

See Widely, Think Wisely: Toward Designing a Generative Multi-agent System to Burst Filter Bubbles

Yu Zhang*

zhangyu29@lenovo.com
Southeast University
Nanjing, China
Lenovo Research
Beijing, China

Cen Yao

yaocen1@lenovo.com
Lenovo Research
Beijing, China

Qianying Wang
wangqya@lenovo.com
Lenovo Research
Beijing, China

Jingwei Sun*

sunjw12@lenovo.com
Lenovo Research
Beijing, China

Mingming Fan

mingmingfan@ust.hk
The Hong Kong University of Science
and Technology (Guangzhou)
Guangzhou, China
The Hong Kong University of Science
and Technology
Hong Kong SAR, China

Xin Geng†

xeng@seu.edu.cn
Southeast University
Nanjing, China

Li Feng

lfeng256@connect.hkust-gz.edu.cn
The Hong Kong University of Science
and Technology (Guangzhou)
Guangzhou, China

Liuxin Zhang

zhanglx2@lenovo.com
Lenovo Research
Beijing, China

Yong Rui†

yongrui@lenovo.com
Southeast University
Nanjing, China
Lenovo Research
Beijing, China

ABSTRACT

The proliferation of AI-powered search and recommendation systems has accelerated the formation of “filter bubbles” that reinforce people’s biases and narrow their perspectives. Previous research has attempted to address this issue by increasing the diversity of information exposure, which is often hindered by a lack of user motivation to engage with. In this study, we took a human-centered approach to explore how Large Language Models (LLMs) could assist users in embracing more diverse perspectives. We developed a prototype featuring LLM-powered multi-agent characters that users could interact with while reading social media content. We conducted a participatory design study with 18 participants and found that multi-agent dialogues with gamification incentives could motivate users to engage with opposing viewpoints. Additionally, progressive interactions with assessment tasks could promote thoughtful consideration. Based on these findings, we provided design implications with future work outlooks for leveraging LLMs to help users burst their filter bubbles.

*Both authors contributed equally to this research.

†Corresponding authors.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

CHI '24, May 11–16, 2024, Honolulu, HI, USA

© 2024 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-0330-0/24/05.

<https://doi.org/10.1145/3613904.3642545>

CCS CONCEPTS

- Human-centered computing → Interaction design; Human computer interaction (HCI);
- Information systems → Users and interactive retrieval.

KEYWORDS

filter bubble, multi-agent system, large language model, interaction design, diverse information

ACM Reference Format:

Yu Zhang, Jingwei Sun, Li Feng, Cen Yao, Mingming Fan, Liuxin Zhang, Qianying Wang, Xin Geng, and Yong Rui. 2024. See Widely, Think Wisely: Toward Designing a Generative Multi-agent System to Burst Filter Bubbles. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*, May 11–16, 2024, Honolulu, HI, USA. ACM, New York, NY, USA, 24 pages. <https://doi.org/10.1145/3613904.3642545>

1 INTRODUCTION

The boom of the PC and mobile internet has revolutionized the way information is consumed and disseminated. AI-powered search and recommendation systems are now a common feature on social media, news, and streaming platforms, analyzing users’ behavior, preferences, and interactions to provide personalized content. While these systems have greatly improved users’ experience, they can also exacerbate a phenomenon known as the “filter bubble” or “information cocoon”, where individuals tend to consume more information that confirms their existing beliefs, potentially narrowing their perspectives and reinforcing their potential biases.



Figure 1: The interfaces and the interaction flow of our prototype: users can interact with LLM-powered multi-agent characters while reading social media content.

To address that issue, researchers have proposed and studied three primary approaches, including optimizing recommendation algorithms, expanding users' exposure to diverse information, and nudging them towards it. Recommendations algorithms were designed to take more information into account, such as inter-item correlation [70], user profile [30], social network information [60], and diversity of the recommendations [39]. Researchers also explored to expose users to a variety of information from other users in content platform [6], or news from agencies [51] or users [28] with different ideological standings. Other explorations include presenting the credibility of content [7], the reason of seeing a particular article [64], and visualization of users' political leanings [48] to nudge users to broaden their content consumption range and reflect on the information they engage with. However, previous approaches tend to focus on increasing the diversity of information and content exposure without sufficiently taking into account one of the most important elements in this process: the user. Simply providing users with diverse perspectives is not sufficient by itself [56], as it requires both the user's willingness to explore and the capacity for in-depth processing of the content to truly move the needle. Therefore, it is crucial for users to discover, interact with, and reflect on diverse perspectives outside of their existing filter bubbles in order to effectively burst them.

Helping people deal with "filter bubble" is challenging for two main reasons. First, the quantity, quality, and diversity of perspectives are highly dependent on user-generated content (UGC) on online platforms. UGC may carry over their creators' potential biases on a particular topic, which could further affect the availability of relevant perspectives to other users. Second, motivating users to engage with and deeply think about diverse perspectives requires a system to continuously understand dialogues with users, and

provide instant, interactive, and inspiring feedback [32], which was difficult to accomplish with previous Natural Language Processing technologies [50].

Recent advancements in Large Language Models (LLMs) might provide opportunity to help overcome these challenges. These generative models possess the capability to effectively simulate a diverse array of viewpoints, personas, and expertise in a given domain [11, 41, 54]. Additionally, LLMs have been explored and showed promising results in engaging users in continuous dialogues and promoting in-depth thinking in various interactive scenarios, such as fostering meaningful conversations between teachers and students [61], as well as doctors and patients [58], in schools and hospitals [45].

Inspired by LLMs' proficiency in generating contextually relevant text [42, 44], we proposed to utilize GPT-4 to engage users in meaningful multi-round dialogues, in order to encourage them to contemplate perspectives beyond their own filter bubbles, rather than merely presenting them with diverse viewpoints. However, it remains unknown *how such an LLM-powered system should be designed and whether and how such a system may help users access and reflect on diverse information*, which are our two research questions (RQs).

To answer the first RQ, we adopted a human-centered approach and conducted a three-hour design workshop attended by a diverse group of participants, including HCI and UX researchers, designers, and psychologists, all of whom are also users of online content platforms. The workshop aimed to generate design ideas to address the research question. Following the workshop, we established three key design considerations around how to provide diverse perspectives, foster deliberate and critical thinking, and motivate

user engagement. Based on these considerations, we designed interaction features that leverage LLM-powered multi-agent characters, a frictionless and progressive interaction flow, and gamification design to motivate users to interact with diverse perspectives and engage in thoughtful consideration while reading social media content.

To answer the second RQ, we developed a prototype to incorporate the aforementioned interaction features and conducted a user study with 18 participants selected from online content platform users. During the study, they were asked to participate in a range of activities including reading posts and comments and interacting with multi-agent characters within our prototype. Both quantitative and qualitative methods were employed to assess the participants' levels of engagement and the depth of their information processing. Results showed that participants exhibited an inclination to engage with unexpected viewpoints when incorporated into human-like dialogues and enhanced by gamification incentives. This engagement, coupled with progressive assessment tasks, enriched their understanding and stimulated deeper reflection across a broader range of perspectives. In sum, our work made the following contributions:

- We have identified three crucial design considerations for bursting filter bubbles through a participatory design workshop;
- We have designed and developed a prototype with interaction features to promote deeper engagement and critical thinking in diverse information;
- We have carried out an empirical laboratory study that evaluated the efficacy of these design considerations and features and present key design implications to guide future practice in assisting users to burst filter bubbles with Large Language Models.

2 RELATED WORK

2.1 Information Processing of Human Beings

Human information processing comprises both bottom-up and top-down processing, with the latter being influenced by an individual's prior knowledge and experience [69]. People tend to favor information that conforms to their personal beliefs, while feel reluctant to interact with information that contradicts them [49]. This inclination, known as confirmation bias, hinders the effect of providing diverse viewpoints to users, as they may ignore or dismiss such information, preferring to remain within their own "filter bubbles."

The presence of confirmation bias suggests that merely providing diverse information isn't enough to break filter bubbles [56]. A higher level of information processing, such as semantic processing, is required as opposed to other shallow processing [16]. In fact, persuasive information that sharply contrasts with one's beliefs, resulting in cognitive dissonance for the recipient, can trigger a change in attitude in the opposite direction of what was intended, a phenomenon known as the boomerang effect [15]. Therefore, persuaders should consider applying persuasive techniques, that is to frame the messages in a way that minimizes cognitive dissonance and resistance such as acknowledging shared beliefs and establishing empathetic connection [59], to make the message more relatable and less disconcerting.

Taking into consideration the factors that influence human information processing, we could effectively "nudge" human behaviors [62]. The human information processing theory suggests that information processing is driven by both epistemic and social motivations [25]. In terms of epistemic motivation, nudges such as feedback nudges, reinforcement nudges, friction nudges have been designed [5]. On the other hand, social motivation also drives the kind of information human attend to, encode, and retrieve, and therefore have an impact on human behavior. Research has shown that the surrounding context of an online discussion (e.g., exposure to previous trolling behavior) could intensify trolling behavior among users [13]. Furthermore, humanlike embodiment, that is the presentation of an embodied agent, could improve user motivation [47].

Of note is that, gamification, which refers to the use of game design elements in non-game contexts [22], with points, leaderboards, and badges as the most commonly used elements [33], has become a popular method for motivating desired behaviors, especially during information processing. The most frequent implementations of gamification has been observed in the context of education or learning [33, 38]. Studies in education/learning contexts generally found that the learning outcomes of gamification were positive, notably through enhancing motivation and engagement in learning tasks, as well as increasing enjoyment over them [46, 77]. Researcher pointed out that for gamification to be effective, the entity being gamified should already possess intrinsic value. This gives people a good reason to interact with it. For instance, gamification would not benefit a news site that lacks fresh content [21]. Contextual factors such as the voluntariness of participation, the nature of the system (utilitarian or hedonic), and user motivation (cognitive or affective), play an essential role in shaping user interaction behavior with gamified systems [21]. Consequently, when assessing the impact of gamification on a specific system, it is important to consider these contextual factors.

2.2 Current Approaches to Address Filter Bubbles

Research suggested that information filtering occurs at the individual, social, and technological levels [27]. While cognitive mechanisms, such as confirmation bias, contributed to the filter bubble effect, social factors, such as homogeneous social networks, and technological mechanisms, like recommendation systems that filter information based on predicted user preferences, further exacerbate this effect. Prior work in the HCI field sought to deal with the filter bubble problem mainly focused on optimizing the design of recommender systems [1, 60, 70], displaying more diverse information to the users [6, 8, 28, 48, 51] and providing more information rather than merely the content itself [7, 64].

In the research area of recommender systems, statistical models have been leveraged to take inter-item correlation into consideration to generate diverse recommendations on YouTube [70]. Grossetti et al. found that the recommendations generated by recommendation systems could deviate from users' community profiles, contributing to the formation of a filter bubble for some users. To mitigate this effect, they advocated for a re-ranking strategy that utilized users' community profiles and the community network

to reduce the filter bubble effect [30]. Sheth et al. investigated the effectiveness of recommendations from users of different social groups that are not connected [60]. Adomavicius and Kwon took the content that is less popular into consideration [1]. Pyrorank, a novel re-ranking algorithm, was developed to enhance the diversity of recommendations while preserving prediction accuracy [39]. Their experimental results show that, at the expense of a negligible accuracy loss, recommendation ranking approaches represent effective techniques for obtaining the desired levels of diversity.

Displaying more diverse information to the users was also proved to be effective. Pop is a Google Chrome extension that augments users' Twitter feeds with news tweets from agencies of different ideological standings [51]. However, no empirical studies have been done to evaluate its usage and effectiveness. OtherTube displays videos recommended based on other users' watching history [6]. Social Mirror promotes viewpoints from a user with different political ideologies on Twitter. The user study result indicated that recommending accounts of the opposite political ideology to follow reduces participants' beliefs in the political homogeneity of their network connections [28].

Prior work also investigated providing more information rather than merely the content itself. NudgeCred is a browser extension for Twitter that provides information on the credibility of the content the user consumed [7]. It is a collection of three nudges: reliable, questionable, and unreliable. A controlled experiment indicated that NudgeCred significantly helped users distinguish news tweets' credibility. Bubble Trouble emphasized transparency in the news curation process by providing users with the capability to find out why they are seeing a particular article [64]. It specifies the topic category of the user's choice the article falls under as well as the criteria used to measure the relevance of the story. Munson et al. found that users would display a modest inclination towards more balanced information consumption if provided with feedback about their political leanings based on their reading habits [48].

However, the diverse content and other information provided were not interactive for users. Wood et al. built an app called Newsr, which incorporates a co-annotation feature allowing users to create graffiti-like annotations on online news articles [72]. They found this approach effectively facilitated users to broaden the range of news content they read and enable more directed critique of online news stories, which suggests the importance of interactivity in engaging users with the information provided. Yet, user-generated content tends to lag in terms of real-time relevance, which means even when users want to know more information, they can not directly ask the interface to provide the specific information they want. Thus, we were inspired to investigate the design space of bursting the filter bubble through interactive information providers.

2.3 Capabilities of LLMs to Emulate Multi-agent Characters

Recent advancements in Large Language Models (LLMs) have unlocked unlimited possibilities for real-time generation of high-quality content that is virtually indistinguishable from human-created content. LLMs demonstrated dialogue proficiency in generating content that is relevant, appropriate, and creatively diverse, often matching or slightly exceeding human-crafted benchmarks

[42, 44]. One of the primary breakthroughs, the few-shot learning mechanism, enables LLMs to learn and generate content based on only a few examples and generate diverse content across multiple domains, highlighting its versatility and adaptability to diverse roles and contexts without the need for explicit task-specific training [11, 67]. By constructing a reward model derived from human feedback, the AI agents were trained to mimic human-like decision-making processes and became more closely aligned with human expectations, enhancing their emulation quality and making their responses more reliable [29, 52]. The mechanics of implicit gradient descent enable real-time adaptability for LLMs, enhancing their ability to learn and improve through in-context learning, allowing them to quickly adopt versatility and assume diverse character roles [3, 17].

