frontiers | Frontiers in Psychology

TYPE Original Research
PUBLISHED 28 March 2024
DOI 10.3389/fpsyg.2024.1382234

OPEN ACCESS

EDITED BY
Martina Rau,
University of Wisconsin-Madison,
United States

REVIEWED BY
Thomas Neumayr,
University of Applied Sciences Upper Austria,
Austria
Matthias Wölfel,
Karlsruhe University of Applied Sciences,
Germany

*CORRESPONDENCE
Yusuke Yamamoto
✉ yusuke_yamamoto@acm.org

RECEIVED 05 February 2024
ACCEPTED 14 March 2024
PUBLISHED 28 March 2024

CITATION
Yamamoto Y (2024) Suggestive answers strategy in human-chatbot interaction: a route to engaged critical decision making. *Front. Psychol.* 15:1382234. doi: 10.3389/fpsyg.2024.1382234

Suggestive answers strategy in human-chatbot interaction: a route to engaged critical decision making

Yusuke Yamamoto*

School of Data Science, Nagoya City University, Nagoya, Japan

In this study, we proposed a novel chatbot interaction strategy based on the *suggestive ending* of answers. This strategy is inspired by the cliffhanger ending narrative technique, which ends a story without specifying conclusions to spark readers' curiosity as to what will happen next and is often used in television series. Common chatbots provide relevant and comprehensive answers to users' questions. In contrast, chatbots with our proposed strategy end their answers with hints potentially interest-triggering users. The suggestive ending strategy aims to stimulate users' inquisition for critical decision-making, relating to a psychological phenomenon where humans are often urged to finish the uncompleted tasks they have initiated. We demonstrated the implication of our strategy by conducting an online user study involving 300 participants, where they used chatbots to perform three decision-making tasks. We adopted a between-subjects factorial experimental design and compared between the following UIs: (1) *plain* chatbot—it provides a generated answer when participants issue a question; (2) *expositive* chatbot—it provides a generated answer for a question, adding short summaries of a positive and negative person's opinion for the answer; (3) *suggestive* chatbot—it provides a generated answer for a question, which ends with a suggestion of a positive and negative person for the answer. We found that users of the *suggestive* chatbot were inclined to ask more questions to the bot, engage in prolonged decision-making and information-seeking actions, and formulate their opinions from various perspectives. These findings vary with the users' experience with *plain* and *expositive* chatbots.

KEYWORDS

human-AI interaction, large language model, behavior change, critical information-seeking, chatbot

1 Introduction

Recent advancements in artificial intelligence (AI), particularly the remarkable evolution of large language models (LLMs), have given rise to a lot of services and applications that support human tasks in various domains. Generative AI with LLMs holds a strong potential for substantially changing human intellectual activities. For example, instruction-tuned LLMs (e.g., ChatGPT) can quickly generate surprisingly natural sentences in response to human questions (Wei et al., 2021). Zylowski and Wölfel (2023) revealed that when specifying personas for ChatGPT in prompts enables it to simulate a variety of personalities and capabilities. OpenAI reported that ChatGPT scored 1,300/1,600 on the SAT[1] by eliciting knowledge in its language model[2]. In 2024, Google released *Gemini Ultra*, the highly capable LLM which outperforms GPT-4 on text-based tasks, including reasoning, reading comprehension, and code generation (Team et al., 2023). Furthermore, an appropriate understanding of LLM applications and their effective use can equally support decision-making and opinion formulation (Wambsganss et al., 2020, 2021; Jakesch et al., 2023; Petridis et al., 2023).

Despite their superlative functionalities, generative AIs with LLMs often generate incorrect, biased, or unrealistic information, a phenomenon known as *hallucination* (Maynez et al., 2020). Overreliance on AIs causes automation bias to users (Goddard et al., 2011), leading to the ubiquitous obliviousness of AI-generated false information (Lakkaraju and Bastani, 2020). Studies have shown that overusing AIs can inhibit the development of users' cognitive skills (Noyes, 2007; Carr, 2014), naturally affecting their critical thinking abilities. As a result, users can be unconsciously led to a specific polarity by opinionated AI assistants for writing (Jakesch et al., 2023). These aspects raise serious educational concerns. For instance, students using generative AI-powered chatbots can accept harmful/incorrect information without doubt, which strongly affects the development of their critical thinking and problem-solving skills (Kasneci et al., 2023).

Although the research on improving the performance of generative AIs with LLMs is under extensive development, undesirable output information remains highly probable (Wei et al., 2021; Nakano et al., 2022; Tay et al., 2022; Wang et al., 2023). This probability is particularly aggravated by the human *confirmation bias*, defined as the tendency to preferentially view or search for information consistent with one's opinions or hypotheses (Kahneman, 2011). Therefore, improving generative AIs should be accompanied by an effective design of human–AI interactions that promote users' cognitive activities for critical decision-making or opinion formulation.

In this study, we proposed a novel human–chatbot interaction strategy, *suggestive ending*, for generative AI-powered chatbot answers to foster decision-making from various perspectives. Our method is inspired by the *cliffhanger ending* narrative technique, which ends a story without specifying conclusions to spark readers' curiosity as to what will happen next. The cliffhanger method is often used in television series. It relates to a psychological phenomenon known as the *Ovsiankina effect*, where humans are often urged to finish the uncompleted tasks they have initiated (Wirz et al., 2023). *Suggestive* bots employed with the proposed strategy output their answers with hints to potentially interest-triggering subjects (Figure 1B). In contrast, common chatbots provide relevant and comprehensive answers to users' questions (Figure 1A). Therefore, when interacting with SUGGESTIVE chatbots on a given theme, users' proactive critical decision-making is stimulated by intentionally leaving room for questions.

We conducted an online user study involving 300 participants to validate our proposed method on the human–AI interaction. Results revealed the following three primary observations.

- When using the SUGGESTIVE bot, participants engaged in decision-making by inputting questions to the bot. This has led participants to spend longer interactions with the SUGGESTIVE bot than the PLAIN (i.e., simply providing relevant answers) and EXPOSITIVE (i.e., providing relevant answers with supplementary information) bots.
- Compared to the PLAIN bot, participants were likely to write longer opinions from various perspectives using the SUGGESTIVE bot.
- When using either the EXPOSITIVE or PLAIN bot, participants showed similar efforts in their decision-making activities.

2 Related work

2.1 AI-assisted decision-making and opinion formation

AI systems developed to assist decision-making and opinion formation have been studied from the viewpoints of supporting interpretation of AI predictions, improving the understanding of arguments, enhancing the efficiency of opinion formulation, searching for supportive information, etc.

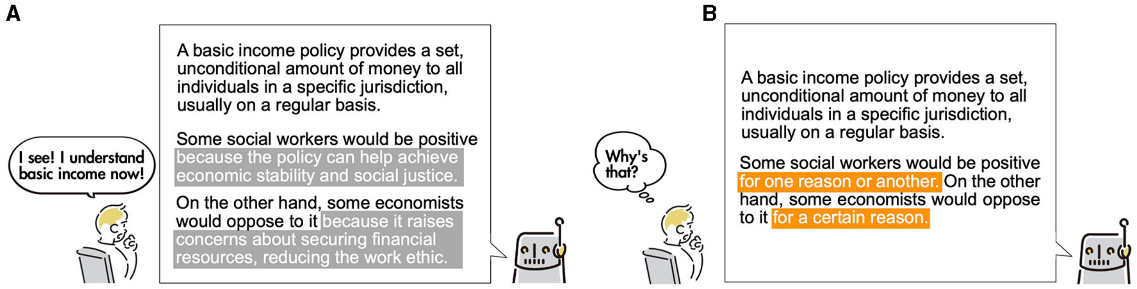
It is essential to understand how and why black-box AIs provide predictions for users to efficiently use them during decision-making. Hence, many researchers have studied *explainable AI* technologies to improve the interpretability of machine learning (ML) models. For instance, for ML behaviors on structured data, researchers have proposed various methods to summarize the contributed features to predictions (Lundberg and Lee, 2017; Fisher et al., 2019) and explain how the models work with data examples (Kim et al., 2016). Lakkaraju and Bastani (2020) reported misleading explanations on black-box MLs as a cause for users trusting even harmful MLs. Therefore, considering the characteristics of human design thinking is important to improve the interpretability of AIs for decision-making.

To better understand the aforementioned arguments, Wambsganss et al. (2021) proposed ARGUETUTOR, a chatbot system that provides users with feedback to identify sentences in their documents that require logical improvement. Furthermore, they proposed an interactive system to visualize the argumentation structure of a given document, thereby helping users make more logical judgments (Wambsganss et al., 2020). Petridis et al. (2023) developed ANGLEKINDLING, a system that supports journalists in exploring points to scrutinize potential negative impacts on press releases using an LLM.

Several investigations have been conducted on *suggestive keyboards* to support efficient opinion formulation. Arnold et al. (2016) proposed a phrase-suggesting method for text composition instead of predicting words following users' input texts. However, *suggestive keyboard* technologies could affect what to write. In another study, Arnold et al. (2020) reported that *suggestive*

[1] The Scholastic Assessment Test (SAT) is a standardized test for college admissions across the United States. It is designed to assess students' proficiency in mathematics, reading, and writing.

[2] https://www.nytimes.com/2023/03/14/technology/openai-new-gpt4.html



A basic income policy provides a set, unconditional amount of money to all individuals in a specific jurisdiction, usually on a regular basis.

Some social workers would be positive because the policy can help achieve economic stability and social justice.

On the other hand, some economists would oppose to it because it raises concerns about securing financial resources, reducing the work ethic.

A basic income policy provides a set, unconditional amount of money to all individuals in a specific jurisdiction, usually on a regular basis.

Some social workers would be positive for one reason or another. On the other hand, some economists would oppose to it for a certain reason.

FIGURE 1
Comparison between *expositive* chatbot and *suggestive* chatbot. **(A)** Expositive chatbots provide relevant answers to questions and brief supplementary information on a specific perspective. **(B)** Suggestive chatbots (the proposed bot) provide answers ending with suggestive messages hinting at something. Suggestive chatbots encourage users to ask spontaneous questions for proactive critical decision-making.

keyboard technologies affect the writers' choices who often follow AI-based text suggestions, while it improves their writing speed. Jakesch et al. (2023) reported that when suggestive keyboards were used, in which an LLM was fine-tuned to suggest positive (negative) phrases, users were likely to write positive (negative) opinions.

In the field of information retrieval (IR), several researchers have investigated systems for searching information to support decision-making. Rinott et al. (2015) proposed a method to search for evidence supporting a given claim from unstructured documents. Liu et al. (2022) proposed CRYSTALLINE, a browser developed to tabulate collected Web information for efficient decision-making.

