversational interaction with a large language model for software development. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (pp. 491–514). Association for Computing Machinery.

Schaekermann, M., Goh, J., Larson, K., & Law, E. (2018). Resolvable vs. irresolvable disagreement: A study on worker deliberation in crowd work. In *Proceedings of the ACM on Human-Computer Interaction* (vol. 2, pp. 1–19). CSCW.

Sekulíc, I., Aliannejadi, M., & Crestani, F. (2021). Towards facet-driven generation of clarifying questions for conversational search. In *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval* (pp. 167–175). Association for Computing Machinery.

Shakeri, H., Neustaedter, C., & DiPaola, S. (2021). *Saga: Collaborative storytelling with gpt-3* [Paper presentation]. Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing (pp. 163–166), Virtual Event, USA.

Shi, W., Min, S., Yasunaga, M., Seo, M., James, R., Lewis, M., & Yih, W-T (2024, June). REPLUG: Retrieval-augmented black-box language models. In K. Duh, H. Gomez, & S. Bethard (Eds.), In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (vol. 1: Long papers, pp. 8371–8384). Association for Computational Linguistics.

Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, *36*(6), 495–504. https://doi.org/10.1080/10447318.2020.1741118

Shneiderman, B., & Maes, P. (1997). Direct manipulation vs. interface agents. *Interactions*, *4*(6), 42–61. https://doi.org/10.1145/267505.267514

Skjuve, M., Følstad, A., & Brandtzaeg, P. B. (2023). The user experience of chatgpt: Findings from a questionnaire study of early users. In *Proceedings of the 5th International Conference on Conversational User Interfaces* (pp. 1–10). Association for Computing Machinery.

Sundararajan, N., & Adesope, O. (2020). Keep it coherent: A meta-analysis of the seductive details effect. *Educational Psychology Review*, *32*(3), 707–734. https://doi.org/10.1007/s10648-020-09522-4

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4

Sweller, J. (2011). Cognitive load theory. In J. P. Mestre & B. H. Ross (Eds.), *Psychology of learning and motivation* (Vol. 55, pp. 37–76). Academic Press. https://doi.org/10.1016/B978-0-12-387691-1.00002-8

Sweller, J. (2020). Cognitive load theory and educational technology. *Educational Technology Research and Development*, *68*(1), 1–16. https://doi.org/10.1007/s11423-019-09701-3

Sweller, J. (2024). Cognitive load theory and individual differences. *Learning and Individual Differences*, *110*, 102423. https://doi.org/10.1016/j.lindif.2024.102423

Walker, M., & Whittaker, S. (1990). Mixed initiative in dialogue: An investigation into discourse segmentation. In *Proceedings of the 28th Annual Meeting on Association for Computational Linguistics* (p. 7078). Association for Computational Linguistics.

Wan, Q., Hu, S., Zhang, Y., Wang, P., Wen, B., & Lu, Z. (2024). "It felt like having a second mind": Investigating human-AI co-creativity in prewriting with large language models. *Proceedings of the ACM on Human-Computer Interaction*, *8*(CSCW1), 1–26. https://doi.org/10.1145/3637361

Wan, Z., Wang, X., Liu, C., Alam, S., Zheng, Y., & Liu, J., Others. (2023). Efficient large language models: A survey. arXiv preprint arXiv:2312.03863. https://doi.org/10.48550/arXiv.2312.03863

Wang, L., Yang, N., & Wei, F. (2023, December). Query2doc: Query expansion with large language models. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 conference on empirical methods in natural language processing* (pp. 9414–9423). Association for Computational Linguistics.

Wasi, A. T., Islam, R., & Islam, M. R. (2024). Ink and individuality: Crafting a personalised narrative in the age of LLMS. In *Proceedings of the Third Workshop on Intelligent and Interactive Writing Assistants* (pp. 43–47). Association for Computing Machinery.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., & Chi, E., Others. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, *35*, 24824–24837. https://doi.org/10.5555/3600270.3602070

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. In *Proceedings of the 30th Conference on Pattern Languages of Programs (PLoP '23)* (Article 5, pp. 1–31). The Hillside Group. https://doi.org/10.5555/3721041.3721046

Wu, T., Jiang, E., Donsbach, A., Gray, J., Molina, A., Terry, M., & Cai, C. J. (2022). *Promptchainer: Chaining large language model prompts through visual programming* [Paper presentation]. Chi Conference on Human Factors in Computing Systems Extended Abstracts (pp. 1–10).