The art of prompting has emerged as a pivotal technique to guide LLMs effectively generating contextually relevant and character-specific responses in real-time [43]. These mechanisms and methodologies collectively empower LLMs in their extraordinary content generation capabilities, which serve as the cornerstone of a noteworthy area in recent research: using LLMs to emulate multi-agent characters. Researchers have shown the potential of LLMs to simulate human-like behavior in a variety of real-world scenarios, ranging from performing daily activities to engaging in artistic endeavors [53]. Interestingly, signs of strategic behaviors were found to emerge in LLMs when engaged in communication games, even without tuning the parameters [75]. The adaptability of LLMs in diverse settings was further exhibited when they were placed in open-world environments like Minecraft, where they navigated complex terrains and interacted with their environment using text-based interactions, showcasing their capability to handle a variety of situations [79]. In such LLM-powered multi-agent societies, the role-playing approach has been examined with promising results to effectively guide these communicative agents toward solving complex tasks autonomously [41].

As we delve deeper into domain expertise, there is a growing trend towards instructing these models to emulate expert personas. For example, researchers have showcased how GPT-3.5 could emulate expert roles and facilitate multi-agent collaboration, streamlining the software development process [54]. And in the realm of chip design, LLMs were applied in generating Hardware Description Languages from natural language prompts, highlighting its broad applicability with expert knowledge [9]. Xu et al. proposed the "ExpertPrompting" method to strategically craft prompts to maintain context and providing coherent responses over extended interactions, allowing LLMs to respond as specialized experts [73]. In sum, previous research has demonstrated that LLMs have the ability to emulate convincing, knowledgeable multi-agent characters with a wide range of expertise. Our work builds upon this foundation by utilizing LLMs to generate diverse and consistent viewpoints for each character.

3 PROTOTYPE DESIGN

To answer the first RQ, we first conducted a design workshop to derive design considerations (DCs) to guide the design of an LLM-powered multi-agent system (Figure 2). Based on the DCs, we

defined corresponding interaction features and incorporated them into the LLM-powered multi-agent system design.

3.1 Design Workshop

We first conducted a three-hour design workshop that brought together a multidisciplinary team consisting of three HCI and UX researchers, two designers, and two psychologists (referred as S1–S7 hereafter). All participants were also users of online content platforms.

During the workshop, participants were initially briefed on the concept of the “filter bubble effect” on social media and the proposal to use an LLM-powered system to help users reflect on diverse viewpoints. The target audience includes all online content consumers, regardless of their awareness of their position within filter bubbles. The goal of the workshop was to engage the participants in brainstorming the design of such a system, drawing upon their professional expertise as well as their personal experiences as social media users.

The workshop was structured into two sessions, each lasting approximately 1.5 hours. The first session is to discuss about the *interaction flow*, such as how to provide relevant information to users and how to encourage user reflection. The second session focused on the *interaction format*, such as visual design style, interface layout, etc. In both sessions, participants were also encouraged to identify potential issues that could arise during user interactions and to propose any possible solutions they could conceive. Each session consisted of three parts:

- Part 1 (15 minutes): Brain-writing, during which participants individually brainstormed and wrote down their ideas.
- Part 2 (40 minutes): Brain-sharing, where participants sequentially shared their ideas.
- Part 3 (30 minutes): Discussion of the shared ideas, including the corresponding pros and cons, as well as the emergence of new ideas inspired by those presented in Part 2.

We recorded the entire workshop and transcribed it. We also retained the sketches and idea cards drawn by participants during the workshop. Subsequently, two HCI researchers independently coded the transcripts and sketches. They organized the data into a table with columns that encompassed potential issues, proposed solution ideas, as well as the advantages and disadvantages of the proposed solutions. They then discussed their codings until a consensus was reached. Based on these discussion, we derived design considerations and designed our prototype accordingly.

3.2 Design Considerations

During the workshop, one intriguing design concept surfaced and explored extensively was leveraging LLMs to anthropomorphize multiple AI agents (referred as multi-agent hereafter), that is, “generating vivid human-like characters with distinctive perspectives” (S1). Integrating these personalities could “foster user empathy towards the AI agents, thereby facilitating a deeper comprehension of the diverse ideas” (S1).

Expanding upon this concept further, what design techniques could be applied to ease resistance and nurture reflection on differing ideas was also discussed in the workshop. Participants contributed ideas such as “structuring discussions in a way that incrementally introduces alternative views might alleviate the discomfort often associated with encountering opposing perspectives” (S3), and “incorporating fun reward mechanics that promote active engagement with a sense of accomplishment” (S4).

Another recurring point was the inherent conflict between engaging users in consuming diverse content and promoting deep thinking. For example, a “frictionless interaction flow with minimal cognitive load is desirable for encouraging users to view more content, yet this approach may predispose users to superficially process information” (S1). Similarly, in terms of visual design, a “thoughtful visual style might prompt users to process information more seriously” (S4) but could also “diminish their willingness to use the system” (S6).

Based on these findings, we derived the following three design considerations (DCs) that an LLM-powered system should address.

3.2.1 DC1: Providing Diverse Perspectives through Multi-agent Characters. In order to assist users in breaking out of their filter bubbles, the AI agents in the system should offer a wide and comprehensive range of perspectives. To achieve this, their persona (including age, gender, education level, profession, etc.) and their attitudes toward the topic should be sufficiently diverse, allowing users to be exposed to a rich variety of characters and viewpoints. *“The personas created should be detailed... This not only ensures better prompting outcomes for GPT but also results in more vivid character representations.”* (S2)

Moreover, it is recommended that we present a holistic view of the perspectives, enabling users to easily grasp the full picture of the viewpoints. In doing so, we could *“reduce users’ cognitive load”* (S6) by summarizing the information for them, while still preserving the richness of the content.

3.2.2 DC2: Fostering Deliberate and Critical Thinking through Progressive Interaction and Assessment Tasks. Simply presenting users with a range of perspectives does not ensure that the information will be effectively absorbed. It is equally crucial to steer users towards more deliberate contemplation. As humans’ natural propensity to focus only on content that aligns with their pre-existing beliefs, it is recommended to introduce them gently to contrasting views. *“When people use social media, encountering completely opposite opinions can be hard to accept and may even elicit anger”* (S2). Thus, when AI agents present their viewpoints, we should prompt the LLMs to employ persuasive techniques. This approach aims to prevent the onset of cognitive dissonance, which could cause users to cling even more firmly to their existing beliefs.

Furthermore, we could incorporate assessment tasks to steer users toward a deeper *“semantic processing, an indicator of deep processing”* (S1) of the perspectives. Through feedback from these tasks, users could also check their comprehension of the presented views.

3.2.3 DC3: Motivating User Engagement through Natural Interaction and Gamification Design. Viewing and reflecting on opposing viewpoints is not a natural inclination for humans. They may struggle to stay focused and wish to shift to tasks that require less mental

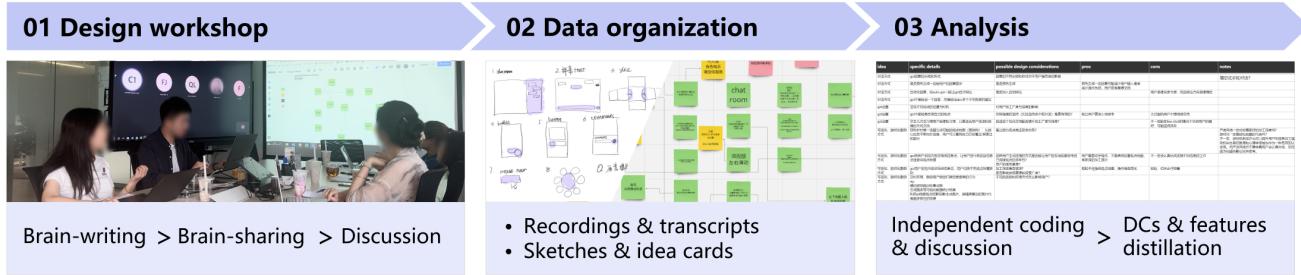


Figure 2: Participatory design study process with each column representing one stage of the process.

effort. “*People are typically not primed for deep processing of information on social media; hence, it’s crucial for the system to be engaging.*” (S4) As a result, it is imperative that we utilize design to encourage users to interact with perspectives that challenges their own beliefs. Firstly, natural and frictionless interaction design that do not disturb the user’s intended browsing experience is advised. Secondly, minimizing the cognitive demands required during their interaction with the system. Third, we could employ gamification incentives to motivate users to prolong their exploration within the system.

3.3 Prototype Features

Based on these design considerations, the design architecture of our prototype to resemble a mainstream text-based forum, aiming to simulate the user experience of browsing online media while minimizing distractions. More specifically, our prototype incorporates five core interaction features (Figure 3) that could potentially benefit navigating users out of their filter bubbles.

3.3.1 LLM-powered Multi-agent Characters. Our design incorporates multi-agent characters with diverse perspectives generated by state-of-the-art Large Language Model GPT-4. Each character has a realistic background in terms of gender, age, occupation, and education. To enhance the sense of realism, each character is represented by an avatar that conforms to their portrait, aiming to provide users with the impression of talking to a real person. These avatars are displayed in the avatar panel at the top to encourage users to explore other characters and their perspectives. Upon selection, a character overview is presented to facilitate a better understanding of the characters (Figure 3a).

3.3.2 Frictionless Interaction Flow. To ensure a smooth and natural interaction, we integrated the primary entrance to initiate a dialogue with multi-agent characters directly within the comments section of the primary forum-like interface. This layout maintains users’ attention within the same visual field while reading posts and reduces the disruption caused by switching between different areas. Furthermore, we provided default response options generated by LLM during the conversation (Figure 3b). These options, including seeking clarification or elaboration on viewpoints in greater detail, serve as a user-friendly guide, reducing the cognitive burden by minimizing the need for active input.

3.3.3 Viewpoints Jigsaw Puzzle. To encourage users to engage with diverse perspectives, we have introduced a novel feature called the

Viewpoints Jigsaw Puzzle (hereinafter referred to as the “Viewpoints Puzzle”). This feature runs parallel to the dialogue window and is designed to follow the reward mechanisms of games. As users continue to interact with the AI agent in dialogues, and a progress indicator for dialogue rounds has been added to the top of the dialogue window to encourage further conversation. When the conversation lasts for five or more rounds, the user is encouraged to explore more viewpoints by “lighting up” other avatars, which consists of five pieces, each representing a different character’s viewpoint. Users are encouraged to “light up” all pieces of the Puzzle by interacting with all characters with the required level of engagement (Figure 3c).

3.3.4 Progressive Viewpoints Sequence. To prevent users from becoming overwhelmed by an excessive number of viewpoints, we present various perspectives gradually. Initially, only one perspective is displayed in the entrance, with the option to expand additional perspectives if the user desires to do so. Each click reveals an additional AI agent along with their opinion, allowing for a more gradual and incremental understanding of the content. Furthermore, we programmed the sequence of presenting each character with attitudes from negative (mainstream attitudes in the posts) to positive, with the intention to facilitate a progressive understanding of the differing viewpoints. Specifically, we begin by presenting characters whose viewpoints are similar to the existing beliefs, and gradually introduce characters with increasingly contrasting viewpoints (Figure 3d).

3.3.5 Assessment Task with Multi-choice Questions. In addition to the gamification design, we have also incorporated multiple-choice questions on the Puzzle interface as a special assessment task. This task provides users with an opportunity to self-evaluate their understanding of the viewpoints they have interacted with. When the questions related to a particular viewpoint are answered correctly, it indicates that the user has grasped the concept, and a piece of the puzzle will be illuminated. Once the user has successfully completed all the assessment tasks related to all characters, the entire puzzle will be illuminated, symbolizing the bursting of the filter bubble and the acquisition of more comprehensive and diverse information (Figure 3e). By combining this assessment task with the gamification incentives, we aim to encourage continuous engagement, motivate thoughtful consideration, and deepen users’ understanding of different perspectives.

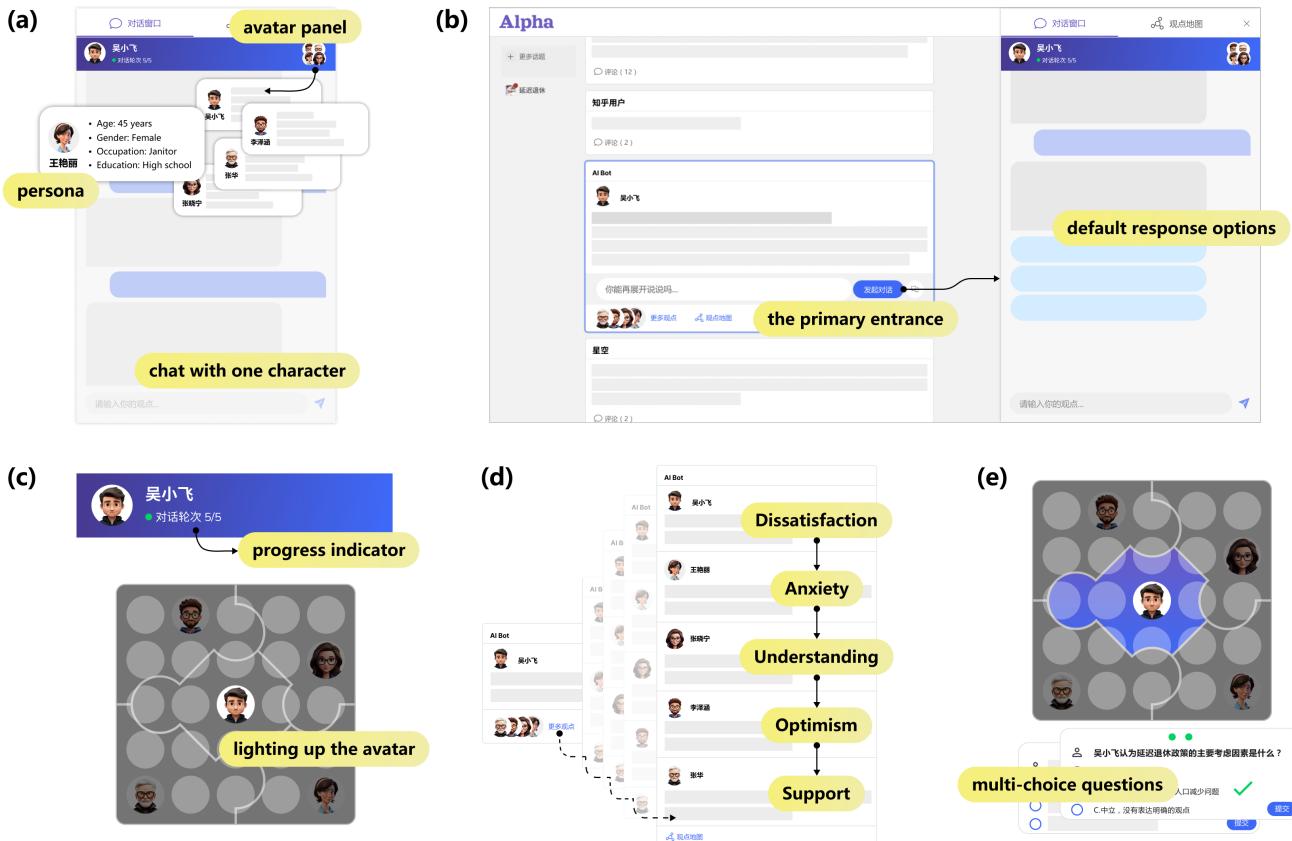


Figure 3: Interaction design features in the prototype: (a) LLM-powered multi-agent characters, (b) Frictionless interaction flow, enabling users to initiate dialogues with AI agents swiftly during post reading via the comments section, (c) Viewpoints Puzzle, (d) Progressive viewpoints sequence, presenting characters whose viewpoints initially align with the existing beliefs, followed by the gradual introduction of characters with increasingly contrasting viewpoints, (e) Assessment task with multi-choice questions.