The aforementioned studies reveal that users with sufficient skills and motivation to properly use advanced AI technologies can obtain useful assistance from these technologies in decision-making and opinion formulation. Otherwise, overreliance on AIs for decision-making causes negative impacts on users, including shortsighted decision making, cognitive downskilling, and opinion radicalization. Therefore, our proposed method focuses on eliciting questions from chatbot users and promoting active opinion formation in the human–chatbot interaction.

2.2 Generative information retrieval

With the emergence of LLMs, changes were introduced to the conventional IR model, which aims to provide a ranked list of relevant documents for a keyword query. *Generative IR* is a new LLMs-based paradigm of generating information to directly answer users' questions. When a question is given, typical generative IR systems (i.e., AI-powered chatbots) extend prompts with likely completions and extract answers from the extended prompts (Najork, 2023).

ChatGPT[3] and Google Gemini[4] are recently developed generative IR applications that have spurred unprecedented universal attention. Nevertheless, ongoing research is highlighting their drawbacks, such as generating incorrect or unrealistic answers, which is known as the *hallucination* phenomenon (Maynez et al., 2020). Metzler et al. (2021) reported several challenges in generative IRs, such as suggesting contexts for generated answers and considering the authority or quality of documents for answer generation. Several methods have been proposed to tackle these challenges, such as tuning LLM models for human-favorable answers (Wei et al., 2021), linking generated answers (or questions) with relevant documents (Nakano et al., 2022; Tay et al., 2022), and improving the interpretability of generative AIs (Sun et al., 2022). Furthermore, Wang et al. (2023) proposed *Shepherd*, an LLM model that provides feedback to improve target LLMs by analyzing the generated texts.

While the aforementioned studies focus on the performance improvement and high functionality of generative IR systems, our proposed method focuses on enhancing users' information-seeking and cognitive activities in generative IRs.

2.3 Enhancing critical information seeking and decision-making

Studies conducted to activate and enhance information-seeking and decision-making abilities can be categorized into two approaches for steering and empowering better judgments: *nudging* (Thaler and Sunstein, 2009; Caraban et al., 2019) and *boosting* (Hertwig and Grne-Yanoff, 2017).

Nudging is defined as "*an approach to alter people's behavior in a predictable way without forbidding any option or substantially changing their economic incentive*" (Thaler and Sunstein, 2009). In the field of HCI, several methods have been reported for the application of this concept, which include supporting critical information seeking (Yamamoto and Yamamoto, 2018; Saito et al., 2020; Ihoriya et al., 2022; Suzuki and Yamamoto, 2022) and enhancing privacy awareness on the Web (Wang et al., 2013; Zimmerman et al., 2019; Yamamoto and Yamamoto, 2020). For example, Yamamoto and Yamamoto (2018) proposed the QUERY PRIMING system, which inserts queries to evoke critical thinking during query completion/recommendation in a search system. Suzuki and Yamamoto (2022) proposed a search user interface (UI) that makes web searchers reflect on their webpage selection criteria and promote content-quality-oriented web searches regardless of visual appearances. Wang et al. (2013) proposed a privacy nudge

[3] ChatGPT: https://chat.openai.com/.

[4] Google Gemini: https://gemini.google.com/.

that shows Facebook profile pictures of the target audience when users post content on Facebook to enhance users' awareness to potential risks.

Nudging supports better decision-making by focusing on related systematic biases. However, *boosting* is an intervention to improve cognitive competence for proactive and rational decision-making (Hertwig and Grne-Yanoff, 2017). Shimizu et al. (2022) proposed *privacy-aware* snippets, which aim to enhance privacy risk judgment in Web searches by providing comprehensive information about sharing conditions of browsing histories. Harvey et al. (2015) reported that providing examples of high-quality queries can help users learn to improve the efficiency of their query formulation. Buçinca et al. (2021) reported that the users' final decision-making performance can be improved if they are required to think by themselves before the AIs provide supportive information for decision-making. Danry et al. (2023) reported that when AIs ask people a simple question to confirm a claim's logical validity, reasoning activities can be activated, and the fallacy identification performance can be improved.

While questioning approaches such as Danry et al. (2023)'s method are explicit boosting (i.e., instructive intervention), our method is regarded as implicit boosting (i.e., modest intervention). Our proposed strategy aims to trigger users' spontaneous questions through their interaction with chatbots, introducing suggestive messages in answers and leaving room for further questioning. We expect that our suggestive ending approach will be perceived as less intrusive than instructive questioning approaches.

3 Research questions

Our proposed *suggestive ending* strategy in chatbots for IR aims to provoke users' questions on a given theme or prior belief, driving them to make theme-dependent critical decisions. Ennis (1987) defined critical thinking as logical and reflective thinking to determine what to believe or do. Furthermore, the author claimed that ideal critical thinkers are disposed to seek reasons, consider entire situations, look for alternatives, and use critical thinking, e.g., deductive reasoning. Several studies revealed the effect of *lateral reading*, a method to check multiple information resources in parallel for critical review on a theme (Meola, 2004; Wineburg and McGrew, 2019; Brodsky et al., 2021). We expect that if chatbots implicitly suggest the existence of things to check at the end of their responses, users would be more willing to critically construct their opinions and gather information for validation compared to cases where chatbots provide detailed explanatory answers.

To explore the validity of our proposed strategy using SUGGESTIVE bots, we considered the following research questions:

RQ1: Do SUGGESTIVE bots engage users in investing additional effort to form their opinions and gather information for decision-making?

RQ2: Do SUGGESTIVE bots encourage users to consider various perspectives when making their decision?

As we are interested in exploring whether SUGGESTIVE bots should actively nudge people to question the details of the bot's ambiguous endings, we also investigated the following research question:

RQ3: Do question (query) suggestions along with suggestive bot's answer promote more critical decision-making?

According to the elaboration likelihood model theory proposed by Petty and Cacioppo (1986), people often pay more attention to information in which they have sufficient knowledge or strong understanding interest. Otherwise, they often use poor judgment for accepting or rejecting the information. Based on this theory, individual factors can affect people's effort and behavior in decision-making tasks as well as suggestive bot's behaviors. Therefore, we have also formulated the following research question:

RQ4: Do individual factors, such as knowledge, interest, and familiarity with the information sought using chatbots, affect associated decision-making tasks?

4 Materials and methods

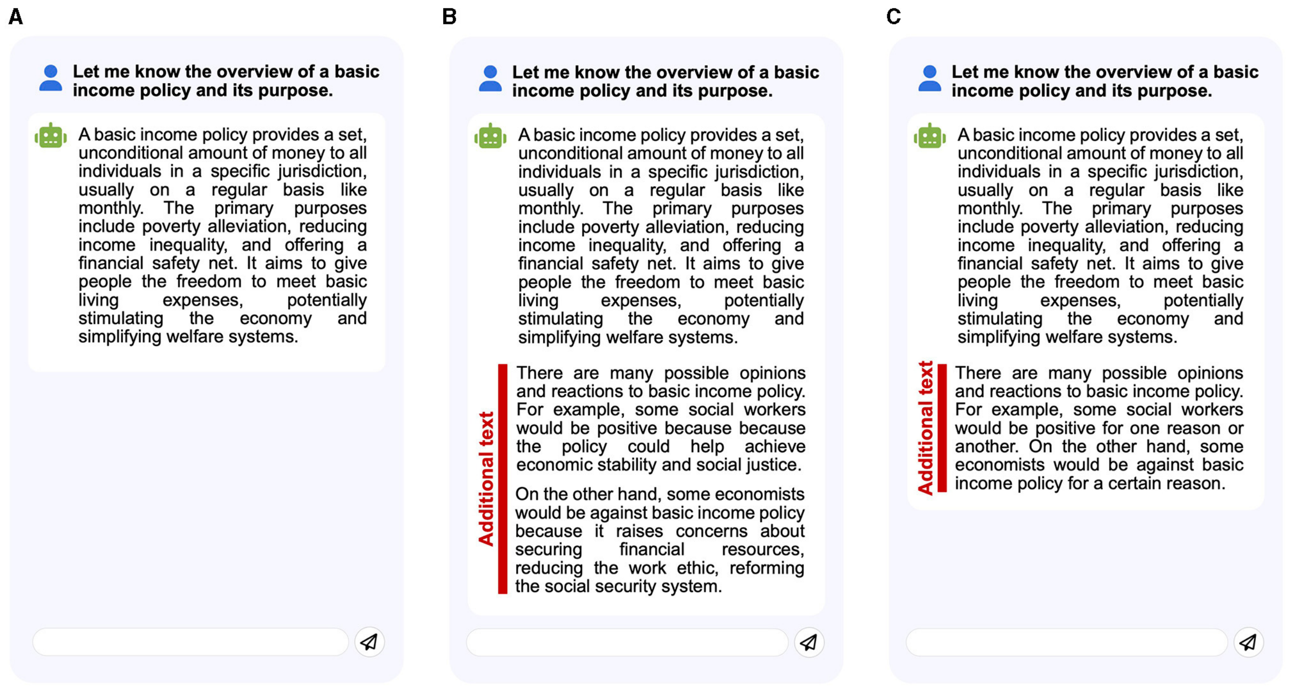
We conducted an online user study to investigate the effect of suggestive ending in AI-powered chatbots on decision-making tasks. The user study was conducted in Japanese (on August 11 and 12, 2023). For this, we adopted a between-subjects factorial experimental design, where the factor is a user interface (UI) condition with four levels:

1. PLAIN chatbot: it provides a generated answer when participants issue a question (query) (Figure 2A).
2. EXPOSITIVE chatbot: provides a generated answer for a question, adding short summaries of a positive and negative person's opinion for the answer (Figure 2B).
3. SUGGESTIVE chatbot: it provides a generated answer for a question, which ends with a suggestion of a positive and negative person for the answer. Unlike the EXPOSITIVE chatbot, this bot does not show short opinion summaries (Figure 2C).
4. SUGGESTIVE^{++} chatbot: as an extension of the SUGGESTIVE chatbot, it provides links to ask about suggested positive/negative people's opinions, following a generated answer (Figure 4).

We conducted a user study on a crowdsourcing platform. Crowdsourcing platforms such as Amazon Mechanical Turk[5] and Prolific[6] enable researchers to recruit a large number of participants via the internet at lower costs compared to traditional survey companies. Consequently, user studies with crowdsourcing have been becoming popular in the communities of Human-Computer Interaction (HCI) (Kittur et al., 2008; Komarov et al., 2013) and Information Retrieval (IR) (Yamamoto and Yamamoto, 2018; Câmara et al., 2021; Roy et al., 2021) as an alternative way to laboratory-based experiments. Numerous studies have examined the reliability of crowdsourcing by comparing crowd workers' performance to that of participants in laboratory settings (Lutz, 2016; Peer et al., 2017; Hettiachchi et al., 2022). These studies have demonstrated that crowdsourcing can be reliable for conducting user studies, provided that the online tasks are designed to control experimental environments and mitigate satisficing behaviors—whereby participants make judgments or complete tasks with

[5] Amazon Mechanical Turk: https://www.mturk.com/.