Wu, T., Terry, M., & Cai, C. J. (2022). AI chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In *Proceedings of the 2022 Chi Conference on Human Factors in Computing Systems* (pp. 1–22). Association for Computing Machinery.

Xi, Y., Liu, W., Lin, J., Cai, X., Zhu, H., Zhu, J., & Yu, Y. (2024). Towards open-world recommendation with knowledge augmentation from large language models. In *Proceedings of the 18th ACM Conference on Recommender Systems* (pp. 12–22). Association for Computing Machinery.

Xu, R., Feng, Y., & Chen, H. (2023). Chatgpt vs. google: A comparative study of search performance and user experience. arXiv preprint arXiv:2307.01135. https://doi.org/10.48550/arXiv.2307.01135

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., & Narasimhan, K. (2023). Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, *36*, 8706–8719. https://doi.org/10.5555/3666122.3666639

Yuan, A., Coenen, A., Reif, E., & Ippolito, D. (2022). *Wordcraft: Story writing with large language models. 27th International Conference on Intelligent User Interfaces* (pp. 841–852). Association for Computing Machinery.

Zamani, H., Dumais, S., Craswell, N., Bennett, P., & Lueck, G. (2020). Generating clarifying questions for information retrieval. In *Proceedings of the Web Conference 2020* (pp. 418–428). Association for Computing Machinery.

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

Human Factors in Computing Systems (pp. 1–21). Association for Computing Machinery.

Zhai, C. (2024). Large language models and future of information retrieval: Opportunities and challenges. In *Proceedings of the 47th international ACM SIGIR conference on research and development in information retrieval* (pp. 481-490). Association for Computing Machinery.

Zhang, A., Chen, Y., Sheng, L., Wang, X., & Chua, T.-S. (2024). On generative agents in recommendation. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1807–1817). Association for Computing Machinery.

Zhu, Y., Yuan, H., Wang, S., Liu, J., Liu, W., Deng, C., & Wen, J.-R. (2023). Large language models for information retrieval: A survey. arXiv Preprint, arXiv:2308.07107. https://doi.org/10.48550/arXiv.2308.07107

About the authors

Soomin Kim obtained her PhD in Communication from Seoul National University. She is currently working as a senior designer. Her research focuses on Human-AI Interaction, conversational agents, and user experience design. She integrates user experience insights with AI approaches in her studies.

Jinsu Eun, a PhD candidate at Seoul National University's Institute of Convergence Science and Technology, excels in data visualization, robotic journalism, and chatbot tech. His research focuses on innovative, user-friendly methods to present complex data, combining advanced technology with user-centered design for effective communication.

Yoobin Elyson Park, with an MA in Communication from Seoul National University, is currently working as a UX Strategist. Her research interests pivot around designing experiences to seamlessly integrate AI into everyday human workflows, leveraging AI for human creativity and productivity, and understanding user perceptions of AI and AI-generated content.

Kwangwon Lee is a master's student in the Interdisciplinary Program in Artificial Intelligence at Seoul National University. His research interest lies in Human-AI Interaction with emphasis on optimal decision making and broadening accessibility through AI. His familiarity and enthusiasm towards the latest ML technology is one of his strengths.

Gyuho Lee is a PhD candidate in the Department of Communication at Seoul National University. His research focuses on user communication and collective behavior within social media platforms. He aims to advance communication research methodologies by incorporating techniques from natural language processing, social network analysis, and data visualization.

Joonhwan Lee is a Professor in the Department of Communication at Seoul National University. He holds a PhD in Human–Computer Interaction (2008) from Carnegie Mellon University. His research includes HCI, social computing, situationally appropriate user interaction, and information visualization. He directs the Human-Computer Interaction + Design Lab (hcid.snu.ac.kr).