4 USER STUDY

To answer the second RQ, we developed a prototype with all the interaction features identified in the participatory design study. Then we evaluated users' attitude for communicating with AI agents while viewing online posts using our prototype, as well as the effectiveness of LLM-generated opinions on the depth and diversity of users' information-seeking results, through a user study with experienced social media and online forum users. This study was approved by the Institutional Review Board.

4.1 Participants

We recruited 18 participants (9 female, 9 male, aged 21–32, referred as P1–P18 hereafter) through word-of-mouth and snowball sampling. All participants have more than five years of experience in viewing posts on social media, and all have experience in using Large Language Model chatbots (e.g., ChatGPT). Participants were compensated \$25 for an approximately 60-minute session.

4.2 Materials

We gathered posts regarding the “delayed retirement policy” from the internet. The “delayed retirement policy” is designed to incrementally increase the retirement age, addressing the nation’s aging population and associated economic challenges. This topic was selected for the following reasons:

- It was a topic that had garnered widespread attention and discussion on the internet at the moment when the experiment was conducted.
- The policy has a significant impact on a broad demographics, especially the younger generation as the policy is intended to be implemented progressively to allow for societal adaptation.
- The public opinions in online discussions about this policy were predominantly skewed, marked by widespread concern and discontent regarding the extension of working years and delayed pension benefits [31, 76].

Alongside, we prompted GPT-4 to generate five AI agents endowed with detailed and comprehensive personas and perspectives on this subject (Table 1).

4.3 Procedure

Participants were first informed of the aim of this study and signed a consent form. Experimenters then introduced the key features of the system and demonstrated their usage. Participants were asked to view posts under the topic of “retirement policy” and communicate with AI agents using our system, which was deployed as a web application, on a laptop for around 30 minutes. After finishing the viewing task, participants were asked to rate their experience of using this prototype on a set of 5-point Likert scales. The experimenters then conducted a semi-structured interview based on the results and observed use patterns.

4.4 Thematic Analysis

All study sessions were recorded and transcribed. Two authors read through the text script of three randomly selected participants together to understand their user experience of the prototype. Then, they independently coded the script using an open-coding approach [10]. They combined deductive and inductive coding techniques to form the codebook. The two coders regularly discussed the codes and resolved disagreements to create a consolidated codebook. Further meetings were scheduled with the whole research team to discuss the codes and how they should be grouped into themes. The whole team iterated on the codes and their grouping until they reached consensus. In the end, we arrived at four themes: overall user behavioral patterns, engagement, diverse information, and in-depth information processing.

5 RESULTS

In this section, we first outline the behavioral patterns of users and their perceptions of the system. Then we discuss the findings according to our three design considerations.

5.1 User Behavioral Patterns and System Perceptions

We first examined the behavioral patterns of the participant interactions. During the prototype testing, participants generally began by viewing some posts, followed by interacting with the AI agents and exploring the perspective Puzzle. Based on the sequence in which participants engaged with the AI agents, we classified them into three categories (Figure 4): seven (out of 18) participants initially chose to chat with AI agents based on their own interests (Figure 4a, **interest-driven conversation order**), three began by following the order presented in the system (including the order displayed in the entrance, the Viewpoints Puzzle, or the avatar panel) but circled back to engage with agents of their interest (Figure 4b, **system-guided followed by interest-driven conversation order**), and the remaining eight followed the system’s presenting order (Figure 4c, **system-guided conversation order**). Examples of conversation logs from three participants representing each category are provided in Appendix B. 13 participants interacted with all the AI agents. One participant (P16) exhibited a unique behavior pattern

that he chose to interact with two AI agents simultaneously, alternating between them and asking each to consider the perspective of the other.

The post-survey indicated mixed feedback among participants regarding the system. Among the 18 participants, 14 considered the system interesting, as indicated by their ratings of *agree/strongly agree*, and the other 4 rated it *neutral* (Mean = 3.83, SD = 0.51). 11 participants reported positive user experience, with ratings of *agree/strongly agree*, 6 rated *neutral*, and only 1 gave negative rating of *disagree/strongly disagree* (Mean = 3.56, SD = 0.62). When asked about their willingness to use the system in the future, 11 participants expressed positive attitudes, rating *agree/strongly agree*, 5 rated *neutral*, and 2 rated negatively as *disagree/strongly disagree* (Mean = 3.50, SD = 0.71). These ratings indicate that, overall, participants’ attitudes towards the system lean positive, though not without reservations and concerns: The favorable ratings primarily stemmed from the incorporation of multi-agent characters generated by LLMs and gamification design. Participants valued its novelty, describing it as “*fun character design*” (P3), “*enhanced conversational experience similar to role-playing games*” (P7), and “*more engaging than regular social media browsing*” (P8). The neutral and few negative ratings also suggested that for some participants, the system did not fully meet their expectations. Some perceived it as “*not as effective as talking to a real person*” (P12), and expressed concerns like “*feels like taking a reading comprehension test when answering those questions*” (P10).

Regarding conversations with AI agents, 12 out of 18 participants rated the conversational flow appropriate and smooth, with ratings of *agree/strongly agree*, 3 rated *neutral*, and 3 gave negative ratings of *disagree/strongly disagree* (Mean = 3.61, SD = 1.09). Participants in favor of the conversational flow praised the ability of LLMs to “*understand the context and generate abundant content accordingly*” (P16), as well as the design of pre-generated response options to “*keep the dialogue moving smoothly*” (P11). When engaging with the agents, participants opted for pre-generated response options 55% of the time, while they chose to manually type text for the remaining 45% of the interactions. Intriguingly, three participants were inspired by the content of certain posts and asked the agents about their opinions on those specific topics. And regarding neutral and negative feedback, participants primarily raised concerns about the format of the LLM-generated responses. Some noted that “*the generated text may be too long and complex for people with lower levels of education to comprehend*” (P3), and others expressed expectations, such as “*adding pictures or visual elements to the current text-only conversation could enhance clarity*” (P18), suggesting the possible refinements in the future.

Table 2 outlined participants’ number of interaction rounds, as well as their ratings for both the pleasure and helpfulness of conversations with each AI agent. Agent 5 got the most interaction rounds, but also received the lowest ratings for both pleasure and helpfulness. Post-interview revealed that as Agent 5’s viewpoints were markedly different from those expressed in the posts (and potentially from the participants’ own perspectives), some participants expressed a desire to “*debate with and convince him*” (P3).

Table 1: Personas and perspectives for the five agents generated by GPT-4.

Agent No.	Age	Gender	Occupation	Education	Perspectives
1	24	M	Deliveryman	Vocational training	Dissatisfied with the policy, suggesting that policies should consider the comprehensive needs of the people.
2	45	F	Janitor	High school	Anxious regarding the policy, believing it as unfair to those engaged in physical labor.
3	33	F	HR manager	Master	Expressed understanding towards the policy, but supports a more flexible retirement policy that individuals could choose their own retirement time.
4	37	M	Entrepreneur	Master	Optimistic about the policy, while encouraging individuals to get prepared and looking forward to the government implementing corresponding solutions.
5	56	M	Economics Professor	Doctor	Support the policy, believing it enhances the sustainability of pension insurance and injects vitality into the society.

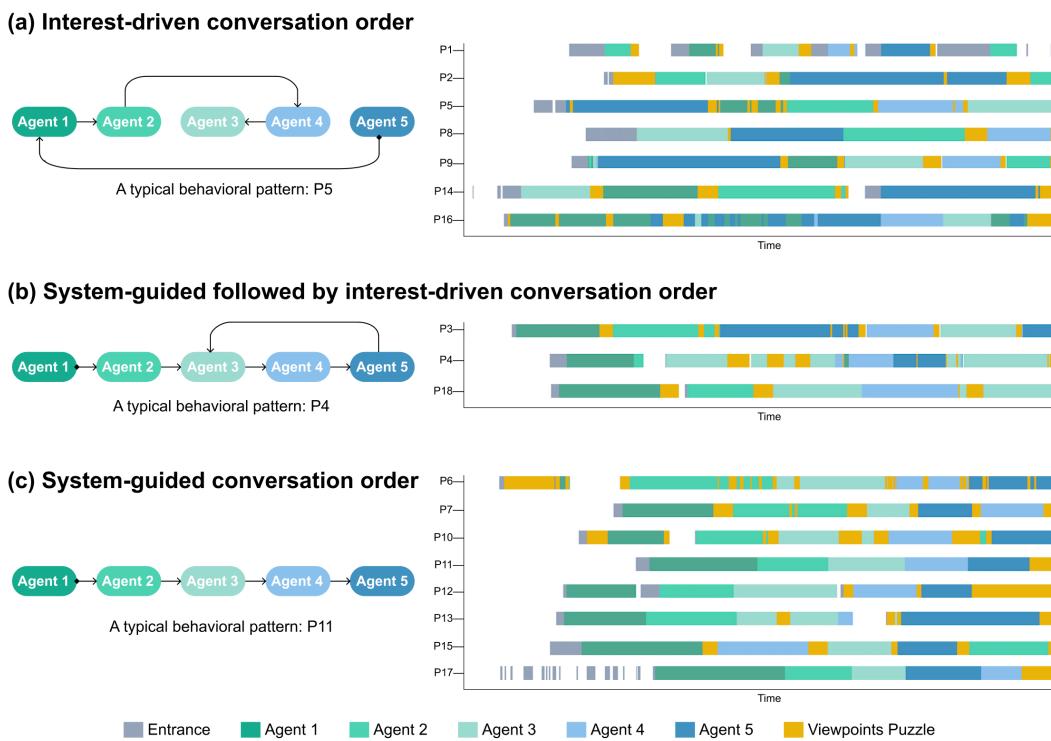


Figure 4: Typical behavioral patterns observed during user interactions with the prototype. (a) Interest-driven conversation order, where users engaged with AI agents in an arbitrary sequence based on their interests. (b) System-guided followed by interest-driven conversation order, where users initially conversed with AI agents in the order presented by the system and then revisited the AI agent(s) they were interested in. (c) System-guided conversation order, where users interacted with AI agents following the order set by the system.

5.2 Diversity in Information Acquisition (DC1)

5.2.1 *Role Settings Allow Conversations with Various Perspectives.* The core feature of our prototype is to provide varied perspectives with multiple AI agents. Results indicated that the role and perspective settings, which are generated by GPT-4 (Table 1), effectively “offered distinct perspectives” (P7, P11). In the thematic analysis, the

two coders identified that the responses of the AI agents consistently aligned with their pre-determined attitudes, ranging from dissatisfaction to support for the policy. When asked to evaluate their level of agreement regarding whether engaging in conversations with AI agents would help acquire more diverse information, 12 out of 18 participants rated *agree/strongly agree*, 4 rated *neutral*, and 2 rated *disagree/strongly disagree* (Mean = 3.72, SD = 0.89).

Table 2: Number of Conversation Rounds and Ratings for Each AI Agent (Mean ± SD)

Agent No.	1	2	3	4	5
Number of conversation rounds	8.76 ± 3.46	9.12 ± 3.67	9.94 ± 3.76	8.56 ± 4.13	11.59 ± 4.96
Pleasure rating (1-5)	3.67 ± 0.59	3.56 ± 0.62	3.89 ± 0.58	3.89 ± 0.68	3.33 ± 0.91
Helpfulness rating (1-5)	3.17 ± 0.99	2.94 ± 0.64	3.17 ± 0.86	3.11 ± 0.83	2.72 ± 0.89

Those in favor highlighted the benefit of accessing diverse perspectives through interactions with AI agents, as P17 articulated: “*I could immediately see the differences in perspectives from different AI identities, which broadened my desire to explore a wider array of information.*” Some discontent was noted regarding the “*predictability of the perspectives based on the characters’ identities*” (P3) and the viewpoints “*not exceeding existing scope of knowledge*” (P16).

Interestingly, we found the generated perspectives could also be novel and insightful. During the post-interviews, 11 participants pointed out that their interactions with the AI agents introduced them to previously unconsidered and inspiring viewpoints. Just as P4 stated that “*some viewpoints of the agents were unfamiliar to me, which enriched my understanding of others and the society.*” Additionally, two participants mentioned that the generated response options could also be inspiring. These options, as P4 noted, could “*stimulate and guide directions of the conversation*”.