[6] https://www.prolific.com/



A

Let me know the overview of a basic income policy and its purpose.

A basic income policy provides a set, unconditional amount of money to all individuals in a specific jurisdiction, usually on a regular basis like monthly. The primary purposes include poverty alleviation, reducing income inequality, and offering a financial safety net. It aims to give people the freedom to meet basic living expenses, potentially stimulating the economy and simplifying welfare systems.

B

Let me know the overview of a basic income policy and its purpose.

A basic income policy provides a set, unconditional amount of money to all individuals in a specific jurisdiction, usually on a regular basis like monthly. The primary purposes include poverty alleviation, reducing income inequality, and offering a financial safety net. It aims to give people the freedom to meet basic living expenses, potentially stimulating the economy and simplifying welfare systems.

There are many possible opinions and reactions to basic income policy. For example, some social workers would be positive because because the policy could help achieve economic stability and social justice.

On the other hand, some economists would be against basic income policy because it raises concerns about securing financial resources, reducing the work ethic, reforming the social security system.

C

Let me know the overview of a basic income policy and its purpose.

A basic income policy provides a set, unconditional amount of money to all individuals in a specific jurisdiction, usually on a regular basis like monthly. The primary purposes include poverty alleviation, reducing income inequality, and offering a financial safety net. It aims to give people the freedom to meet basic living expenses, potentially stimulating the economy and simplifying welfare systems.

There are many possible opinions and reactions to basic income policy. For example, some social workers would be positive for one reason or another. On the other hand, some economists would be against basic income policy for a certain reason.

FIGURE 2
Chatbot user interfaces in our user study. **(A)** PLAIN bot, **(B)** EXPOSITIVE bot, and **(C)** SUGGESTIVE bot. The differential information between chatbots is represented by red bars (The chatbots did not show these red bars and texts to the participants).

minimal effort. In light of these findings, we conducted a user study with a crowdsourcing service to examine the effectiveness of our proposed method. Note that we implemented an instructional manipulation check (IMC), a popular technique to identify inattentive crowd workers, to ensure the integrity of our data collection process. Furthermore, we rejected crowd workers using mobile/tablet devices so that all participants could perform tasks on the same layout on their PCs.

Participants were randomly allocated into one of the above four UIs. They then conducted tasks to summarize their opinions about three randomly allocated themes. To consider individual differences, we measured and analyzed personal factors as covariates, including the frequency of using chatbots for information seeking, interest in task themes, and familiarity with the themes. We designed this user study following the research ethics guidelines of our affiliated organization.

4.1 Themes for decision-making tasks

We prepared eight themes for decision-making tasks and one theme for practice tasks. The themes were prepared from the website of the National Association of Debate in Education, Japan. We selected the frequently used nine themes in debate competitions for high school students in Japan, as listed in Table 1. As presented in Table 1, the impressions of participants indicated their unfamiliarity with most themes on average. Moreover, the interests of participants were slightly inclined to positive polarity on average (excluding *making doggy bags available at restaurants*).

4.2 Chatbots

The aforementioned four UI conditions (chatbots) employed ChatGPT, OpenAI instruction-tuned LLM, via Azure OpenAI Service GPT API (gpt-3.5-turbo[7]) to generate answers for participants' questions. In particular, we used an LLM prompt-engineering technique in the SUGGESTIVE, SUGGESTIVE^{++}, and EXPOSITIVE bots to complement additional information with plain answers for questions.

One possible way in our proposed *suggestive ending* strategy in chatbots is to suggest perspectives for decision-making explicitly, such as key issues (Câmara et al., 2021; Petridis et al., 2023) and positive/negative aspects for themes (Liao and Fu, 2014; Liao et al., 2015). However, such explicit suggestions are revealing and do not encourage users to proactively reflect on what they should think for their decision-making. On the one hand, studies in the field of learning science indicate that contents should leave proper room for questioning and discussion so that people would be willing to learn a theme and deepen their knowledge (King, 1992). On the other hand, it is difficult for users to find important questions and perspectives for a theme if they lack knowledge and interest.

Therefore, we designed two types of chatbots, namely, (SUGGESTIVE and SUGGESTIVE^{++}), to provide direct answers to users' questions and additional suggestions on the existence of positive and negative people for a theme, respectively. The two chatbots never suggest the kind of perspectives the

[7] Azure OpenAI Service: https://azure.microsoft.com/en-us/products/ai-services/openai-service-b.

TABLE 1 Themes for decision-making tasks and corresponding participants' impressions.

Theme	Interest	Familiarity	#Exp. perspectives
Introduction of daylight saving time	3.77 (1.50)	3.63 (1.44)	5
Introduction of carbon tax	3.77 (1.60)	2.34 (1.37)	5
Charging for ambulance	4.81 (1.23)	2.74 (1.34)	5
Making doggy bags available at restaurants	2.95 (1.64)	1.72 (1.14)	4
Restrictions on whale fishing	3.88 (1.50)	3.22 (1.40)	6
Sales promotion of genome-edited food	3.59 (1.65)	2.05 (1.17)	4
Expanding acceptance of foreign workers	4.73 (1.27)	3.30 (1.38)	6
Restrictions on fake news	4.58 (1.37)	3.37 (1.30)	4
Introduction of universal basic income system (for practice task)	NA	NA	NA

Interest and familiarity use a seven-point scale (1, not at all; 4, neutral; 7, very much). Numbers in the table indicate the mean and standard deviation (in parentheses). #Exp. perspectives mean the number of expected perspectives for each theme.

positive/negative people can have before users explicitly ask about them.

4.2.1 Suggestive bot

This chatbot suggests examples of positive and negative people for a decision-making theme when the participants ask an *initial question*, an overview of a given theme, and its purpose (Figure 2C). As described in Section 4.3, just after each decision-making task started, we predefined an initial question (query) about an overview of a theme and set it in the query box of the chatbot. When accepting the initial question, the SUGGESTIVE bot generated an answer for the question. The bot then suggested an example of a positive and negative persons at the end of the generated answer using the following sentence:

> *"There are many possible opinions and reactions to [THEME]. For example, Some [POSITIVE PERSON] would be positive for one reason or another. However, some [NEGATIVE PERSON] would be against [THEME] for a certain reason".*

The SUGGESTIVE bot finds an example of positive/negative people for a theme as follows:

1. The bot generates an answer (referred to as *initial answer*) for an initial question about a theme by simply fetching Azure OpenAI API with the initial question.
2. The bot gathers a list of people who might have positive/negative feelings for the *initial answer* using the prompt illustrated in Figure 3A.
3. The bot randomly picks up a positive and negative person.

Before the user study, we cached an initial answer and a list of positive/negative people for each theme in Table 1. During the study, the SUGGESTIVE bot used the cached results for suggestive answer generation so as not to fail due to OpenAI API error.

4.2.2 Suggestive++ bot

The SUGGESTIVE^{++} bot is an extension of the SUGGESTIVE bot. When providing participants with initial answers with suggestive endings, SUGGESTIVE^{++} displays links to question what opinions a suggested positive/negative person might have for a given theme (referred to as *suggestive links*). Once the participants click a suggestive link to a positive/negative person, the SUGGESTIVE^{++} bot displays the person's opinions against a task theme (Figure 4).

As illustrated in Figure 3B, each positive/negative person's opinion is generated via OpenAI's API using the prompt to question what opinions the person might have for the *initial answer* text. Similar to *initial answers*, the SUGGESTIVE^{++} bot suggests an example of a positive and a negative person at the end of the generated opinions. In addition, the bot lists *suggestive links* to other people's opinions. In other words, once they click a *suggestive link*, the participants could see other *suggestive links*. Similar to the case of *initial answers*, we generated and cached positive/negative people's opinion texts for the task themes before the user study. We expected that the SUGGESTIVE^{++} bot could encourage the participants to recall more easy-to-query questions than SUGGESTIVE bot.

4.2.3 Expositive bot

In addition to suggesting the existence of positive/negative people, the EXPOSITIVE shows one-line summaries of their opinions in the *initial answers* as supplementary information. Participants using EXPOSITIVE bots can briefly learn the possible perspectives or opinions of a positive and negative person without additional questioning.

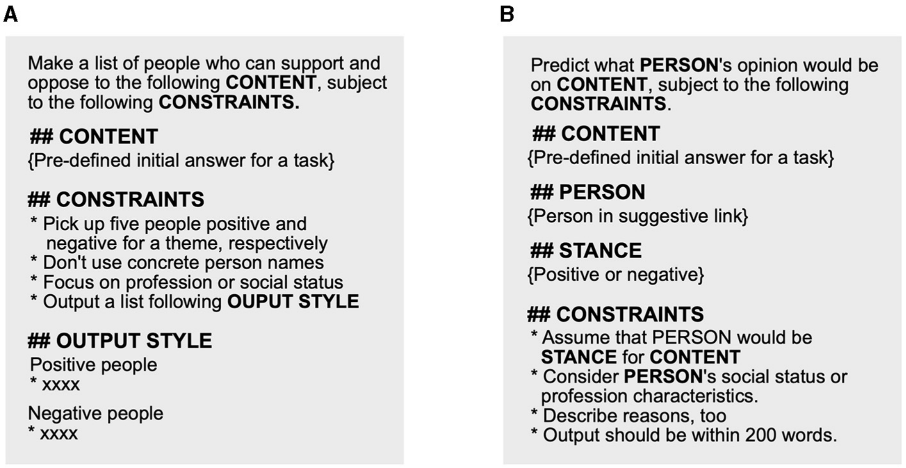
The following is the procedure of *initial answer* generation in the EXPOSITIVE bot:

1. Similar to the SUGGESTIVE bot, the EXPOSITIVE bot generates a plain *initial answer* and a list of positive/negative people for a given theme.
2. Similar to the SUGGESTIVE^{++} bot, the EXPOSITIVE bot generates opinions for randomly selected positive and negative persons.
3. Each opinion is summarized in a one-line sentence via the GPT API.
4. The EXPOSITIVE bot puts the summarized sentences for a positive and negative person at the end of the *initial answer*.

Note that we cached summarized one-line opinions before the study, similar to *initial answers*.

4.2.4 Plain bot

PLAIN bot is a control UI. Unlike the other three chatbots, this chatbot generates simple answers to participant queries via the GPT API. For *initial questions*, the bot displays the cached *initial answers*.



A

Make a list of people who can support and oppose to the following **CONTENT**, subject to the following **CONSTRAINTS.**

## CONTENT
{Pre-defined initial answer for a task}

## CONSTRAINTS
* Pick up five people positive and negative for a theme, respectively
* Don't use concrete person names
* Focus on profession or social status
* Output a list following **OUPUT STYLE**

## OUTPUT STYLE
Positive people
* xxxx

Negative people
* xxxx

B

Predict what **PERSON**'s opinion would be on **CONTENT**, subject to the following **CONSTRAINTS**.