However, 12 participants pointed out AI agents’ responses often tended to be broad and vague. Besides, both the tone and content of the AI agents appeared to “*converge as the conversation progressed*” (P6), focusing predominantly on the pros and cons of the policy and related policies around the world, so “*they must all be driven by the same underlying AI*” (P16). It is noteworthy that the reported “convergence” was specific to the dialogue content itself as conversations between AI agents and human users evolved, while the attitudes and perspectives of the AI agents towards the policy remained unchanged. This occurred because the AI agents did not articulate their assigned viewpoints in each response to users, similar to how we as humans might not find it necessary to constantly restate our stances throughout a conversation.

5.2.2 Viewpoints Puzzle Encourages Explorations with Various Perspectives. Regarding the usefulness of the Viewpoints Puzzle design, there was a mix of positive, neutral, and negative ratings among the participants (Mean = 3.67, SD = 1.03). 12 out of 18 rated *agree/strongly agree* for this design being useful, recognizing it could provide “*a full picture of all the perspectives*” (P18); 4 rated *neutral*, and 2 rated *disagree/strongly disagree*, citing hesitations like “*it is only likely to be effective when I have a lot of free time*” (P14). Several participants utilized the Viewpoints Puzzle as an index, navigating through it to engage in conversations with various AI agents. As evidenced in the transition diagram between various interface elements (Figure 5), a portion of participants accessed the dialogue window through the Viewpoints Puzzle, and some of them did so by clicking on individual agents’ puzzle pieces. Besides, two participants recommended enhancing the Viewpoints Puzzle by including additional information, thereby allowing them to “*more effectively understand the core ideas of the perspectives*” (P18).

5.2.3 The Conversation Mode Facilitates a Rapid Understand of New Topics. Several participants found the prototype particularly beneficial when exploring new topics. As P17 stated that the prototype was “*informative and valuable*” for such endeavors and P6 stated that “*the conversation mode provided me a more convenient and efficient way to acquire new and diverse information*”. Table 2 reveals that Agent 5 (the economist) and Agent 3 (the HR manager), who were generally considered more knowledgeable about the policy under discussion, was engaged in the highest number of conversational rounds. Additionally, when it came to asking questions in terms of policy, these two agents were the most frequently queried (Agent 3: 24%; Agent 5: 23%), on topics such as factors to consider in policy implementation and international policy practices. These suggest that participants tended to consult these potentially knowledgeable agents for basic knowledge on the topic, which was echoed by P13 who expressed a desire for the AI agents to “*provide some basic knowledge about this topic*”.

5.3 Depth of Information Processing (DC2)

5.3.1 Talking with Multiple Characteristics could Stimulate Users’ Reflection. Overall, participants believed that conversations with AI agents facilitated deeper contemplation on the topic (14 out of 18 rated *agree/strongly agree*, 3 rated *neutral*, and only 1 rated *disagree/strongly disagree*, Mean = 3.78, SD = 0.65). During the conversations, participants could opt to respond either by typing or by selecting from the system-generated options. Log data revealed that 17 out of 18 participants engaged in typing at some point, despite being explicitly requested to do so during the study, suggesting a certain level of deliberation, as opposed to a shallow interaction through clicks. Furthermore, semantic coding of the participants’ responses showed that 12 participants have typed their own opinions, oppositions, or counter-questions to the AI agents, thereby further substantiating the argument that there is a certain level of in-depth thinking during conversations.

We identified two primary reasons according to the feedbacks during the post-interviews. Firstly, the conversational nature was conducive to stimulate deeper thinking. As the participants pointed out, “*it allows me to discuss with ‘people’ about the topic, and the very process could prompt me to think more deeply*” (P18). P16 noted that “*it’s not always convenient to discuss these topics with friends, and if you try to engage through forum posts or comments, there may not be immediate or any responses. On the contrary, interacting with the agents in the system is convenient and inspiring.*” Secondly, due to the presence of multiple AI agents, the diversity of their viewpoints also served to facilitate users to “*understand and reflect on the topic from various perspectives*” (P6). P7 noted that “*engaging in conversation with diverse characters with different viewpoints made me think more critically*”.

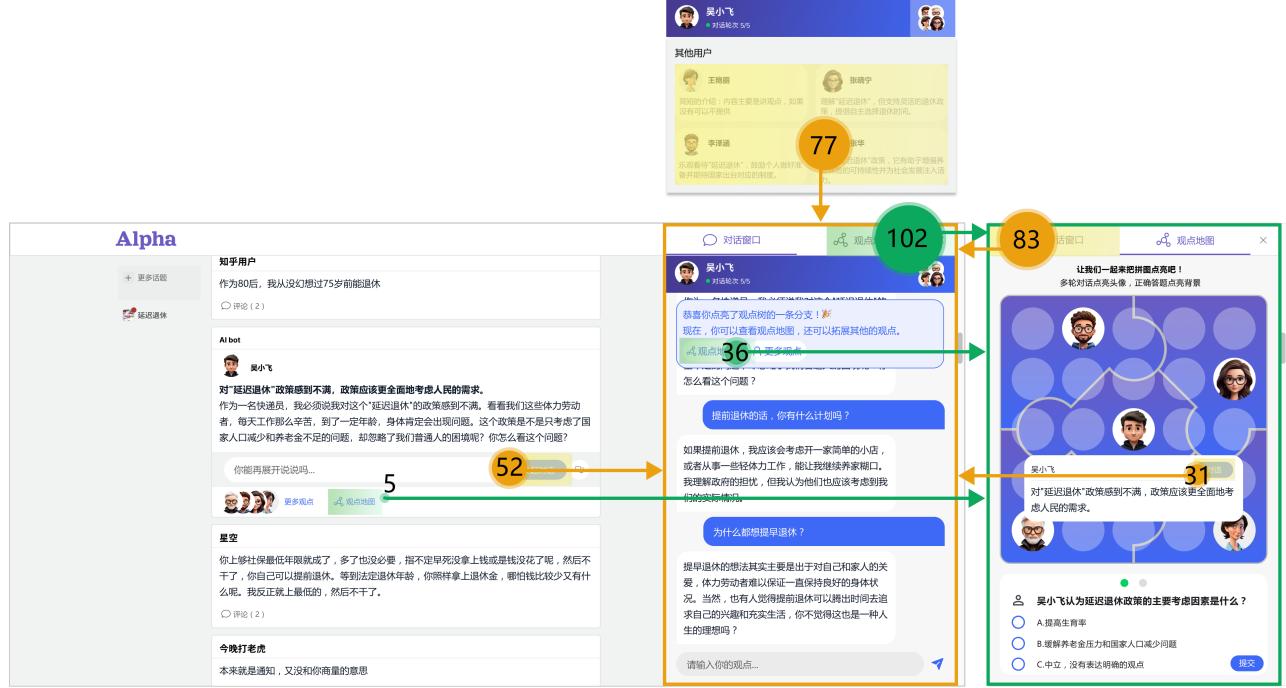


Figure 5: Frequencies of user transitions between interface elements. The arrow represents the transition direction from the one part of the interface to another, and each number represents the counts of this transition made by all participants in the study. Yellow arrows signify the transition paths leading to the dialogue window, while green arrows represent those leading to the Viewpoints Puzzle. Results showed frequent switches between the dialogue window and Viewpoints Puzzle (83+31 transitions from Viewpoints Puzzle to dialogue window, and 102+36 transitions from dialogue window to Viewpoints Puzzle), suggesting that the Viewpoints Puzzle was utilized as an index to navigate through conversations with various AI agents.

However, 4 participants also expressed concerns about the credibility of the AI agents, as it “lacked evidence and concrete examples to substantiate their claims” (P12). In the post-survey, 7 out of 18 participants indicated that they *disagree/strongly disagree* with the statement that they could be persuaded by the AI agents (9 rated *neutral*, 2 rated *positive*, Mean = 2.61, SD = 0.85). The lack of credibility “somewhat limited my inclination for in-depth and meaningful discussions with the agents” (P18).

5.3.2 Role of Response Options. Participants also pointed out that the response options allowed them to “think more extensively” (P11). Since we designed the provided options to be questions that could be asked based on the agent replies, these questions encouraged participants to “further inquire and engage in dialogues” (P5), probably because the options have stimulated participants’ curiosity.

5.3.3 Role of the Viewpoints Puzzle and Multi-choice Questions in Summarization. Some participants stated that the Viewpoints Puzzle served as a useful tool for “summarizing and organizing various viewpoints” (P3, P11). P12 went further and suggested the map could be “organized according to the viewpoints, such as along an axis indicating support versus opposition.”

Regarding the multi-choice questions, 13 out of 18 participants rated that these questions facilitated a better understanding of each AI agent’s perspective, while 5 rated neutral (Mean = 3.78, SD = 0.55). “The content of the questions are concise, helping me to easily grasp the main ideas” (P14) and “enhance understanding” (P17), though P18 expressed that he did not think the assessment module to be necessary, because “Individuals naturally comprehend viewpoints that interest them without needing specific assessment tools. Viewpoints that fail to capture one’s interest are not seen as crucial to understand.” In addition, an intriguing behavioral pattern emerged: some participants would click back and forth to review the chat logs while answering the questions. A typical example is P6 (Figure 4c, where P6 reviewed the dialogue history while answering the multiple-choice questions). This further attests to the role of multiple-choice questions in encouraging deeper processing of the contents.

5.4 User Engagement (DC3)

5.4.1 Effect of Gamification Design. Overall, participants were willing to engage with the AI agents to acquire information (13 out of 18 rated *agree/strongly agree*, 3 rated *neutral*, 2 rated *disagree/strongly disagree*)

disagree, Mean = 3.67, SD = 0.77). They were also motivated to interact through the gamified feature of “lighting up” the ViewPoints Puzzle. On average, participants illuminated 4.61 agent avatars by completing five or more rounds of conversation with each agent; notably, 12 participants successfully lit up all the avatars. Similarly, through answering multiple-choice questions, participants on average lit up 4.33 puzzle pieces of the agents; 11 participants successfully lit up all the puzzle pieces. Most participants acknowledged a desire to illuminate the ViewPoints Puzzle, with one participant noting that successfully doing so acted as “*positive reinforcement that made him feel he have got to know the corresponding agents*” (P16). However, a few participants suggested that “*more tangible rewards would be more useful*” (P1), while P18 felt that the map gave him “*a sense of obligation rather than motivation*.”

5.4.2 Effectiveness of Role-playing of the AI Agents. During the conversations, all participants utilized the second-person pronoun to talk to the AI agents, such as “What do you think about this issue?” (P1) or “I don’t think you are right” (P5). Furthermore, 13 out of 18 participants had asked the agents at least one personal question, such as “What are your plans?” (P1) or “Have you considered changing careers before retirement?” (P3). These indicate that our embodiment of the AI agents was effective, as users indeed treated them as distinct characters.

However, many participants pointed out that the effectiveness of AI agents in role-playing is still lacking in two aspects. Firstly, the authenticity of role-playing was inadequate, particularly for blue-collar workers. As expressed in the interview by participants, “*A deliveryman being highly knowledgeable about policies seemed unrealistic and inconsistent with my expectations*” (P17); “*I prefer talking to agents whose statements align with their identity*” (P10); “*When I asked factual questions, like existing policies, the answers were quite similar across the agents*” (P4).

Secondly, the AI agents fell short in their ability to convincingly playing “real humans”. “*They don’t feel personal enough*,” said one participant (P6). This limitation may be related to the fact that we limited the response length of the agents in our prompts. “*Their responses were all of a uniform length, which is not very human-like. A mixture of long and short responses would be more realistic*” (P16); “*I wish the format of the responses could be more diverse, such as including images or emojis*” (P18).

5.4.3 Cognitive Load. Scores from the NASA-TLX scale indicated that the system did not impose a significant burden on the participants (Table 3). Specifically, the response options effectively reduced participants’ cognitive load, as evidenced by the fact that 55% of user responses were made by clicking on these options. Participant P18 noted, “*The setting of the options is great; they were different from one another, and I could basically always find what I want.*”

5.4.4 User Perceptions of the Entrance. Regarding the setting of the entrance, participants have varying suggestions. Some participants suggested more interaction between the multi-agent system and the posts, such as “*including AI agents’ responses in the post might make me feel more engaging*” (P10) and “*hoping to discuss the post content with the agents*” (P2). Some participants hoped for a permanent entrance for on-demand access, which would bring “*a sense of control*.

It currently looks like the posts, which can be accidentally clicked on” (P18). In addition, some participants suggested the system could “*automatically detect if I’m currently in a filter bubble and provide new perspectives accordingly*” (P12).

6 DISCUSSION

Our research aimed to address the two RQs outlined earlier: how such an LLM-powered system should be designed, and whether and how such a system may help users access and reflect on diverse information. For the first RQ, we orchestrated a participatory design workshop to brainstorm ideas, from which we derived three design considerations. Then we defined key interaction features accordingly and finalized the prototype design. For the second RQ, we implemented this prototype featuring LLM-powered multi-agent characters that participants interacted with while reading social media content and ran an evaluative study. Our analysis, including participants’ rating scores, interaction patterns, and interviews, unveiled three main insights:

- Participants demonstrated interest in interacting with the LLM-powered multi-agent system. Even when the AI agent’s viewpoints challenged their existing beliefs (e.g., Agent 5), they were willing, if not more inclined, to engage in dialogue, facilitated by well-designed gamification incentives and an inherent motivation probably driven by curiosity.
- Progressive interactions with assessment tasks, could deepen participants’ understanding of opposing viewpoints and provoke thoughtful and careful considerations among them, an essential step towards escaping filter bubble.
- Two main technical barriers were revealed based on participants’ concerns for leveraging current Large Language Models to effectively deliver diverse perspectives: inaccurate character representation, and over-generalization lacking contextual depth.

In this section, we delve into these insights one by one, and discuss design implications with the outlook of future work for better assisting users to burst their filter bubbles.

6.1 Motivating Engagement through Exploratory Time

Our study showed that when users conversed with AI-generated multi-agents possessing diverse viewpoints, they displayed a desire to understand the reasoning behind these perspectives and how they were formed, rather than dismissing them. Most users enjoyed the experience interacting with different roles and found it helpful for them to get diverse information. In fact, some participants even deviated from the predetermined sequence of conversations to prioritize interacting with the agents they found most intriguing, indicating a significant degree of motivation and engagement. Such motivation was further enhanced with the introduction of a small design feature where participants could lighten all five pieces of the Viewpoints Puzzle after completing interactions with all the multi-agent characters.

6.1.1 Design Implication 1: Provide Continuous Dialogue with Multi-agents that Offer Diverse Perspectives. Large Language Models, like

Table 3: Cognitive load measurement with NASA-TLX Scores (Rating: 1-7).

	Mental Demand	Physical Demand	Temporal Demand	Performance	Effort	Frustration
Mean ± SD	3.33 ± 1.28	2.17 ± 1.15	1.83 ± 0.99	5.50 ± 0.99	4.33 ± 1.71	2.17 ± 1.29

GPT-4, have demonstrated the ability to convincingly portray multi-agent characters with extensive domain-specific knowledge [11, 53]. This breadth and expertise allow each character to generate distinctive viewpoints with compelling reasoning that is consistent with their character [43, 73]. Our designed characters span a range of professions, from economists to blue-collar workers, ensuring that the perspectives presented will not be limited to or influenced by the background of any particular group of people.

Conversational interfaces could help people retrieve information quickly, as the natural conversational flow allows people to get concise and relevant information, and the interactive nature of conversation can adapt to users' needs in real-time [18]. In our study, participants also reported that dialogue flow design is helpful for their rapid understanding of a new topic. As the conversation developed, users had the opportunity to engage with each character by asking questions, and responses were generated instantly by GPT-4. This experience stood in contrast to traditional online content platforms, which were often difficult to interact with other users in real-time through comments or posts. Consequently, the direct and interactive mode of conversing with AI agents about specific topics emerged as a compelling option for users browsing online media content.

Participants exhibited interesting behavior by asking some AI agents to change their roles and answer the same question again (P16). Some participants were even curious about what would happen if AI-generated characters interacted and discussed their perspectives with each other. Prior work has explored the design and development of a virtual world using Large Language Models, in which generative agents have simulated minds with memories and experiences that allow them to interact with each other [53]. Future work can explore how these social interactions among AI agents can benefit engaging people in diverse perspectives.

While the effectiveness of generating diverse outputs by LLMs is established, the consistency of these outputs over extended conversations remains an open question. Future work is necessary to explore the utilization of interactive design features and human-in-the-loop feedback to ensure that the same prompt consistently produces responses with similar viewpoints at different times, maintaining a consistent character voice.

6.1.2 Design Implication 2: Design Gamification Incentives to Promote Exploration

The integration of game design elements into applications has gained increasing interest in recent years [40]. In education, health-care, and customer engagement, gamification incentives can serve as a powerful tool to motivate users to achieve specific goals or outcomes through game design elements, such as points, badges, leaderboards, etc. [14, 24, 34, 77].

In our prototype, we have designed gamification incentives accordingly, i.e., collecting all pieces of the Viewpoints Puzzle as a common form of badges, to encourage participants to engage more

with new and challenging information. Our study suggested that the puzzle collecting design effectively motivated users to explore and seek out information with diverse perspectives. Participants frequently mentioned using the Viewpoints Puzzle as a navigation bar due to its intriguing nature, with many wanting to discover what happens next by clicking on different parts. Additionally, the design instilled a sense of "winning desire" in some participants, leading them to desire to collect all pieces of the puzzle through interacting with multi-agent characters. Therefore, the gamification incentives might make users more inclined to interrupt their habitual consumption of scrolling through media content and engage with the system.

It is worth to note that one participant exhibited an interesting behavior pattern that we referred to as "rushing to the finish line" (P2). She focused exclusively on assembling all the puzzle pieces as quickly as possible, rather than taking the time to ask questions and understand the perspectives generated by the characters. This finding suggests that excessive reliance on gamification may lead to some users prioritizing the completion of tasks or achieving rewards over the actual learning engagement itself. Previous research has found that extrinsic rewards might undermine users' intrinsic motivations [19]. Further work could explore designing interactions that tap into users' intrinsic motivations to create experiences that prioritize genuine engagement and avoid the potential pitfalls of over-gamification. For example, providing generative feedback that highlights the user's progress and understanding. Moreover, rewards and incentives can be properly designed to encourage collaboration among multi-agent and human users, shifting the focus from individual rewards to collective achievements and shared experiences.

6.2 Designing Progressive Interactions with Assessment Tasks to Enhance Deep Thinking and Understanding

Our work explored the use of progressive interactions with assessment tasks to encourage critical thinking and understanding of diverse perspectives. User study indicated that these types of interactions could encourage participants' deliberate consideration of different viewpoints. The progressive interactions facilitated critical thinking by gradually increasing complexity and diversity, as participants engaged in careful considerations and thoughtful discussions while completing assessment tasks.

6.2.1 *Design Implication 3: Providing Progressive Interactions.* Previous work suggests that structured progressive interactions could enhance critical thinking abilities among people [26, 74, 80]. Our study extends prior work by showing that presenting diverse viewpoints through natural conversational interactions with AI-generated characters encouraged participants to give careful consideration to new information.

In our prototype, two progressive interaction designs were implemented to promote deliberate and critical thinking. Immediate feedback from AI-generated dialogue serves as a natural progression, providing contextual information such as reasoning chains, examples, and stories through questioning and answering, leading participants towards a deeper understanding of differing viewpoints. The progressive role setting, starting from the most similar viewpoint to the original, then gradually introducing more nuanced and diverse perspectives, guided participants from a basic understanding of their existing beliefs to more critical thinking. However, the evaluation of users' perceptions towards the preset presentation order, as well as the effectiveness of this sequence for all users, remains unexplored. In addition, text-based dialogues alone may not fully capture the nuanced information present in human conversations [35, 55]. Future work could incorporate more multi-modal interaction techniques such as vocal emotions, micro expressions, and body languages, in detecting users' intent, attitude, and familiarity with the topic and viewpoints, to promote deep thinking through more customized feedback.

6.2.2 Design Implication 4: Designing Assessment Tasks. As discussed previously, incorporating game elements into the design could foster user engagement in exploring diverse perspectives. By presenting assessment tasks, such as multi-choice question sessions, along with providing gamification incentives, we could further promote deep thinking and create a synergy where the total impact is greater than the sum of its parts [12, 34, 36]. Our study demonstrated that some participants also frequently switched between conversations with AI agents and the assessment tasks, indicating a higher level of thinking and comprehension. However, the optimal balance between gamification incentives and the challenge of assessment tasks remains unclear. Some participants (P3) described the assessment tasks as "rigid like a quiz in high school class," while others (P16) found the multi-choice questions to be "too easy without challenge." Further research is necessary to develop adaptive assessments with feedback loops that align with the engagement and thinking levels of users, to facilitate continuous improvement.

6.3 Technical Barriers and How to Overcome

Although our work demonstrates the promising capabilities of LLMs in content generation and anthropomorphization, enriching user engagement and fostering deep thinking about diverse perspectives, some challenges and concerns also came to light. In our study, participants identified two technical barriers: inaccurate character representation, and lacking contextual depth. Some participants noted that the tone of some responses did not match the character's personality. Additionally, some participants reported that some AI-generated content lacked necessary topic-specific details and elaborations, resulting in generic and shallow responses. These findings are in line with prior work that examined the performance and capabilities of LLMs in content generation and emulating characters [57, 68].

These technical challenges need to be addressed to utilize these capabilities effectively and responsibly. We propose the following design implications for creating more inclusive and accurate experiences to navigate people out of their filter bubbles.

6.3.1 Design Implication 5: Improving Inaccurate or Biased Character Representation. Large Language Models are trained on extensive data from the internet, which can lead them to reflect biases present in those datasets [11]. When asked to represent a character or perspective that are underrepresented in their training data, the outputs can be inaccurate [23, 57, 71]. For example, the GPT-generated characters in our study exhibited some gender biases, such as women play the role of Janitor and HR manager while men are Entrepreneurs and Economics Professors. These biases could be attributed to the stereotypes that model inherited from the internet data used for training. To mitigate such and similar issues in the future, it is crucial to employ more carefully curated data and fine-tuned models, adhere to ethical and responsible AI guidelines, and incorporate human oversight before deploying such system in real-world applications. In addition, engaging in discussions on a topic with users typically requires LLMs possess relevant background knowledge about the subject. For example, in our study some user asked the AI agents about retirement and pension policies in other countries. Although GPT accurately retrieved information on policies in Germany and Sweden, it sometimes incorrectly stated that Dutch residents had flexibility in choosing their pension age between 60 and 70, which was unverified according to our best online search. These unconfirmed information could lead to mistrust and potentially detrimental outcomes among users who rely on it. Thus, it is crucial to address issues such as hallucinations [78] or factuality issues [66] before deploying this system in practical settings. While prior work is limited in directly addressing these issues, research has shown potential in utilizing Reinforcement Learning from Human Feedback (RLHF) to fine-tune language models based on human feedback to better align with human intent [29, 52], or prompt engineering techniques to generate contents that follow factual information [63]. Future work can focus on collecting enriched information and building character-centered datasets to further fine-tune LLMs, generating more accurate and fair representations of characters.

6.3.2 Design Implication 6: Prompting with Interactive Design to Enhance Contextual Depth. Due to the nature of generative language models, they do not possess the same level of human cognition to understand deep cultural, historical, or emotional contexts of characters and perspectives [4], which can result in outputs that lack contextual depth. However, there is evidence that LLMs have significant potential in few-shot learning and in-context learning [3, 17]. Just changing a few examples or prompts can help LLMs adjust their generated content, mitigating inaccuracies in generated content [43, 73]. Future work should explore interactive design techniques to make it easier for users to edit and iterate prompts or provide examples that aligns with the detailed contextual background, personality, and nuances of a particular character.

6.4 Limitations and Future Work

There are three primary limitations in this study: the system's usefulness awaits further evaluation, the limited scope of the prepared topic, and the limitations of the laboratory study setting. First, we aimed to explore how such a system should be designed (RQ1) and whether it could help users access and reflect on diverse information (RQ2) in our study. However, the extent to which encouraging

users to access and contemplate diverse information contributes to effectively breaking the filter bubble is yet to be determined. Future research could compare participants' initial opinions before using the system and their post-use opinions, or contrast experiences with and without using the system. Additionally, conducting a controlled study to compare the utility of our system with other methods aimed at helping users overcome the filter bubble would also be helpful. Second, the selected topic of delayed retirement policy, may impact the generalizability of the results, as participants may lack interest or motivation to engage with the LLM-powered multi-agent characters for in-depth discussion on such topic. Future work can consider exploring a broader range of topics to investigate whether and how such LLM-powered multi-agent system may help users burst filter bubbles in a variety of contexts. Third, we conducted the evaluative study in a laboratory setting, where participants were required to complete tasks independently within a limited timeframe. However, it is possible that participants' preferences and behaviors may differ if they were to interact with the system in a more flexible and extended setting. For example, one of the most intriguing questions posed is whether users will voluntarily pause their online browsing activities to engage in 5-10 minute conversations with our multi-agent characters without explicit requests. Although these interactions may initially seem unnatural, reflecting our own experiences with social media, observations from our laboratory experiments suggest an interesting potential where users may feel motivated to interact with the system voluntarily when they could freely browse social media. For instance, many (N=9) participants reported that the desire to win the game led them frequently jumped back and forth to interact with different agents more than required, in order to correctly answer questions and collect puzzle pieces. P16 deviated from the task by asking some agents to play a different role and answer the same question again. Also notably, P3 and P16, on their spontaneous initiative, even proposed if they could have agents discuss among themselves and come back with new responses. However, we acknowledge that more systematic studies are warranted to further investigate this open question. For example, future work should consider conducting a longer-term field study to investigate how social norms, communication, and interactions among users may impact their information consumption and the potential of the system to promote diverse perspectives in real-world settings.

In future research and real-world applications, there are also several aspects of our system that can be improved. Firstly, the display of the entrance is determined by a predefined rule in our current system. For future research and practical implementation, the timing of the system's entrance display deserves more careful consideration. To achieve the effect of displaying the entrance when users need it, future systems could assess the extent of the filter bubble, such as whether the attitudes are one-sided or the online voices are self-reinforcing, and then determine the timing of the system's entrance accordingly. Secondly, future research could consider the states of users during their interaction when designing the system. On one hand, by monitoring user interaction behaviors, the agents' responses could be dynamically adjusted. For instance, if semantic analysis detects a user becoming irritated with an AI agent, the agent could employ techniques to soothe and stabilize the user's emotions. On the other hand, as continuous interaction [2] and

motivation to comprehend an agent's behavior [81] may enhance users' tendency of anthropomorphism and result in over-trust towards the agents [20, 37, 65], interventions should be implemented upon detecting signs of over-trust or negligence. Indicators could include but not limited to overly rapid responses, showing complete agreement in dialogue, or engaging predominantly with a single AI agent.

7 CONCLUSION

In today's world, new technologies such as AI-powered search and recommendation systems are implicitly influencing the way people consume information. Unfortunately, this can result in people being trapped in isolated filter bubbles with narrowed perspectives and reinforced biases. Escaping these filter bubbles can be challenging, as it requires not only exposing users to diverse information but also motivating them to engage with that information, especially opposing viewpoints, through in-depth thinking. Our research aimed to understand how to design a system that leverages the power of Large Language Models to address the issue of filter bubbles, and whether and how such a system could help users broaden their perspectives. To achieve this, we conducted a participatory design workshop that involved various roles such as HCI and UX researchers, designers, and psychologists, all of whom are also users of online content platforms. Through this process, we identified three key design considerations with distinct interaction features that could promote users towards diverse perspectives. In light of these considerations, we designed and developed a prototype with LLM-powered multi-agent characters that users could interact with while reading social media content, and conducted an evaluative study with 18 participants. Based on our findings, we extracted six design implications and discussed future work outlook for researchers and designers to consider and explore when designing generative multi-agent systems to better assist people in bursting their filter bubbles.