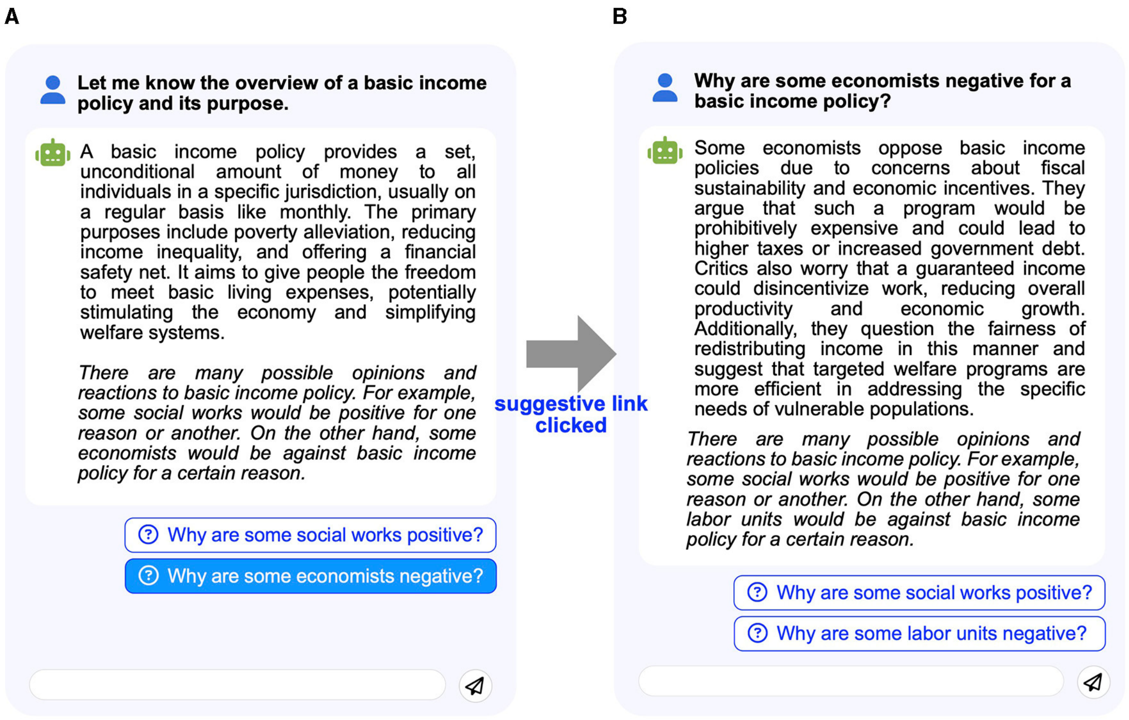
## CONTENT
{Pre-defined initial answer for a task}

## PERSON
{Person in suggestive link}

## STANCE
{Positive or negative}

## CONSTRAINTS
* Assume that PERSON would be **STANCE** for **CONTENT**
* Consider **PERSON**'s social status or profession characteristics.
* Describe reasons, too
* Output should be within 200 words.

FIGURE 3
LLM prompts used to generate information for SUGGESTIVE and SUGGESTIVE^{++} bots. **(A)** Prompt to gather positive/negative people. **(B)** Prompt to find target person's opinion.



Let me know the overview of a basic income policy and its purpose.

A basic income policy provides a set, unconditional amount of money to all individuals in a specific jurisdiction, usually on a regular basis like monthly. The primary purposes include poverty alleviation, reducing income inequality, and offering a financial safety net. It aims to give people the freedom to meet basic living expenses, potentially stimulating the economy and simplifying welfare systems.

There are many possible opinions and reactions to basic income policy. For example, some social works would be positive for one reason or another. On the other hand, some economists would be against basic income policy for a certain reason.

Why are some social works positive?
Why are some economists negative?

Why are some economists negative for a basic income policy?

Some economists oppose basic income policies due to concerns about fiscal sustainability and economic incentives. They argue that such a program would be prohibitively expensive and could lead to higher taxes or increased government debt. Critics also worry that a guaranteed income could disincentivize work, reducing overall productivity and economic growth. Additionally, they question the fairness of redistributing income in this manner and suggest that targeted welfare programs are more efficient in addressing the specific needs of vulnerable populations.

There are many possible opinions and reactions to basic income policy. For example, some social works would be positive for one reason or another. On the other hand, some labor units would be against basic income policy for a certain reason.

Why are some social works positive?
Why are some labor units negative?

FIGURE 4
SUGGESTIVE^{++} bot behavior when a participant clicks the suggestive link. **(A)** Response to initial question. **(B)** Response to suggestive click.

4.2.5 Common setting to all UI conditions

To ensure that the OpenAI API responses for a question were not truncated every time participants issued questions to the chatbot[8], we added a prompt-limited answer length of 400 Japanese characters (about 200 English words) for their questions. If the chatbot did not receive responses from the API within 10s, the chatbot displayed the following message: "The query has failed. Please reissue your question." In all UI conditions, we cached the API results to new queries for stable chatbot responses to participant queries. We configured the bots to display the generated answers within 7–10s when using the cached results.

[8] The prompt was set invisible to the participants.

4.3 Procedure

First, the participants were asked to read an overview of our user study and the treatment of their collected data on a crowdsourcing website. After agreeing to a consent form, the participants were transferred to our website to start their participation in the user study. Each participant was then randomly allocated to a UI environment and three decision-making tasks. To ensure that all participants view our system information with the same layout, only PC-based log-ins were allowed (i.e., no participant could access the study if one uses a tablet or a smart phone).

Then, the participants read a description of a task flow and the chatbot used in the study. Assuming that some participants were unfamiliar with chatbot systems for IR, we made the description of our chatbot system as comprehensible as possible. Moreover, we required participants to click a "read next" button every time they read a portion of the description to ensure that they read it completely.

Next, the participants were asked to conduct a practice task to familiarize themselves with their allocated chatbot. In the practice task, the participants were asked to summarize their opinions on introducing a universal basic income system in Japan.

Afterward, the participants performed the three main tasks for the three themes randomly allocated to them from the nine themes listed in Table 1. The main task order was randomized for each participant. In each main task, the participants performed the following three subtasks for each of the allocated themes: (1) pre-questionnaire, (2) decision-making, and (3) post-questionnaire tasks.

In the pre-questionnaire task, the participants ranked their interest and knowledge of each main task theme using a seven-point Likert scale (1, not at all; 7, very much).

Subsequently, the following scenario was presented to each participant ([THEME] is a task theme):

> *"Imagine the following case. The introduction of* [THEME] *has been discussed in your city. After the discussion in the city council, the city decided whether* [THEME] *is introduced or not, based on the interview with several residents. You are selected as an interviewee and need to explain whether you support* [THEME]. *Your answer will have a substantial influence on the city policy. So, you are about to collect information about* [THEME] *for your decision-making by using the latest chatbot system. Collect necessary information with the chatbot. When you reach a satisfactory conclusion, summarize your opinion with reasons and fill it in on an answer form."*

After reading the scenario, the participants were invited to start the decision-making task by clicking a dedicated button. The browser opened a webpage, where the participants interacted with the allocated chatbot and reported their opinions. At this stage, we set an initial question such as *"Let me know the overview of* [THEME] *and its purpose"* in the chatbot query box. Thus, all participants would ask the chatbot the initial question and collect information if necessary. The participants then reported their opinions when they reached their conclusion.

In the post-questionnaire task, we surveyed how many times our chatbot failed to generate answers during the decision-making task. These situations occurred because the chatbot occasionally failed to fetch the OpenAI GPT API within a limited time. For this survey, we asked the participants the following question: *How many times did you see the message "The query failed. Please enter your question again."* The participants reported the error frequency on a five-point Likert scale (1, never; 2, once; 3, 2–3 times; 4, 4–5 times; 5, more than five times).

At the end of the three main tasks, we administered an exit questionnaire to obtain feedback regarding the chatbot systems. The participants also answered the daily usage of chatbot systems for IR and demographic questions related to gender, age, and education.

4.4 Participants

We recruited 300 participants using Lancers.jp[9], a Japanese crowdsourcing service. Nevertheless, 18 participants were excluded from the analysis because 1 participant violated an instructional manipulation check (IMC) (Oppenheimer et al., 2009) in the exit questionnaire, 15 participants had more than one chatbot failure case in responding to their queries, and 2 participants completed the tasks without using our chatbot. Thus, only 282 participant responses were analyzed. All participants were Japanese (male = 191; female = 87; others= 4). Most participants were in their 30s and 40s (20s = 5.7%; 30s = 27.3%; 40s = 44.0%; 50s = 17.7%; others = 5.4%). Furthermore, about half of the participants reported that they never used chatbots for IR, such as ChatGPT, Google Gemini, and Bing Copilot[10] (never used = 45.7%; once every several months = 10.3%; once a month = 13.5%; once a week = 16.3%; once every several days = 7.8%; several times a day = 6.3%). Participants were randomly assigned to one of the four UI conditions (PLAIN = 66; EXPOSITIVE = 71; SUGGESTIVE = 78; SUGGESTIVE^{++} = 67). They used their PC or Mac to join our online user study. All participants who completed the tasks received 400 Japanese yen (approximately $2.75). On average, the participants finished all tasks within 48 min (*median*: 43 min).

4.5 Measurements

4.5.1 Task duration

We measured the *task duration*, corresponding to the time spent on a decision-making task per theme. Task duration is often used to examine how much effort users make in learning during the information-seeking process (Câmara et al., 2021). In our study, we defined the *task duration* as the time span from the moment when chatbot interfaces were displayed to the moment when the participants reported their opinions.

During the user study, participants engage in a critical learning activity, requiring them to not only look up unfamiliar topics but also analyze the task theme and summarize their opinions

[9] https://www.lancers.jp/

[10] Bing Copilot: https://www.microsoft.com/en-us/bing.

from various perspectives. This type of learning is often referred to as *critical learning* (Lee et al., 2015). Within the information retrieval community, researchers often use task/search duration as a measure of critical learning engagement and effort during information-seeking activities (Yamamoto and Yamamoto, 2018; Câmara et al., 2021; Roy et al., 2021). However, studies have shown that people interacting with chatbots, like ChatGPT, tend to spend less time on search tasks compared to conventional web search engines (Xu et al., 2023). Therefore, we consider that task/search duration could be a valuable metric to assess how effectively our chatbot strategy promotes critical learning during conversational searches.

4.5.2 Search frequency

We measured the *search frequency*, corresponding to the number of times the participants issued queries to the chatbots during their decision-making tasks. Similar to task/search duration, this metric is also often used to evaluate how willing people are to learn a topic in the fields of information retrieval and human-computer interaction. The *query issue count* can be regarded as how the participants came up with questions in their minds while interacting with the chatbots for their decision-making. We also measured the recommended queries (i.e., *suggestive links*) that the participants with SUGGESTIVE^{++} bot clicked as well as the queries that the participants filled in the chat box by themselves.

4.5.3 Opinion length

We examined how many tokens (terms) are contained in the participants' reported opinions. In the study, we asked the participants to report their opinions with reasons without setting minimum requirements for opinion length. We assumed that the more persuasive opinions the participants were encouraged to write, the longer their opinions would be. Therefore, we calculated the token-based length of participant opinions using *MeCab*, a Japanese morphological analyzer[11].

4.5.4 Perspective in opinion

We calculated the number of perspectives in the participants' opinions to investigate whether they summarized their opinions from various perspectives. This approach aligns with the concept of *T-Depth*, a metric introduced by Wilson and Wilson (2013), designed to evaluate the coverage of subtopics in participant opinions. *T-Depth* has been used in several studies to measure learning outcomes during information-seeking activities (Wilson and Wilson, 2013; Roy et al., 2021). Our indicator is a simplified version of *T-Fact*; it focuses only on the number of distinct perspectives rather than seeing how deeply participants mention each subtopic. This simplification stems from the challenge of objectively evaluating the depth of opinion on subtopics.

The themes listed in Table 1 are popular debating topics in Japan. Therefore, many books and webpages organize and list perspectives for discussion of themes. Our research group members collected and aggregated perspectives for each theme from the Web. Then, they used the list of aggregated perspectives to manually check which aspect appeared in each participant's opinion. It should be noted that the number of perspectives varied depending on the themes. Therefore, we rescaled the number of appearing perspectives in participant opinions by the expected maximum value (the number of collected perspectives per theme in Table 1).

As Sharma et al. (2024) have shown, conversational searches facilitated by LLMs often lead people to inquire about biased topics, resulting in more selective search behaviors. Therefore, we consider that the number of perspectives is a significant indicator to how effective our chatbot strategy is to promote more diverse information-seeking.

4.6 Statistical analyses

To examine the effect of suggestive endings in AI-powered chatbots, we analyzed the collected behavior logs and participant questionnaire responses using an analysis of covariance (ANCOVA). We conducted an ANCOVA to examine the main effect of *UI conditions* on the following measurements: (1) task duration, (2) questioning (search) frequency, (3) token length of opinions, and (4) the number of aspects in opinions. In the ANCOVA, we treated *familiarity*, *interest* of task themes, and *use frequency of chatbot for IR* as covariates to control personal factors. In *post hoc* tests, we conducted the Benjamini–Hochberg false discovery rate (FDR) correction (Benjamini and Hochberg, 2000) to make multiple comparisons between the UI conditions. In the ANCOVA and *post hoc* tests, we conducted log transformation for task duration, questioning frequency, and token length of opinions since the data did not follow Gaussian distributions.

5 Results

5.1 Task duration

We investigated the time and effort invested by participants for the decision-making tasks. Figure 5A illustrates the mean and standard error of the task duration. The ANCOVA result showed a significant impact of the UI conditions on the task duration per task, after controlling individual factors ($F_{(3,839)} = 6.28$, $p < 0.001$). Moreover, we observed a statistically significant difference between interest in themes (a covariate) on task duration ($F_{(1,839)} = 5.29$, $p0.05$).

The *post hoc* tests revealed that the participants using the SUGGESTIVE bot spent 114s longer in decision-making tasks compared to those using the PLAIN bot (*mean*: 680.3s vs. 566.6s; $p(suggestive - plain) < 0.001$). Moreover, *post hoc* tests showed that participants using the SUGGESTIVE^{++} and EXPOSITIVE bots completed the tasks significantly faster than those using the SUGGESTIVE bot (*mean*: 574.3s vs. 532.3s vs. 680.3s; $p(suggestive - suggestive^{++}) < 0.05$, $p(suggestive - expositive) < 0.01$). Nonetheless, no significant differences were observed between the SUGGESTIVE^{++} and PLAIN bots and between the EXPOSITIVE and PLAIN bots.

In summary, the SUGGESTIVE bot encouraged the participants to conduct a decision-making task more slowly than any other UIs. By contrast, the SUGGESTIVE^{++} and EXPOSITIVE bots did not seem to affect the participants' behavior. These findings suggest

[11] MeCab: https://taku910.github.io/mecab/ (in Japanese).

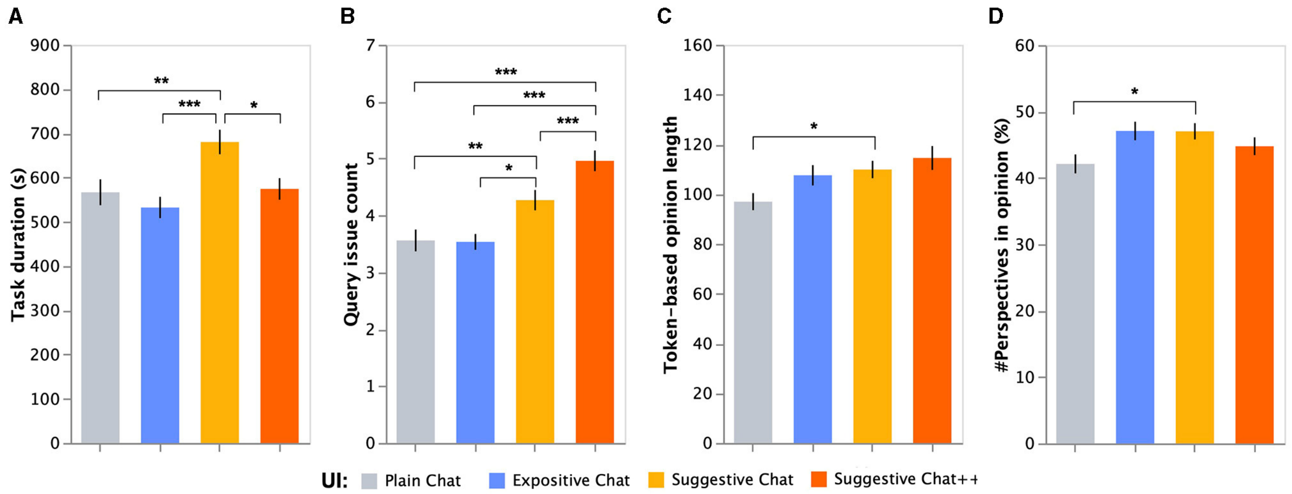


FIGURE 5
Mean and standard errors of **(A)** task duration, **(B)** search frequency, **(C)** opinion length, and **(D)** perspective count in opinion, depending on UI conditions. ***, significance level at 0.001; **, significance level at 0.01; *, significance level at 0.05, adjusted using the Benjamini–Hochberg FDR corre ction).

that the participants using the SUGGESTIVE bot appeared to invest more effort in collecting clues for their decision-making or organizing their opinions compared to those using the PLAIN and SUGGESTIVE^{++} bots.

5.2 Search frequency

We investigated how frequently the participants asked their chatbots to collect clues for the decision-making tasks. Figure 5B illustrates the mean and standard error of the search frequency. The ANCOVA result showed that the UI conditions had a significant impact on query issue count after controlling individual factors ($F_{(3,839)} = 17.7, p < 0.001$). No statistically significant differences were observed between interest, knowledge of themes, and daily usage of chatbots for IR on query issue count.

The *post hoc* tests revealed that the participants significantly asked more questions to the SUGGESTIVE bot compared to the PLAIN and EXPOSITIVE bots (*mean*: 4.27 vs. 3.56 vs. 3.54; $p(suggestive - plain) < 0.01$, $p(suggestive - expositive) < 0.05$). Furthermore, we observed that SUGGESTIVE^{++} bots encouraged participants to ask the bot more frequently compared with the SUGGESTIVE, EXPOSITIVE, and PLAIN bots (*mean*: 4.96 vs. 4.27 vs. 3.56 vs. 3.54; $p(suggestive^{++} - suggestive) < 0.001$, $p(suggestive^{++} - expositive) < 0.001$, $p(suggestive^{++} - plain) < 0.001$). It should be noted that the participants using the SUGGESTIVE^{++} bot queried with suggestive links at 3.55 times per task and queried without the links (querying by themselves) at 2.04 times per task on average. SUGGESTIVE^{++} bot enables people to ask the bot questions just using suggestive links, whereas people using SUGGESTIVE bot have to think about questions and type them in the bot by themselves. Therefore, These statistics show that participants using SUGGESTIVE^{++} bot were quite willing to use the suggestive links during the tasks. The *post hoc* test results revealed that the EXPOSITIVE bot promoted active searches compared to the PLAIN bot.

The above findings suggest that if the answer of the SUGGESTIVE bot ended with a suggestion regarding the existence of positive/negative opinions, participants were willing to ask questions to the bot more than what they would do with the PLAIN and EXPOSITIVE bots, which proactively and explicitly describe positive/negative opinions. This tendency could be stronger if the SUGGESTIVE^{++} bot displayed links to issue queries for viewing detailed information on positive/negative opinions.

5.3 Opinion length

The length of the participant reports submitted as task answers was considered as a metric to examine the decision-making level promoted by the four chatbot types. Figure 5C illustrates the mean and standard error of the token-based opinion length. The ANCOVA result showed that the UI conditions had a significant impact on token-based opinion length after controlling individual factors ($F_{(3,839)} = 2.80, p < 0.05$). We observed that two individual factors (covariates), i.e., interest in themes ($F_{(1,839)} = 9.21$, $p < 0.01$) and knowledge on themes ($F_{(1,839)} = 4.62$, $p < 0.05$), significantly affected the opinion length.

The *post hoc* tests revealed that the participants using the SUGGESTIVE bot wrote longer opinions compared to those using the PLAIN bot (*mean*: 109.9 tokens vs. 97.0 tokens; $p(suggestive - plain) < 0.05$). No significant difference was observed between the SUGGESTIVE^{++} and PLAIN bots ($p > 0.05$), although the mean opinion length of the SUGGESTIVE^{++} bot was higher than that of the SUGGESTIVE bot (*mean*: 114.5 vs. 109.9). Furthermore, no significant difference was observed between the EXPOSITVE and PLAIN bots.

These results indicate that if the participants found the existence of positive/negative people for themes using SUGGESTIVE bots, they were likely to explain their opinion with more words than those using the PLAIN bot, which just answered given questions straightforwardly. In addition, the results indicate that the EXPOSITVE bot did not have a large influence on opinion volume, despite providing richer answers to initial questions than the PLAIN bot.

5.4 Perspectives in opinion

We investigated how many possible perspectives appeared in the participants' submitted opinions to examine if they wrote their opinions from various perspectives. Figure 5D illustrates the mean and standard error of the perspective count. The ANCOVA result revealed that the UI conditions had a significant impact on the rescaled number of perspectives in opinion after controlling individual factors ($F_{(3,838)} = 2.82$, $p < 0.05$). No statistical significance was observed in individual factors (interest in themes, knowledge of themes, daily usage of chatbots for IR).