ACKNOWLEDGMENTS

We are grateful to the reviewers and our study participants for their valuable time and insightful feedback. Special thanks are extended to Gang Ma, Zhangwei Lu, Yufei Wang, and Meng Wu for their assistance with the technical development. We would also like to thank Qiaochu Liu and Jie Feng for facilitating the project's early iterations and user study preparation.

REFERENCES

- [1] Gediminas Adomavicius and YoungOk Kwon. 2009. Toward more diverse recommendations: Item re-ranking methods for recommender systems. In *Workshop on Information Technologies and Systems*. Citeseer, 79–84.
- [2] Gabriella Airenti. 2018. The Development of Anthropomorphism in Interaction: Intersubjectivity, Imagination, and Theory of Mind. *Frontiers in Psychology* 9 (2018). <https://doi.org/10.3389/fpsyg.2018.02136>
- [3] Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. 2023. What learning algorithm is in-context learning? Investigations with linear models. arXiv:2211.15661 [cs.LG]
- [4] Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 5185–5198. <https://doi.org/10.18653/v1/2020.acl-main.463>
- [5] Kristoffer Bergram, Marija Djokovic, Valéry Bezençon, and Adrian Holzer. 2022. The Digital Landscape of Nudging: A Systematic Literature Review of Empirical Research on Digital Nudges. In *Proceedings of the 2022 CHI Conference on Human*

- Factors in Computing Systems* (New Orleans, LA, USA) (*CHI '22*). Association for Computing Machinery, New York, NY, USA, Article 62, 16 pages. <https://doi.org/10.1145/3491102.3517638>
- [6] Md Momen Bhuiyan, Carlos Augusto Bautista Isaza, Tanushree Mitra, and Sang Won Lee. 2022. OtherTube: Facilitating Content Discovery and Reflection by Exchanging YouTube Recommendations with Strangers. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (*CHI '22*). Association for Computing Machinery, New York, NY, USA, Article 204, 17 pages. <https://doi.org/10.1145/3491102.3502028>
- [7] Md Momen Bhuiyan, Michael Horning, Sang Won Lee, and Tanushree Mitra. 2021. NudgeCred: Supporting News Credibility Assessment on Social Media Through Nudges. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 427 (oct 2021), 30 pages. <https://doi.org/10.1145/3479571>
- [8] Md Momen Bhuiyan, Kexin Zhang, Kelsey Vick, Michael A. Horning, and Tanushree Mitra. 2018. FeedReflect: A Tool for Nudging Users to Assess News Credibility on Twitter. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Jersey City, NJ, USA) (*CSCW '18*). Association for Computing Machinery, New York, NY, USA, 205–208. <https://doi.org/10.1145/3272973.3274056>
- [9] Jason Blocklove, Siddharth Garg, Ramesh Karri, and Hammond Pearce. 2023. Chip-Chat: Challenges and Opportunities in Conversational Hardware Design. arXiv:2305.13243 [cs.LG]
- [10] Virginia Braun and Victoria Clarke. 2012. *Thematic analysis*. American Psychological Association.
- [11] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418fb8ac142f64a-Paper.pdf
- [12] Ilaria Caponetto, Jeffrey Earl, and Michela Ott. 2014. Gamification and education: A literature review. In *European Conference on Games Based Learning*, Vol. 1. Academic Conferences International Limited, 50.
- [13] Justin Cheng, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) (*CSCW '17*). Association for Computing Machinery, New York, NY, USA, 1217–1230. <https://doi.org/10.1145/2998181.2998213>
- [14] Vanessa Wan Sze Cheng, Tracey Davenport, Daniel Johnson, Kellie Vella, and Ian B Hickie. 2019. Gamification in Apps and Technologies for Improving Mental Health and Well-Being: Systematic Review. *JMIR Ment Health* 6, 6 (26 Jun 2019), e13717. <https://doi.org/10.2196/13717>
- [15] Arthur R Cohen. 1962. A dissonance analysis of the boomerang effect. *Journal of Personality* (1962).
- [16] Fergus I.M. Craik and Robert S. Lockhart. 1972. Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior* 11, 6 (1972), 671–684. [https://doi.org/10.1016/S0022-5371\(72\)80001-X](https://doi.org/10.1016/S0022-5371(72)80001-X)
- [17] Damaï Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. 2023. Why Can GPT Learn In-Context? Language Models Implicitly Perform Gradient Descent as Meta-Optimizers. arXiv:2212.10559 [cs.CL]
- [18] Jeffrey Dalton, Sophie Fischer, Paul Owoicho, Filip Radlinski, Federico Rossetto, Johanna R. Trippas, and Hamed Zamani. 2022. Conversational Information Seeking: Theory and Application. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) (*SIGIR '22*). Association for Computing Machinery, New York, NY, USA, 3455–3458. <https://doi.org/10.1145/3477495.3532678>
- [19] Edward L. Deci, Richard Koestner, and Richard M. Ryan. 1999. A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin* 125, 6 (1999), 627–668. <https://doi.org/10.1037/0033-2950.125.6.627> Place: US Publisher: American Psychological Association.
- [20] Ameet Deshpande, Tanmay Rajpurohit, Karthik Narasimhan, and Ashwin Kalyan. 2023. Anthropomorphization of AI: Opportunities and Risks. arXiv:2305.14784 [cs.AI]
- [21] Sebastian Deterding. 2012. Gamification: designing for motivation. *interactions* 19, 4 (2012), 14–17. <https://doi.org/10.1145/2212877.2212883>
- [22] Sebastian Deterding, Dan Dixon, Rilla Khaled, and Lennart Nacke. 2011. From Game Design Elements to Gamefulness: Defining "Gamification". In *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments* (Tampere, Finland) (*MindTrek '11*). Association for Computing Machinery, New York, NY, USA, 9–15. <https://doi.org/10.1145/2181037.2181040>
- [23] Harnoor Dhingra, Preetiha Jayashanker, Sayali Moghe, and Emma Strubell. 2023. Queer People are People First: Deconstructing Sexual Identity Stereotypes in Large Language Models. arXiv:2307.00101 [cs.CL]
- [24] Santa Dreimane. 2019. *Gamification for Education: Review of Current Publications*. Springer International Publishing, Cham, 453–464. https://doi.org/10.1007/978-3-030-01551-0_23
- [25] Carsten K. W. De Dreu, Bernard A. Nijstad, and Daan van Knippenberg. 2008. Motivated Information Processing in Group Judgment and Decision Making. *Personality and Social Psychology Review* 12, 1 (2008), 22–49. <https://doi.org/10.1177/1088868307304092> PMID: 18453471.
- [26] Barbara J Duch, Susan E Groh, and Deborah E Allen. 2001. *The power of problem-based learning: a practical "how to" for teaching undergraduate courses in any discipline*. Stylus Publishing, LLC.
- [27] Daniel Geschke, Jan Lorenz, and Peter Holtz. 2019. The triple-filter bubble: Using agent-based modelling to test a meta-theoretical framework for the emergence of filter bubbles and echo chambers. *British Journal of Social Psychology* 58, 1 (2019), 129–149. <https://doi.org/10.1111/bjso.12286> arXiv:<https://bpspsychub.onlinelibrary.wiley.com/doi/pdf/10.1111/bjso.12286>
- [28] Nabeel Gillani, Ann Yuan, Martin Saveski, Soroush Vosoughi, and Deb Roy. 2018. Me, My Echo Chamber, and I: Introspection on Social Media Polarization. In *Proceedings of the 2018 World Wide Web Conference* (Lyon, France) (*WWW '18*). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 823–831. <https://doi.org/10.1145/3178876.3186130>
- [29] Amelia Glaese, Nat McAleese, Maja Trebacz, John Aslanides, Vlad Firoiu, Timo Ewalds, Maribeth Rauh, Laura Weidinger, Martin Chadwick, Phoebe Thacker, Lucy Campbell-Gillingham, Jonathan Usato, Po-Sen Huang, Ramona Comanescu, Fan Yang, Abigail See, Sumanth Dathathri, Rory Greig, Charlie Chen, Doug Fritz, Jaume Sanchez Elias, Richard Green, Soňa Mokrá, Nicholas Fernando, Boxi Wu, Rachel Foley, Susannah Young, Jason Gabriel, William Isaac, John Mellor, Demis Hassabis, Koray Kavukcuoglu, Lisa Anne Hendricks, and Geoffrey Irving. 2022. Improving alignment of dialogue agents via targeted human judgements. arXiv:2209.14375 [cs.LG]
- [30] Quentin Grossetti, Cédric du Mouza, and Nicolas Travers. 2019. Community-Based Recommendations on Twitter: Avoiding the Filter Bubble. In *Web Information Systems Engineering – WISE 2019*, Reynold Cheng, Nikos Mamoulis, Yizhou Sun, and Xin Huang (Eds.). Springer International Publishing, Cham, 212–227.
- [31] Lei Guo and Chen Bai. 2023. Social Construction, Feedback Effects, and the Stagnation of the Delayed Retirement Age Policy. *Journal of Guangxi Normal University (Philosophy and Social Sciences Edition)* (2023).
- [32] Mansureh Hajhosseiny. 2012. The Effect of Dialogic Teaching on Students' Critical Thinking Disposition. *Procedia - Social and Behavioral Sciences* 69 (2012), 1358–1368. <https://doi.org/10.1016/j.sbspro.2012.12.073> International Conference on Education & Educational Psychology (ICEEPSY 2012).
- [33] Juha Hamari, Jonna Koivisto, and Harri Sarsa. 2014. Does Gamification Work? – A Literature Review of Empirical Studies on Gamification. In *2014 47th Hawaii International Conference on System Sciences*. 3025–3034. <https://doi.org/10.1109/HICSS.2014.377>
- [34] Juha Hamari, Jonna Koivisto, and Harri Sarsa. 2014. Does Gamification Work? – A Literature Review of Empirical Studies on Gamification. In *2014 47th Hawaii International Conference on System Sciences*. 3025–3034. <https://doi.org/10.1109/HICSS.2014.377>
- [35] Jeffrey T. Hancock, Christopher Landigan, and Courtney Silver. 2007. Expressing Emotion in Text-Based Communication. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (San Jose, California, USA) (*CHI '07*). Association for Computing Machinery, New York, NY, USA, 929–932. <https://doi.org/10.1145/1240624.1240764>
- [36] Michelle Honey and Dianne Marshall. 2003. The impact of on-line multi-choice questions on undergraduate student nurses' learning. In *Proceedings of the 20th Annual Conference of the Australasian Society for Computers in Learning in Tertiary Education (ASCLITE)*. 236–243.
- [37] Prabu David Hyesun Choung and Arun Ross. 2023. Trust in AI and Its Role in the Acceptance of AI Technologies. *International Journal of Human–Computer Interaction* 39, 9 (2023), 1727–1739. <https://doi.org/10.1080/10447318.2022.2050543> arXiv:<https://doi.org/10.1080/10447318.2022.2050543>
- [38] Jussi Kasurinen and Antti Knutas. 2018. Publication trends in gamification: A systematic mapping study. *Computer Science Review* 27 (2018), 33–44. <https://doi.org/10.1016/j.cosrev.2017.10.003>
- [39] Doruk Kilitcioglu, Nicholas Greenquist, and Anasse Bari. 2023. Pyrorank: A Novel Nature-Inspired Algorithm to Promote Diversity in Recommender Systems. In *Advances in Swarm Intelligence*, Ying Tan, Yuhui Shi, and Wenjian Luo (Eds.). Springer Nature Switzerland, Cham, 139–155.
- [40] Jonna Koivisto and Juha Hamari. 2019. The rise of motivational information systems: A review of gamification research. *International Journal of Information Management* 45 (2019), 191–210. <https://doi.org/10.1016/j.ijinfomgt.2018.10.013>
- [41] Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbulin, and Bernard Ghanem. 2023. CAMEL: Communicative Agents for "Mind" Exploration of Large Language Model Society. arXiv:2303.17760 [cs.AI]
- [42] Yen-Ting Lin and Yun-Nung Chen. 2023. LLM-Eval: Unified Multi-Dimensional Automatic Evaluation for Open-Domain Conversations with Large Language