The *post hoc* tests showed that the participants using the SUGGESTIVE bot referred to significantly more perspectives in their opinions than those using the PLAIN bot (*mean*: 47.0% vs. 42.1% of possible perspectives; $p(\textit{suggestive} - \textit{plain}) < 0.05$). No significant difference was observed between the SUGGESTIVE^{++} and PLAIN bots ($p > 0.05$), although the participants using the SUGGESTIVE^{++} bot did more chatbot searches compared to those using the PLAIN and SUGGESTIVE bots. Moreover, no significant difference was observed between the EXPOSITIVE and PLAIN bots ($p > 0.05$), although the mean opinion length of the EXPOSITIVE bot was higher compared to that of the SUGGESTIVE bot (*mean*: 47.1% vs 47.0%).

These results indicate that the participants using the SUGGESTIVE bot were likely to summarize their opinions from various viewpoints compared to those using the PLAIN bot. Furthermore, the results indicate that regardless of the richer answers provided by the EXPOSITVE bot to initial questions compared to the PLAIN bot, the participants did not formulate their opinions from multiple perspectives.

5.5 Qualitative analysis

We analyzed the free-form responses in the exit questionnaire to explore the participants' strategies for their decision-making. In the exit questionnaire, the participants were asked to report how they organized and summarized their opinions during the decision-making tasks. Our research group members conducted an open coding (Lewins and Silver, 2014) for the participants' reports to explore the types of participant strategies.

5.5.1 Examination from various perspectives

Some participants stated that they made decisions based on various perspectives (e.g., advantages and disadvantages of a given theme). The following comments are from participants who reported that they considered various perspectives (translated from Japanese to English):

> *(P19 with suggestive bot) "I was careful not to favor one side over the other by making the chatbot present information on both pros and cons. I also verified my prior knowledge, comparing the chatbot responses with my own views."*

> *(P47 with expositive bot) "To write solid opinions, I collected information from two perspectives: pros/cons and positive/negative opinions."*

Meanwhile, the following comment is from a participant who was thought not to consider various perspectives:

> *(P11 with plain bot) "After deciding my stance, either for or against a given theme, I used the chatbot to collect information supporting my stance."*

We examined the ratio of participants who clearly commented that they considered various perspectives during the tasks depending on the UI conditions. The ratios were 52.1%, 34.9%, 62.8%, and 62.7% for the participants using the EXPOSITIVE bot, the PLAIN bot, the SUGGESTIVE bot, and the SUGGESTIVE^{++} bot, respectively. The χ^2 tests with the Bonferroni adjustment revealed that the ratios of the SUGGESTIVE and SUGGESTIVE^{++} bots were significantly higher than that of the PLAIN bot ($p(\textit{suggestive} - \textit{plain}) < 0.05/6$; $p(\textit{suggestive}^{++} - \textit{plain}) < 0.05/6$). These results indicate that if the chatbots implicitly suggested the existence of positive/negative opinions, the participants could be more careful about various perspectives in their decision-making. By contrast, even if the EXPOSITVE bot complemented a brief summary about a positive and a negative person's opinion to initial answers, the participants did not try to make their decisions from multiple perspectives.

5.5.2 How to use chatbots

Different participants used the chatbots for different reasons. Some participants used the chatbots to learn about unknown concepts from the chatbot's answers, as represented by the following comments:

> *(P26 with expositive bot) "I read the chatbot's answer. Then I queried the chatbot to summarize my answer if I came up with questions."*

> *(P229 with plain bot) "I asked the bot about what I was curious about or did not understand and then summarized my opinion."*

Other participants used the bots to collect clues for their decision-making. Some participants thoughtfully considered various perspectives or weighed the pros and cons of the given themes to inform their decisions as follows:

> *(P59 with suggestive bot) "I compared opinions from both supporters and opponents. Then I organized those opinions closer to my own thinking."*

> *(P170 with suggestive^{++} bot) "I made sure to check both positive and negative opinions before forming my own view. I queried the chatbot about positive opinions and negative opinions by turns."*

Some participants also tried to corroborate their opinions (prior beliefs) with the chatbots to gather supportive data and expected counterarguments such as the following:

> *(P102 with suggestive^{++} bot) "I started by reviewing the provided theme overview and determined my stance. I then searched for supportive reasons and opposing ones and selected persuasive arguments to consolidate my own opinion. If I didn't find a decisive reason in the first search, I conducted a further, more in-depth survey using the bot."*

> *(P148 with suggestive bot) "Firstly, I received an overview of the theme and then inquired about the details of opposing opinions. After that, I formulated my arguments, constructing a rebuttal."*

As already described, the participants' comments in the exit questionnaire indicate that the SUGGESTIVE and SUGGESTIVE^{++} bots promoted the participants' awareness of decision-making from both positive and negative perspectives on the given themes. The following comments indicate that participants thought suggestive links provided by the SUGGESTIVE^{++} bot are useful in searching for detailed information regarding positive/negative perspectives efficiently:

> *(P6 with suggestive^{++} bot) "Once I understood the overview, the task was not so difficult. I didn't come up with new questions by myself, so I relied more on suggestive links rather than manually asking the bot questions."*

> *(P281 with suggestive^{++} bot) "The system allowed me to choose between pro and con opinions (links) on the theme. I used this feature to validate my own views."*

However, the way of using suggestive links depended on the participants. Some participants using the SUGGESTIVE^{++} bot implied that they used either links to positive opinions or links to negative opinions as follows:

> *(P87 with suggestive^{++} bot) "I used the system to investigate what risks might be involved. If the risks (negative opinions) on a theme were low, I tried to have positive opinions on it."*

> *(P241 with suggestive^{++} bot) "Firstly, I determined whether I was in favor or against the given theme, and then I used the chatbot to search for data supporting my view."*

The SUGGESTIVE bot did not provide suggestive links compared with the SUGGESTIVE^{++} bot; however, it prompted the participants to ask about or reflect on positive/negative people's opinions on the theme:

> *(P203 with suggestive bot) "In answer to the initial question, the bot showed a description suggesting that I should seek further clarification on positive and negative opinions, so I started by following the suggestion."*

> *(P280 with suggestive bot) "Firstly, I was curious about what the pros and cons might be, so I searched for those aspects. While considering the pros and cons of the theme, I checked current statistics or data to ensure that I tried to form a fair opinion."*

As the below comment suggests, some participants using the EXPOSITIVE bot would not be willing to ask questions as they would feel that the bot provided sufficient information for their decision-making:

> *(P277 with expositive bot) "Most information from the bot was usable as-is, so I actively used them."*

5.5.3 Complaints

A few participants complained that the chatbot's answers sometimes seemed wrong or unreliable, thereby hoping that the bots could provide more detailed information and evidence. Moreover, a few participants complained that the chatbot's information was difficult to read.

6 Discussion

After controlling the individual factors, our study results revealed that the SUGGESTIVE and SUGGESTIVE^{++} bots significantly influence the participants' behavior and attitude in their decision-making tasks.

As for **RQ1**, the SUGGESTIVE bot caused the participants to spend the longest time in decision-making tasks among the four UI conditions. Moreover, the SUGGESTIVE bot promoted more frequent information seeking compared to the PLAIN and EXPOSITIVE bots. It also encouraged the participants to write longer texts regarding their opinions compared to those using the PLAIN bot. Therefore, we conclude that the SUGGESTIVE bot can encourage users to put more effort into formulating their opinions and gathering information for decision-making from time and content perspectives.

As for **RQ2**, our qualitative analysis revealed that more participants using the SUGGESTIVE and SUGGESTIVE^{++} bots were aware of both the pros/cons perspectives in their decision-making compared to those using the PLAIN bot. Furthermore, our behavior analysis showed that the participants using the SUGGESTIVE bot were likely to refer to more perspectives in their opinion reports compared to those using the PLAIN bot, whereas the SUGGESTIVE^{++} did not indicate such a tendency. We conclude that the SUGGESTIVE bot can encourage users to formulate their decision from various viewpoints.

As for **RQ3**, the SUGGESTIVE^{++} bot, providing links to survey positive/negative people's opinions along with the suggestive answers, promoted more frequent search activities compared to any other UI. In addition, the SUGGESTIVE^{++} bot significantly reduced the time cost for the tasks compared to the SUGGESTIVE bot. In the exit questionnaire, 62.7% of participants using the SUGGESTIVE^{++} bot reported that they tried to formulate their opinions as objectively as possible from both sides of pros and cons. However, the behavior analysis result showed that the SUGGESTIVE^{++} bot did not encourage participants to report long opinions with various perspectives compared to the SUGGESTIVE bot. These results indicate that the SUGGESTIVE^{++} bot did not substantially promote critical decision-making activities, although it could improve information-seeking efficiency. We believe that such noncritical behaviors can be attributed to the cognitive bias in information seeking (White, 2013; Azzopardi, 2021), such as *selective exposure* (Liao et al., 2015) and *confirmation bias* (Kahneman, 2011; Suzuki and Yamamoto, 2021). The comments of P87 and P241 in the qualitative analysis suggest the influence of selective exposure and confirmation bias on the users' behaviors. However, it is worth noting that our interpretations above are based only on the submitted task reports and the participants' reflective comments in the exit questionnaire. To ensure whether the SUGGESTIVE^{++} bot can promote critical decision-making, a further study of the cognitive process during decision-making tasks with the chatbots should be conducted via laboratory experiments.

As for the EXPOSITIVE bot, we found no significant effects compared to the PLAIN bot. When querying a theme overview at the beginning of the tasks, the participants using the EXPOSITIVE bot saw a brief summary of a positive/negative person's opinions without additional queries. In other words, the bot explicitly complemented short, two-sided information for task themes, although the complemented information is not sufficient to make critical judgments on the task themes. However, as P277 suggested, the EXPOSITIVE bot creates the impression of providing sufficient information. This drives participants to pick up only their favorable information to summarize their opinions. Therefore, even if the participants used the EXPOSITVE bot, they would not exert much effort toward critical decision-making.

As for **RQ4**, we confirmed that the knowledge of themes affected time efforts in decision-making tasks, while the interest in themes affected the length of reported opinions. These results indicate that knowledge of and interest in themes could affect the amount of effort in decision-making with AI-powered chatbots.

In the end, we conclude that *suggestive endings*, which hint at something in chatbot interaction, can draw more spontaneous questions from users and encourage them to formulate their opinions from various perspectives rather than provide definitive answers or predefined questions (such as in the SUGGESTIVE++ bot).

7 Limitations and potential challenges

Our study showed that the suggestive ending strategy in a human–chatbot interaction can be useful in enhancing critical decision-making. However, the study has some limitations and several challenges still exist toward better AI-based decision-making support.

One limitation is an experimental environment. We used a crowdsourcing platform for our user study. Although user studies with crowdsourcing have been more popular, this approach has several concerns, including the demographic biases of crowd workers, the presence of lazy participants, and the control of experimental environments (task times and devices for experiments) (Ross et al., 2010). As a result, our study's participant pool might not accurately reflect the general population, and some participants might not have performed the tasks seriously.

Another limitation the display timing of suggestive endings. In the study, the SUGGESTIVE bot provided answers with suggestive endings only for the initial questions. Therefore, we need to investigate the effects depending on the timing and context of suggestive ending presentations. Moreover, we relied only on the analysis of participants' behaviors during the main tasks and their comments in the exit questionnaire to understand their strategy for decision-making. Think-aloud protocols and stimulated recalls should be conducted in laboratory experiment settings to understand the cognitive decision-making process with chatbots better.

A possible challenge is the topic on which chatbot hints. In the study, we focused on suggesting who is positive or negative for a theme, aiming to make participants aware of the pros/cons viewpoints and to draw spontaneous questions to foster their understanding of the theme (e.g., *"[Occupation name] people can be positive for [THEME] with a certain reason"*). However, other factors can affect critical decision-making and information seeking. For example, researchers in information and media literacy have stated that currency, relevance, authority, accuracy, and purpose are important to check for critical judgment on the quality of claims and information (Musgrove et al., 2018). Therefore, the chatbots should determine a focused factor and create associated suggestive endings depending on the context of decision-making. For example, if users are encouraged to explore various information from the currency viewpoints, a possible suggestive ending can be *"The above opinion was mainstream in the 2010s, but completely different opinions are prevalent in the 2020s"*. A remaining issue is a method to automatically generate effective suggestive endings.

The second challenge is related to the proper use of chat strategies to enhance cognitive activities. In this study, we focused on hinting at something in chatbot answers to draw spontaneous questions from users. However, there can be other ways to draw cognitive efforts toward critical decision-making, such as AI questioning and forcible time setting for thinking (Buçinca et al., 2021). As for the AI-questioning approach, devising what and how to make chatbots ask would enable them to promote various cognitive activities, such as logical reasoning (Danry et al., 2023) and reflecting on lacking issues of one's view (Okuse and Yamamoto, 2023). Nevertheless, explicit questioning might make users intrusive and uncomfortable depending on the frequency, timing, or user personality. Furthermore, even if chatbots provide questions and suggestive endings for users, some users may have difficulties in finding answers and related information by themselves (Odijk et al., 2015). Therefore, the chatbots for decision-making support should use explicit questioning (instructive intervention), suggestive endings in answers (modest intervention), and detailed explanations, depending on the situation and users' personal factors. Moreover, the chatbots should encourage users to perform Web searches without an overreliance on the bots as necessary so that users can corroborate their opinions and the bot's answers from various sources.

8 Conclusion

Although people use generative AIs with LLMs to readily obtain information relevant to their requirements, their overreliance on AIs can cause shortsighted decision-making and weaken cognitive skills. Our proposed SUGGESTIVE chatbot encourages people to have spontaneous questions for critical decision-making on a given theme by ending an answer that hints at potentially interest-triggering points.

The online user study revealed that the SUGGESTIVE bot encouraged participants to exert more effort in developing their opinions and gathering information for decision-making compared with simple chatbots. Moreover, the study showed that the SUGGESTIVE bot encouraged participants to make their decisions from various perspectives. We did not observe such a tendency with the EXPOSITVE chatbot, which complemented information from a specific perspective. These findings indicate that AI-powered chatbots can better enhance human decision-making with *suggestive endings*, which leave room for questions and discussions rather than definitive explanations to a question (query).

Our proposed method has several challenges for improvement. These include investigation on how to use suggestive endings, questioning, and definitive explanations depending on situations and laboratory studies to understand the cognitive processes during decision-making tasks using our chatbot strategy. However, we believe that *suggestive endings* in chatbot answers constitute a good strategy for AI-powered chatbots to enhance critical information seeking and decision-making.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study. Written informed consent was obtained from the individual(s) for the publication of any potentially identifiable images or data included in this article.

Author contributions

YY: Conceptualization, Formal Analysis, Methodology, Software, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This work was supported in part by the Grants-in-Aid for Scientific Research (21H03554, 21H03775, and 22H03905) from the MEXT of Japan and Daiko Foundation.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Arnold, K. C., Chauncey, K., and Gajos, K. Z. (2020). "Predictive text encourages predictable writing, in *Proceedings of the 25th International Conference on Intelligent User Interfaces, IUI '20* (New York, NY, USA: Association for Computing Machinery), 128–138. doi: 10.1145/3377325.3377523

Arnold, K. C., Gajos, K. Z., and Kalai, A. T. (2016). "On suggesting phrases vs. predicting words for mobile text composition, in *Proceedings of the 29th Annual Symposium on User Interface Software and Technology, UIST '16* (New York, NY, USA: Association for Computing Machinery), 603–608. doi: 10.1145/2984511.2984584

Azzopardi, L. (2021). "Cognitive biases in search: a review and reflection of cognitive biases in information retrieval, in *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval, CHIIR '21* (New York, NY, USA: Association for Computing Machinery), 27–37. doi: 10.1145/3406522.3446023

Benjamini, Y., and Hochberg, Y. (2000). On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J. Educ. Behav. Stat.* 25, 60–83. doi: 10.2307/1165312

Brodsky, J. E., Brooks, P. J., Scimeca, D., Todorova, R., Galati, P., Batson, M., et al. (2021). Improving college students' fact-checking strategies through lateral reading instruction in a general education civics course. *Cogn. Res.* 6, 1–18. doi: 10.1186/s41235-021-00291-4

Buçinca, Z., Malaya, M. B., and Gajos, K. Z. (2021). To trust or to think: Cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proc. ACM Hum. Comput. Inter.* 5, 1–21. doi: 10.1145/3449287

Câmara, A., Roy, N., Maxwell, D., and Hauff, C. (2021). "Searching to learn with instructional scaffolding, in *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval, CHIIR '21* (New York, NY, USA: Association for Computing Machinery), 209–218. doi: 10.1145/3406522.3446012

Caraban, A., Karapanos, E., Gonçalves, D., and Campos, P. (2019). "23 ways to nudge: a review of technology-mediated nudging in human-computer interaction, in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19* (New York, NY, USA: Association for Computing Machinery), 1–15. doi: 10.1145/3290605.3300733

Carr, N. (2014). *The Glass Cage: How Our Computers are Changing Us*. New York: WW Norton and Company.

Danry, V., Pataranutaporn, P., Mao, Y., and Maes, P. (2023). "Don't just tell me, ask me: Ai systems that intelligently frame explanations as questions improve human logical discernment accuracy over causal ai explanations, in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23* (New York, NY, USA: Association for Computing Machinery), 1–13. doi: 10.1145/3544548.3580672

Ennis, R. H. (1987). "A taxonomy of critical thinking dispositions and abilities, in *Series of books in psychology. Teaching thinking skills: Theory and practice*, eds. J. B. Baron, and R. J. Sternberg (New York: W H Freeman/Times Books/ Henry Holt and Co.), 9–26.

Fisher, A., Rudin, C., and Dominici, F. (2019). All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *J. Mach. Learn. Res.* 20, 1–81.

Goddard, K., Roudsari, A., and Wyatt, J. C. (2011). Automation bias: a systematic review of frequency, effect mediators, and mitigators. *J. Am. Med. Inform. Assoc.* 19, 121–127. doi: 10.1136/amiajnl-2011-000089

Harvey, M., Hauff, C., and Elsweiler, D. (2015). "Learning by example: training users with high-quality query suggestions, in *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '15* (New York, NY, USA: Association for Computing Machinery), 133–142. doi: 10.1145/2766462.2767731

Hertwig, R., and Grne-Yanoff, T. (2017). Nudging and boosting: Steering or empowering good decisions. *Perspect. Psychol. Sci.* 12, 973–986. doi: 10.1177/1745691617702496

Hettiachchi, D., Kostakos, V., and Goncalves, J. (2022). A survey on task assignment in crowdsourcing. *ACM Comput. Surv.* 55, 1–35. doi: 10.1145/3494522

Ihoriya, H., Suzuki, M., and Yamamoto, Y. (2022). "Mitigating position bias in review search results with aspect indicator for loss aversion, in *Proceedings of the 2022 International Conference on Human-Computer Interaction, HCII '22* (Berlin, Heidelberg: Springer-Verlag), 17–32. doi: 10.1007/978-3-031-06509-5_2

Jakesch, M., Bhat, A., Buschek, D., Zalmanson, L., and 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, CHI '23* (New York, NY, USA: Association for Computing Machinery), 1–11. doi: 10.1145/3544548.3581196

Kahneman, D. (2011). *Thinking, Fast and Slow*. New York: NY: Macmillan.

Kasneci, E., Sessler, K., Kchemann, S., Bannert, M., Dementieva, D., Fischer, F., et al. (2023). Chatgpt for good? On opportunities and challenges of large language models for education. *Learning Indiv. Differ.* 103:102274. doi: 10.1016/j.lindif.2023.102274

Kim, B., Khanna, R., and Koyejo, O. (2016). "Examples are not enough, learn to criticize! criticism for interpretability, in *Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS '16* (Red Hook, NY, USA: Curran Associates Inc.), 2288–2296.

King, A. (1992). Facilitating elaborative learning through guided student-generated questioning. *Educ. Psychol.* 27, 111–126. doi: 10.1207/s15326985ep2701_8

Kittur, A., Chi, E. H., and Suh, B. (2008). "Crowdsourcing user studies with mechanical turk, in *Proceedings of the 2008 ACM Conference on Human Factors in Computing Systems, CHI '08* (New York, NY, USA: Association for Computing Machinery), 453–456. doi: 10.1145/1357054.1357127

Komarov, S., Reinecke, K., and Gajos, K. Z. (2013). "Crowdsourcing performance evaluations of user interfaces, in *Proceedings of the 2013 ACM Conference on Human Factors in Computing Systems, CHI '13* (New York, NY, USA: Association for Computing Machinery), 207–216. doi: 10.1145/2470654.2470684

Lakkaraju, H., and Bastani, O. (2020). "How do i fool you?": Manipulating user trust via misleading black box explanations, in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, AIES '20* (New York, NY, USA: Association for Computing Machinery), 79–85. doi: 10.1145/3375627.3375833

Lee, H.-J., Lee, J., Makara, K. A., Fishman, B. J., and Hong, Y.-I. (2015). Does higher education foster critical and creative learners? An exploration of two universities in South Korea and the USA. *Higher Educ. Res. Dev.* 34, 131–146. doi: 10.1080/07294360.2014.892477

Lewins, A., and Silver, C. (2014). *Using Software in Qualitative Research: A Step-By-Step Guide*. London: SAGE publications Ltd.