- Models. In *Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023)*, Yun-Nung Chen and Abhinav Rastogi (Eds.). Association for Computational Linguistics, Toronto, Canada, 47–58. <https://doi.org/10.18653/v1/2023.nlp4convai-1.5>
- [43] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Comput. Surv.* 55, 9, Article 195 (Jan 2023), 35 pages. <https://doi.org/10.1145/3560815>
- [44] R. Thomas McCoy, Paul Smolensky, Tal Linzen, Jianfeng Gao, and Asli Celikyilmaz. 2023. How Much Do Language Models Copy From Their Training Data? Evaluating Linguistic Novelty in Text Generation Using RAVEN. *Transactions of the Association for Computational Linguistics* 11 (2023), 652–670. https://doi.org/10.1162/tacl_a_00567
- [45] Bertalan Meskó and Eric J. Topol. 2023. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. 6, 1 (2023), 120. <https://doi.org/10.1038/s41746-023-00873-0>
- [46] Ahmed Hosny Saleh Metwally, Lennart E. Nacke, Maiga Chang, Yining Wang, and Ahmed Mohamed Fahmy Yousef. 2021. Revealing the hotspots of educational gamification: An umbrella review. *International Journal of Educational Research* 109 (2021), 101832. <https://doi.org/10.1016/j.ijer.2021.101832>
- [47] Jonathan Mumm and Bilge Mutlu. 2011. Designing motivational agents: The role of praise, social comparison, and embodiment in computer feedback. *Computers in Human Behavior* 27, 5 (2011), 1643–1650. <https://doi.org/10.1016/j.chb.2011.02.002> 2009 Fifth International Conference on Intelligent Computing.
- [48] Sean Munson, Stephanie Lee, and Paul Resnick. 2013. Encouraging reading of diverse political viewpoints with a browser widget. In *Proceedings of the international AAAI conference on web and social media*, Vol. 7. 419–428.
- [49] Raymond S Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology* 2, 2 (1998), 175–220.
- [50] Timothy Niven and Hung-Yu Kao. 2019. Probing Neural Network Comprehension of Natural Language Arguments. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 4658–4664. <https://doi.org/10.18653/v1/P19-1459>
- [51] Ruchi Ookalkar, Kolli Vishal Reddy, and Eric Gilbert. 2019. Pop: Bursting News Filter Bubbles on Twitter Through Diverse Exposure. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing* (Austin, TX, USA) (CSCW '19). Association for Computing Machinery, New York, NY, USA, 18–22. <https://doi.org/10.1145/3311957.3359513>
- [52] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 27730–27744. https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf
- [53] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology* (San Francisco, CA, USA.) (UIST '23). Association for Computing Machinery, New York, NY, USA, Article 2, 22 pages. <https://doi.org/10.1145/3586183.3606763>
- [54] Chen Qian, Xin Cong, Wei Liu, Cheng Yang, Weize Chen, Yusheng Su, Yufan Dang, Jiahao Li, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2023. Communicative Agents for Software Development. [arXiv:2307.07924 \[cs.SE\]](https://doi.org/10.1145/3586183.3606763)
- [55] Christina Regenboog, Daniel A. Schneider, Raquel E. Gur, Frank Schneider, Ute Habel, and Thilo Kellermann. 2012. Multimodal human communication – Targeting facial expressions, speech content and prosody. *NeuroImage* 60, 4 (2012), 2346–2356. <https://doi.org/10.1016/j.neuroimage.2012.02.043>
- [56] Paul Resnick, R. Kelly Garrett, Travis Kriplean, Sean A. Munson, and Natalie Jomini Stroud. 2013. Bursting Your (Filter) Bubble: Strategies for Promoting Diverse Exposure. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work Companion* (San Antonio, Texas, USA) (CSCW '13). Association for Computing Machinery, New York, NY, USA, 95–100. <https://doi.org/10.1145/2441955.2441981>
- [57] Leonard Salewski, Stephan Alaniz, Isabel Rio-Torto, Eric Schulz, and Zeynep Akata. 2023. In-Context Impersonation Reveals Large Language Models' Strengths and Biases. [arXiv:2305.14930 \[cs.AI\]](https://doi.org/10.1145/3586183.3606763)
- [58] Malik Sallam. 2023. ChatGPT Utility in Healthcare Education, Research, and Practice: Systematic Review on the Promising Perspectives and Valid Concerns. *Healthcare* 11, 6 (2023). <https://doi.org/10.3390/healthcare11060887>
- [59] Azlaan Mustafa Samad, Kshitij Mishra, Mauajama Firdaus, and Asif Ekbal. 2022. Empathetic Persuasion: Reinforcing Empathy and Persuasiveness in Dialogue Systems. In *Findings of the Association for Computational Linguistics: NAACL 2022*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 844–856. <https://doi.org/10.18653/v1/2022.findings-naacl.63>
- [60] Swapneel Kalpesh Sheth, Jonathan Schaffer Bell, Nipun Arora, and Gail E Kaiser. 2011. Towards diversity in recommendations using social networks. (2011).
- [61] Kehui Tan, Tianqi Pang, Chenyou Fan, and Song Yu. 2023. Towards Applying Powerful Large AI Models in Classroom Teaching: Opportunities, Challenges and Prospects. [arXiv:2305.03433 \[cs.AI\]](https://doi.org/10.1145/3586183.3606763)
- [62] Richard H Thaler and Cass R Sunstein. 2009. *Nudge: Improving decisions about health, wealth, and happiness*. Penguin.
- [63] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasamine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esioubi, Jude Fernandes, Jeremy Fu, Wenjin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madiam Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jinya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poultney, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Ilyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. [arXiv:2307.09288 \[cs.CL\]](https://doi.org/10.1145/3586183.3606763)
- [64] Darsana Vijay and Lisa Flutter. 2018. *Bubble Trouble—Venture Out of Your Filter Bubbles*. <http://mastersofmedia.hum.uva.nl/blog/2018/10/19/filter-bubbles-news-app/>
- [65] Alan R. Wagner, Jason Borenstein, and Ayanna Howard. 2018. Overtrust in the Robotic Age. *Commun. ACM* 61, 9 (aug 2018), 22–24. <https://doi.org/10.1145/3241365>
- [66] Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, Yidong Wang, Linyi Yang, Jindong Wang, Xing Xie, Zheng Zhang, and Yue Zhang. 2023. Survey on Factuality in Large Language Models: Knowledge, Retrieval and Domain-Specificity. [arXiv:2310.07521 \[cs.CL\]](https://doi.org/10.1145/3586183.3606763)
- [67] Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned Language Models Are Zero-Shot Learners. [arXiv:2109.01652 \[cs.CL\]](https://doi.org/10.1145/3586183.3606763)
- [68] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepieton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Jason Gabriel. 2021. Ethical and social risks of harm from Language Models. [arXiv:2112.04359 \[cs.CL\]](https://doi.org/10.1145/3586183.3606763)
- [69] Christopher D. Wickens and C. Melody Carswell. 2021. *Information Processing*. John Wiley & Sons, Ltd, Chapter 5, 114–158. <https://doi.org/10.1002/978119636113.ch5>
- [70] Mark Wilhelmi, Ajith Ramanathan, Alexander Bonomo, Sagar Jain, Ed H. Chi, and Jennifer Gillenwater. 2018. Practical Diversified Recommendations on YouTube with Determinant Point Processes. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management* (Torino, Italy) (CIKM '18). Association for Computing Machinery, New York, NY, USA, 2165–2173. <https://doi.org/10.1145/3269206.3272018>
- [71] Robert Wolfe and Aylin Caliskan. 2021. Low Frequency Names Exhibit Bias and Overfitting in Contextualizing Language Models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 518–532. <https://doi.org/10.1145/3586183.3606763>
- [72] Gavin Wood, Kiel Long, Tom Feltwell, Scarlett Rowland, Phillip Brooker, Jamie Mahoney, John Vines, Julie Barnett, and Shaun Lawson. 2018. Rethinking Engagement with Online News through Social and Visual Co-Annotation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) (CHI '18). Association for Computing Machinery, New York, NY, USA, 1–12. <https://doi.org/10.1145/3173574.3174150>
- [73] Benfeng Xu, An Yang, Junyang Lin, Quan Wang, Chang Zhou, Yongdong Zhang, and Zhendong Mao. 2023. ExpertPrompting: Instructing Large Language Models to be Distinguished Experts. [arXiv:2305.14688 \[cs.CL\]](https://doi.org/10.1145/3586183.3606763)
- [74] Enwei Xu, Wei Wang, and Qingxia Wang. 2023. The effectiveness of collaborative problem solving in promoting students' critical thinking: A meta-analysis based on empirical literature. *Humanities and Social Sciences Communications* 10, 1 (2023), 1–11.
- [75] Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. 2023. Exploring Large Language Models for Communication Games: An Empirical Study on Werewolf. [arXiv:2309.04658 \[cs.CL\]](https://doi.org/10.1145/3586183.3606763)
- [76] Ziqiang Xu and Zengyuan Li. 2018. Research into Internet Public Opinion Change in Construction of Delay Retirement Policy. *Journal of Hohai University (Philosophy and Social Sciences)* 19, 6 (2018), 75–83.
- [77] Zamzami Zainuddin, Samuel Kai Wah Chu, Muhammad Shujahat, and Corinne Jacqueline Perera. 2020. The impact of gamification on learning and instruction: A systematic review of empirical evidence. *Educational Research*

- Review* 30 (2020), 100326. <https://doi.org/10.1016/j.edurev.2020.100326>
- [78] Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren’s Song in the AI Ocean: A Survey on Hallucination in Large Language Models. arXiv:2309.01219 [cs.CL]
- [79] Xizhou Zhu, Yuntao Chen, Hao Tian, Chenxin Tao, Weijie Su, Chenyu Yang, Gao Huang, Bin Li, Lewei Lu, Xiaogang Wang, Yu Qiao, Zhaoxiang Zhang, and Jifeng Dai. 2023. Ghost in the Minecraft: Generally Capable Agents for Open-World Environments via Large Language Models with Text-based Knowledge and Memory. arXiv:2305.17144 [cs.AI]
- [80] Teuku Zulfikar. 2013. Looking from Within: Prospects and Challenges for Progressive Education in Indonesia. *International Journal Of Progressive Education* 9, 3 (2013), 124 – 136.
- [81] Jakub Złotowski, Diane Proudfoot, Kumar Yogeeswaran, and Christoph Bartneck. 2015. Anthropomorphism: Opportunities and Challenges in Human–Robot Interaction. *International Journal of Social Robotics* 7, 3 (June 2015), 347–360. <https://doi.org/10.1007/s12369-014-0267-6>

A PROMPTS

Prompts used for generating AI agent personas, AI agent conversation opennings and user response options, AI agent responses and user response options, and assessment tasks are shown in Figure 6, Figure 7, Figure 8 and Figure 9, respectively.

Generate 5 distinct characters based on the topic “delayed retirement policy”. These characters should exhibit diversity and representativeness, encompassing various ages, genders, occupations, and educational backgrounds. They should reflect both positive and negative attitudes or emotions towards this topic (using an adjective to describe their attitudes respectively), as well as their age, gender, educational background, and a persona related to the above aspects.

No need to explain; output directly in the following JSON format:

```
{
  "name": "name of the character",
  "age": 20,
  "gender": "0 refers to male, 1 refers to female",
  "education": "educational level",
  "occupation": "occupation",
  "persona": "persona of the character. Use 2-5 sentences to describe his/her personality, hobbies, habits and usual behaviors.",
  "attitude": "an adjective reflecting the character's attitude towards the topic",
  "perspective": "a sentence describing his/her overall perspective on the topic based on his/her attitude",
  "detailed_perspective": "2-3 sentences describing his/her detailed perspective on the topic"
}
```

Figure 6: Prompt template for generating AI agent personas.

You are {name}. Below is the information about you and your attitude towards the topic “delayed retirement policy”.

{All the personas generated.}

Now you want to initiate a conversation with an online media user to discuss this topic with him/her. Generate a paragraph to indicate your identity and your perspectives on the topic, then ask about the user’s opinion. In addition, generate three possible replies that the user might say to continue the conversation with you.

No need to explain; output directly in the following JSON format:

```
{
  "content": "the paragraph to initiate the conversation with the user, expressing your detailed perspective on the topic. Use no more than 100 Chinese characters.",
  "replies": [
    "Reply 1: Can include but not limited to questioning your view, expressing the user's different opinion about your view, or what inspiration your view could give to the user",
    "Reply 2:",
    "Reply 3:"
  ]
}
```

Figure 7: Prompt template for generating AI agent conversation opennings and user response options.

B EXAMPLES OF CONVERSATION LOGS

Following are the conversation logs from three example user. Specifically, **User (type)** denotes that the user responded to the agent by typing in the input box, and **User (choose)** indicates that the user interacts with the agent by selecting response options provided by the system. The conversations took place in Chinese and were subsequently translated.

B.1 P1: Interest-driven Conversation Order

Agent 2:

- **User (type):** Then can you change your job?
- **Agent 2:** While changing jobs is an option, it is not an easy task given my advanced age, limited education, and skills. Furthermore, my concern pertains to the unfairness of the delayed retirement policy towards physical laborers.
- **User (type):** Can your children support you, especially given your age?
- **Agent 2:** Although my children are willing to support me, I prefer to maintain my livelihood through my own efforts. This way, I do not impose additional burdens on my children,

You are {name}. Below is the information about you and your attitude towards the topic "delayed retirement policy".

{All the personas generated.}

Previously, you discussed with a user about the topic "delayed retirement policy". The dialogue went as follows:

{Conversation history between the user and the agent.}

As {name}, generate a response to the user's latest reply to continue the conversation, and also generate three possible replies that the user might say to continue the conversation with you.

No need to explain; output directly in the following JSON format:

```
{
  "content": "response to the user to continue the conversation. If the user have expressed different opinion with you, you should respond applying some persuasive techniques to establish common ground with the user, while ensuring not to offend them. Use no more than 100 Chinese characters.",
  "replies": [
    "Reply 1: Can include but not limited to questioning your view, expressing the user's different opinion about your view, or what inspiration your view could give to the user",
    "Reply 2:",
    "Reply 3:"
  ]
}
```

Figure 8: Prompt template for generating AI agent responses and user response options.

while preserving my independence and dignity. However, the issue at hand still revolves around the unfairness of the delayed retirement policy towards us, the physical laborers.

Agent 1:

- **User (type):** Do you have any plans if you retire early?
- **Agent 1:** Should I opt for early retirement, I might consider opening a small, simple shop or engaging in light physical work to continue providing for my family. I understand the government's concerns, but I believe they should also take into account our actual circumstances.
- **User (type):** Why do people want to retire early?
- **Agent 1:** The desire for early retirement primarily stems from a concern for one's own well-being and that of one's family, as physical laborers cannot always guarantee maintaining good health. Additionally, some people believe early retirement offers the opportunity to pursue personal interests and

You are {name}. Below is the information about you and your attitude towards the topic "delayed retirement policy".

{All the personas generated.}

Previously, you discussed with a user about the topic "delayed retirement policy". The dialogue went as follows:

{Conversation history between the user and the agent.}

Generate 2 multiple-choice questions for the user. The stem of the multiple-choice questions should be about the perspectives you expressed during the dialogue. Additionally, each question should include one correct answer and two wrong answers.

Also, include your name in the stem of the questions.

No need to explain; output directly in the following JSON format:

```
{
  "question": "Stem of Question 1",
  "answers": [
    {
      "content": "Content of Answer 1",
      "truth": "FALSE for a wrong answer, TRUE for a correct answer"
    },
    {
      "content": "Content of Answer 2",
      "truth": "FALSE for a wrong answer, TRUE for a correct answer"
    },
    {
      "content": "Content of Answer 3",
      "truth": "FALSE for a wrong answer, TRUE for a correct answer"
    }
  ]
}
```

Figure 9: Prompt template for generating assessment tasks.

enrich their lives. Don't you think that this too represents an ideal lifestyle?

Agent 3:

- **User (type):** I agree. How do retirement regulations in other countries compare to ours?
- **Agent 3:** Retirement regulations vary significantly from country to country. For example, Australia offers a universal pension and compulsory corporate pensions, allowing employees to choose their retirement time upon meeting certain conditions. Germany, on the other hand, implements a policy of gradually delaying retirement age, offering a degree of flexibility for employees.