Liao, Q. V., and Fu, W.-T. (2014). "Can you hear me now? mitigating the echo chamber effect by source position indicators, in *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW '14* (New York, NY, USA: Association for Computing Machinery), 184–196. doi: 10.1145/2531602.2531711

Liao, Q. V., Fu, W.-T., and Mamidi, S. S. (2015). "It is all about perspective: An exploration of mitigating selective exposure with aspect indicators, in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, CHI '15* (New York, NY, USA: Association for Computing Machinery), 1439–1448. doi: 10.1145/2702123.2702570

Liu, M. X., Kittur, A., and Myers, B. A. (2022). "Crystalline: Lowering the cost for developers to collect and organize information for decision making, in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, CHI '22* (New York, NY, USA: Association for Computing Machinery), 1–16. doi: 10.1145/3491102.3501968

Lundberg, S. M., and Lee, S.-I. (2017). "A unified approach to interpreting model predictions, in *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS '17* (Red Hook, NY, USA: Curran Associates Inc), 4768–4777.

Lutz, J. (2016). The validity of crowdsourcing data in studying anger and aggressive behavior. *Soc. Psychol.* 47, 38–51. doi: 10.1027/1864-9335/a000256

Maynez, J., Narayan, S., Bohnet, B., and McDonald, R. (2020). "On faithfulness and factuality in abstractive summarization, in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL '20* (Association for Computational Linguistics), 1906–1919. doi: 10.18653/v1/2020.acl-main.173

Meola, M. (2004). Chucking the checklist: a contextual approach to teaching undergraduates web-site evaluation. *Portal* 4, 331–344. doi: 10.1353/pla.2004.0055

Metzler, D., Tay, Y., Bahri, D., and Najork, M. (2021). Rethinking search: Making domain experts out of dilettantes. *SIGIR Forum* 55, 1–27. doi: 10.1145/3476415.3476428

Musgrove, A. T., Powers, J. R., Rebar, L. C., and Musgrove, G. J. (2018). Real or fake? Resources for teaching college students how to identify fake news. *College Undergr. Libr.* 25, 243–260. doi: 10.1080/10691316.2018.1480444

Najork, M. (2023). "Generative information retrieval, in *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '23* (New York, NY, USA: Association for Computing Machinery). doi: 10.1145/3539618.3591871

Nakano, R., Hilton, J., Balaji, S., Wu, J., Ouyang, L., Kim, C., et al. (2022). Webgpt: Browser-assisted question-answering with human feedback. arXiv:2112.09332

Noyes, J. (2007). "Automation and decision making, in *Decision Making in Complex Environments* (New York: CRC Press), 73–82. doi: 10.1201/9781315576138-7

Odijk, D., White, R. W., Hassan Awadallah, A., and Dumais, S. T. (2015). "Struggling and success in web search, in *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, CIKM '15* (New York, NY, USA: Association for Computing Machinery), 1551–1560. doi: 10.1145/2806416.2806488

Okuse, Y., and Yamamoto, Y. (2023). "Chatbot to facilitate opinion formation in web search, in *Proceedings of the 25th HCI International Conference, HCII '23* (Berlin, Heidelberg: Springer-Verlag), 568–582. doi: 10.1007/978-3-031-35132-7_43

Oppenheimer, D. M., Meyvis, T., and Davidenko, N. (2009). Instructional manipulation checks: detecting satisficing to increase statistical power. *J. Exper. Soc. Psychol.* 45, 867–872. doi: 10.1016/j.jesp.2009.03.009

Peer, E., Brandimarte, L., Samat, S., and Acquisti, A. (2017). Beyond the turk: Alternative platforms for crowdsourcing behavioral research. *J. Exper. Soc. Psychol.* 70, 153–163. doi: 10.1016/j.jesp.2017.01.006

Petridis, S., Diakopoulos, N., Crowston, K., Hansen, M., Henderson, K., Jastrzebski, S., et al. (2023). "Anglekindling: supporting journalistic angle ideation with large language models, in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23* (New York, NY, USA: Association for Computing Machinery), 1–16. doi: 10.1145/3544548.3580907

Petty, R. E., and Cacioppo, J. T. (1986). The elaboration likelihood model of persuasion. *Adv. Exper. Soc. Psychol.* 19, 123–205. doi: 10.1016/S0065-2601(08)60214-2

Rinott, R., Dankin, L., Alzate Perez, C., Khapra, M. M., Aharoni, E., and Slonim, N. (2015). "Show me your evidence - an automatic method for context dependent evidence detection, in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing* (Lisbon, Portugal: Association for Computational Linguistics), 440–450. doi: 10.18653/v1/D15-1050

Ross, J., Irani, L., Silberman, M. S., Zaldivar, A., and Tomlinson, B. (2010). "Who are the crowdworkers? shifting demographics in mechanical turk, in *Proceedings of the 2010 ACM Conference on Human Factors in Computing Systems, CHI EA '10* (New York, NY, USA: Association for Computing Machinery), 2863–2872. doi: 10.1145/1753846.1753873

Roy, N., Torre, M. V., Gadiraju, U., Maxwell, D., and Hauff, C. (2021). "Note the highlight: Incorporating active reading tools in a search as learning environment, in *Proceedings of the 2021 ACM Conference on Human Information Interaction and Retrieval, CHIIR '21* (New York, NY, USA: Association for Computing Machinery), 229–238. doi: 10.1145/3406522.3446025

Saito, F., Shoji, Y., and Yamamoto, Y. (2020). "Highlighting weasel sentences for promoting critical information seeking on the web, in *Proceedings of the 21st International Conference on Web Information Systems Engineering, WISE '20* (Berlin, Heidelberg: Springer-Verlag), 424–440. doi: 10.1007/978-3-030-34223-4_27

Sharma, N., Liao, Q. V., and Xiao, Z. (2024). Generative echo chamber? Effects of llm-powered search systems on diverse information seeking. arXiv:2402.05880

Shimizu, Y., Ohki, T., and Yamamoto, Y. (2022). "Privacy-aware snippets: Enhancing assessment of balance between privacy risks and benefits in web search, in *Proceedings of the 2022 ACM Conference on Information Technology for Social Good, GoodIT '22* (New York, NY, USA: Association for Computing Machinery), 24–31. doi: 10.1145/3524458.3547231

Sun, J., Liao, Q. V., Muller, M., Agarwal, M., Houde, S., Talamadupula, K., et al. (2022). "Investigating explainability of generative ai for code through scenario-based design, in *27th International Conference on Intelligent User Interfaces, IUI '22* (New York, NY, USA: Association for Computing Machinery), 212–228. doi: 10.1145/3490099.3511119

Suzuki, M., and Yamamoto, Y. (2021). Characterizing the influence of confirmation bias on web search behavior. *Front. Psychol.* 12, 1–11. doi: 10.3389/fpsyg.2021.771948

Suzuki, M., and Yamamoto, Y. (2022). "Don't judge by looks: Search user interface to make searchers reflect on their relevance criteria and promote content-quality-oriented web searches, in *Proceedings of the 2022 ACM Conference on Information Technology for Social Good, GoodIT '22* (New York, NY, USA: Association for Computing Machinery), 1–8. doi: 10.1145/3524458.3547222

Tay, Y., Tran, V. Q., Dehghani, M., Ni, J., Bahri, D., Mehta, H., et al. (2022). Transformer memory as a differentiable search index, in *Advances in Neural Information Processing Systems, NeurIPS'22*, 21831–21843.

Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., et al. (2023). Gemini: A family of highly capable multimodal models. arXiv:2312.11805

Thaler, R. H., and Sunstein, C. R. (2009). *Nudge: Improving decisions about health, wealth, and happiness*. London: Penguin.

Wambsganss, T., Kueng, T., Soellner, M., and Leimeister, J. M. (2021). "Arguetutor: an adaptive dialog-based learning system for argumentation skills, in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI '21* (New York, NY, USA: Association for Computing Machinery), 1–13. doi: 10.1145/3411764.3445781

Wambsganss, T., Niklaus, C., Cetto, M., Söllner, M., Handschuh, S., and Leimeister, J. M. (2020). "Al: an adaptive learning support system for argumentation skills, in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI '20* (New York, NY, USA: Association for Computing Machinery), 1–14. doi: 10.1145/3313831.3376732

Wang, T., Yu, P., Tan, X. E., O'Brien, S., Pasunuru, R., Dwivedi-Yu, J., et al. (2023). Shepherd: A critic for language model generation. arXiv:2308.04592

Wang, Y., Leon, P. G., Scott, K., Chen, X., Acquisti, A., and Cranor, L. F. (2013). "Privacy nudges for social media: an exploratory facebook study, in *Proceedings of the 22nd International Conference on World Wide Web, WWW '13 Companion* (New York, NY, USA: Association for Computing Machinery), 763–770. doi: 10.1145/2487788.2488038

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., et al. (2021). Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

White, R. (2013). "Beliefs and biases in web search, in *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '13* (New York, NY, USA: Association for Computing Machinery), 3–12. doi: 10.1145/2484028.2484053

Wilson, M. J., and Wilson, M. L. (2013). A comparison of techniques for measuring sensemaking and learning within participant-generated summaries. *J. Am. Soc. Inform. Sci. Technol.* 64, 291–306. doi: 10.1002/asi.22758

Wineburg, S., and McGrew, S. (2019). Lateral reading and the nature of expertise: Reading less and learning more when evaluating digital information. *Teach. College Rec.* 121, 1–40. doi: 10.1177/016146811912101102

Wirz, D. S., Ort, A., Rasch, B., and Fahr, A. (2023). The role of cliffhangers in serial entertainment: An experiment on cliffhangers' effects on enjoyment, arousal, and intention to continue watching. *Psychol. Popular Media* 12, 186–196. doi: 10.1037/ppm0000392

Xu, R., Feng, Y., and Chen, H. (2023). Chatgpt vs. google: A comparative study of search performance and user experience. arXiv:2307.01135 doi: 10.2139/ssrn.4498671

Yamamoto, Y., and Yamamoto, T. (2018). "Query priming for promoting critical thinking in web search, in *Proceedings of the 2018 Conference on Human Information Interaction and Retrieval, CHIIR '18* (New York, NY, USA: Association for Computing Machinery), 12–21. doi: 10.1145/3176349.3176377

Yamamoto, Y., and Yamamoto, T. (2020). "Personalization finder: A search interface for identifying and self-controlling web search personalization, in *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020, JCDL '20* (New York, NY, USA: Association for Computing Machinery), 37–46. doi: 10.1145/3383583.3398519

Zimmerman, S., Thorpe, A., Fox, C., and Kruschwitz, U. (2019). "Privacy nudging in search: Investigating potential impacts, in *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval, CHIIR '19* (New York, NY, USA: Association for Computing Machinery), 283–287. doi: 10.1145/3295750.3298952

Zylowski, T., and Wölfel, M. (2023). "An NLP analysis of ChatGPT's personality simulation capabilities and implications for human-centric explainable AI interfaces, in *Proceedings of the 6th International Conference on Natural Language and Speech Processing, ICNLSP '23* (Association for Computational Linguistics), 168–177.