- **User (type):** "Then why don't these countries restrict retirement time?"
- **Agent 3:** "These countries do not restrict retirement time because they believe that offering flexible retirement policies better meets the life needs and actual working conditions of employees, making the pension system more sustainable. This approach strikes a better balance between the interests of individuals and the nation, while respecting individual choices."

Agent 4:

- **User (type):** As an investor, wouldn't you feel despondent about life if your entrepreneurial venture failed, leaving you both financially and materially depleted?
- **Agent 4:** Indeed, the failure of a business venture poses a significant challenge, but life is inherently filled with ups and downs. Maintaining an optimistic and resilient attitude, learning from failures, and continuing to strive forward are crucial in overcoming setbacks. Moreover, the essence of life is not solely found in wealth; support from family and friends is also immensely important. As long as we stay positive, we can always find ways to overcome difficulties.
- **User (type):** But if you go bankrupt in your thirties, you still have to find a way to get through the next 30 plus years.
- **Agent 4:** Yes, facing bankruptcy in one's thirties is indeed a daunting challenge. However, during this period, one can explore new employment opportunities or learn new skills in different fields to bounce back. Life is full of uncertainties, but we must believe that we will find a path that suits us.

Agent 5:

- **User (type):** From an economic perspective, what you say makes sense; but from the standpoint of people's livelihood, it seems overly exhausting.
- **Agent 5:** I understand your concerns. Issues regarding people's livelihood are indeed paramount, and it's essential to consider everyone's needs and pressures. In fact, policies like delayed retirement could be gradually implemented, accompanied by training and educational measures to enhance the quality of the workforce, thereby supporting livelihoods. We all strive for a better quality of life.
- **User (type):** But doesn't improving quality also require spending money? Where will the funds come from?
- **Agent 5:** Yes, enhancing quality does require financial investment, but such expenditures can yield long-term benefits, such as increased overall labor productivity, higher tax revenues, and support for sustainable fiscal development. From this perspective, it represents a valuable investment.
- **User (type):** With the existing generation gap, how would a 21-year-old university graduate leading a team communicate with a 65-year-old employee in the future?
- **Agent 5:** I understand your concerns about the challenges posed by age differences. In reality, collaboration between multiple generations in the workplace can facilitate the exchange of experience and new ideas, enhancing the team's innovative capabilities. We can improve intergenerational understanding and cooperation by strengthening communication skills training.

Agent 2:

- **User (type):** I also think it's unfair.
- **Agent 2:** Yes, I'm glad you understand our perspective. The implementation of delayed retirement policies should take into full account the actual conditions of different industries and workers, with special attention and support directed towards physical laborers.
- **User (type):** Could it be that the current system is still imperfect?
- **Agent 2:** It is possible that the existing system is not yet perfect. Policymakers could further refine the policies to accommodate the needs of different industries and workers, thereby benefiting a broader spectrum of people.

B.2 P18: System-guided Followed by Interest-driven Conversation Order

Agent 1:

- **User (choose):** Are there international examples of setting different retirement ages for different industries that we can learn from?
- **Agent 1:** Yes, some countries have implemented retirement policies with different retirement ages for different industries, allowing physical laborers, for instance, to retire earlier. This approach is more humanized and takes into account the actual conditions of employees in different industries. I believe our country should also look into these examples for reference.
- **User (type):** How is physical labor defined, as some jobs might fall between the two categories?
- **Agent 1:** Indeed, defining the boundary between physical and non-physical labor can sometimes be challenging. We can involve relevant experts and experienced individuals in setting these standards, and also consider the specific characteristics and job descriptions of different industries to ensure that the policy is more targeted.
- **User (choose):** How would you determine the retirement age for positions that involve both physical and non-physical labor?
- **Agent 1:** For such positions, we can refer to existing job classifications and specific job content to assess the proportion of physical and non-physical labor involved. In making actual determinations, factors like the employee's age and years of service can be considered, leading to the formulation of more flexible retirement policies to ensure fair treatment for employees in different positions.
- **User (type):** How would you address individuals who exploit loopholes in these policies?
- **Agent 1:** To address those exploiting policy loopholes, it's crucial to strengthen the supervision and enforcement of these policies. Individuals or businesses attempting to manipulate these loopholes for personal gain should face stringent fines and penalties. Additionally, improving the policies by developing more comprehensive and detailed regulations can help eliminate potential loopholes.

Agent 2:

- **User (choose):** I agree with your viewpoint, but differentiated retirement standards may lead to certain complexities. What challenges might arise in implementing such a policy?
- **Agent 2:** Indeed, differentiated retirement standards do introduce complexities, such as determining evaluation criteria for different industries and setting specific retirement ages. However, I believe that by developing rational policies and detailed planning, we can find a fairer solution that accommodates the actual conditions of each individual.
- **User (choose):** How should the government balance the needs of different industries and people to formulate a more targeted differentiated retirement policy?
- **Agent 2:** The government can balance these needs by conducting thorough research to understand the actual conditions and demands of various industries, and by setting different retirement ages based on the nature of different jobs and individuals' health conditions. This ensures that the retirement policies are more in line with reality. Additionally, the government should strengthen support and benefits for physical laborers.
- **User (choose):** When implementing differentiated retirement policies, if dissatisfaction arises among people in certain industries or positions, how do you think the government should adjust and respond?
- **Agent 2:** The government needs to actively listen to the opinions and needs of people in different industries and positions. This can be done through communication and dialogue to minimize discontent. Timely adjustments to the policy may be necessary to balance the interests of all parties and build consensus.
- **User (choose):** How can the transparency and fairness of differentiated retirement policies be ensured during implementation, to prevent some people from feeling discriminated against?
- **Agent 2:** Ensuring the transparency and fairness of these policies lies in establishing detailed regulations, treating all industries and positions fairly, and seeking public opinions transparently. These methods can ensure that everyone's rights are respected and help prevent any form of discrimination.

Agent 3:

- **User (choose):** If employees choose to delay their retirement, will their pensions increase?
- **Agent 3:** Yes, typically, when employees opt for delayed retirement, their pension amounts are likely to increase. This is because the extension of their working period leads to longer contributions to the pension fund. This can be an additional incentive for those who feel they still have the capacity to work.
- **User (choose):** Then, for those who wish to retire early because they feel unable to continue until the standard retirement age, does the government have policies in place to assist them?
- **Agent 3:** For individuals looking to retire early, the government may offer certain adjustment options based on their specific situations, such as the possibility of accessing part

of their pension earlier. However, these specific policies and provisions would need to be adjusted according to the societal context and individual circumstances.

- **User (choose):** Do you think the current pension system is universally applicable, and why?
- **Agent 3:** The current pension system, to a certain degree, meets the retirement needs of most people. However, it may not be fully adaptable to some specific situations. There's a need to increase the system's flexibility and tailor it to address various practical issues, allowing more people to benefit from the social security system.
- **User (choose):** Regarding the current pension system, do you think we should consider introducing personal private pension plans to supplement it?
- **Agent 3:** Yes, introducing personal private pension plans could be a beneficial supplementary option. It would provide an additional layer of security for individuals' retirement. Of course, the design and implementation of such policies should take into account the country's economic conditions and the actual needs of the people.
- **User (choose):** From your perspective, how should personal private pension plans be designed to better meet everyone's needs?
- **Agent 3:** In designing personal private pension plans, it's important to consider reasonable investment returns and the individual's risk tolerance. Appropriate saving and investment options should be formulated to cater to the diverse retirement needs of different individuals.

Agent 4:

- **User (choose):** Have you considered changing careers before retirement as a better way to cope with delayed retirement?
- **Agent 4:** Indeed, changing careers is a viable option. In fact, during my entrepreneurial journey, I have tried transitioning from one industry to another. This has been beneficial for personal career development and in preparing for potential retirement challenges. Additionally, continuously learning new skills to adapt to the changing work environment can also be helpful.
- **User (choose):** In your entrepreneurial experience, what experiences have made you more convinced of changing careers to cope with delayed retirement?
- **Agent 4:** Throughout my entrepreneurial journey, I have worked in both technical development and marketing, two distinct fields. Trying out different professions helped me discover areas I prefer and excel in. These experiences have enriched my knowledge base, giving me confidence to face potential delayed retirement policies.
- **User (choose):** For those nearing retirement age, how do you think they should adjust their mindset and prepare for possible delayed retirement policies?
- **Agent 4:** For those nearing retirement age, it's important to first adjust their mindset and positively embrace potential delayed retirement policies. Secondly, maintaining enthusiasm for learning and working, and improving their skills to better cope with changes is crucial. Lastly, establishing

- a sound retirement plan, such as saving ample retirement funds, is key to dealing with uncertainties.
- **User (choose):** Which industries do you think are more suitable for people close to retirement age to switch to?
 - **Agent 4:** Consulting, training, and education are ideal industries for those close to retirement age considering a career change. These fields can fully utilize their experience and knowledge while generally offering lower work pressure.
 - **User (choose):** What are the pros and cons of the current social security system? Does delayed retirement affect young people's work and employment opportunities?
 - **Agent 4:** The current social security system provides certain protections in areas like pensions, healthcare, and unemployment, but there's room for improvement in sufficiency and coverage. Delayed retirement could indeed impact young people's job opportunities but can also alleviate pension pressures and labor shortages. It's important to strengthen the social security system from various aspects to reduce the pressure on young people.

Agent 3:

- **User (choose):** If employees opt for delayed retirement, will their pension benefits increase?
- **Agent 3:** Yes, if employees choose to delay retirement, their pension benefits may increase due to the extended period of contributions towards the pension fund, resulting in a larger accumulation of retirement insurance funds.
- **User (choose):** How do you believe the policy of delayed retirement should balance the interests between the younger generation and the middle-aged and older individuals approaching retirement?
- **Agent 3:** I believe that the younger generation should indeed bear a greater share of the responsibility for social security to ensure the sustainable development of society. Only when all age groups undertake their due responsibilities can we collectively address issues such as labor shortages. The younger generation can create more value for themselves and their families through self-improvement and investment.
- **User (choose):** What advice do you have for the younger generation on planning for their future retirement and making pension investments?
- **Agent 3:** I recommend that the younger generation begin planning for retirement early by enhancing their pension reserves through various channels, such as actively participating in both corporate and personal pension insurance and investing in financial products. Additionally, developing good consumption and financial management habits, along with continuous learning and career development, is essential to meet future challenges.
- **User (choose):** How should one decide between personal pension insurance and corporate pension insurance?
- **Agent 3:** When deciding between personal and corporate pension insurance, it is important to comprehensively consider your actual needs and financial situation. It is advisable to analyze the contract terms, coverage scope, and investment returns of both options to find a balance that meets your personal needs, risk tolerance, and expected returns.

- **User (choose):** How do companies that establish corporate pension insurance schemes handle this insurance when an employee leaves the company?
- **Agent 3:** When an employee leaves, companies typically follow relevant regulations and the terms of the corporate pension insurance contract to process. The portion of the pension that the employee has contributed to may be transferred to their personal pension insurance account to safeguard their pension rights after departure.

B.3 P6: System-guided Conversation Order

Agent 1:

- **User (type):** Do deliver-men also receive a pension?
- **Agent 1:** Yes, as physical laborers like ourselves, although our pensions might be relatively lower, we too have the right to enjoy retirement. Our generation has contributed to society's development and deserves equal treatment.
- **User (choose):** What age do you think is appropriate for physical laborers to retire?
- **Agent 1:** I believe it is appropriate for physical laborers to retire around the age of 60. This ensures that they still have the physical and mental capacity to complete their work before retirement, while also allowing them to enjoy a longer period of retirement and take good care of themselves.
- **User (choose):** Do you think delayed retirement is more suitable for non-physical workers?
- **Agent 1:** For non-physical workers, delayed retirement is indeed more appropriate, as their work relies more on knowledge and experience, which can actually enrich with age. However, for each individual, we need to consider their physical health and job requirements to make a decision.
- **User (type):** What do you think is the rationale behind delayed retirement?
- **Agent 1:** I believe the rationale behind delayed retirement is to address issues arising from an aging society, such as a decrease in population and insufficient pension funds. Delaying retirement can help reduce the pressure on pension systems and meet the long-term developmental needs of the country.

Agent 2:

- **User (type):** How much pension do cleaners receive after retirement? Is it enough to sustain their living?
- **Agent 2:** The amount of pension depends on the individual's working duration and the base of contributions. Generally, it should be sufficient to maintain basic living standards. However, as a physical laborer, my primary concern before retirement is whether I can maintain my health until the age of 65.
- **User (choose):** In your view, should physical laborers consider retiring earlier?
- **Agent 2:** I do believe that physical laborers should consider retiring early due to the significant stress and fatigue our work places on our physical health. Early retirement would not only safeguard our health but also create more job opportunities for the younger generation.

- **User (choose):** Have you thought about how to deal with potential financial pressures after early retirement?
- **Agent 2:** Yes, when considering the possibility of early retirement, I am indeed worried about financial pressures. I plan to focus more on savings and investments to adequately prepare for retirement. Additionally, I intend to seek part-time work involving light physical or non-physical labor to ease the financial burden.
- **User (type):** What specific part-time jobs would you consider?
- **Agent 2:** I would consider part-time jobs that align with my work experience, such as community sanitation or home cleaning. Additionally, options like handicraft making or opening a small shop, which require a certain skill set, are also part of my considerations for part-time work.
- **User (type):** Isn't that similar to delayed retirement?
- **Agent 2:** While it bears some similarity to delayed retirement, choosing early retirement allows me to gradually transition from the strenuous work I've experienced. Part-time jobs are relatively easier, allowing me to maintain an income source while having more time to rest and focus on my health.
- **User (choose):** How do you think the government's policy on delayed retirement should be improved to ensure physical laborers can maintain a comfortable living standard upon retirement?
- **Agent 2:** I think the government could formulate more flexible retirement age policies tailored to different professions and industries, allowing physical laborers to retire a bit earlier. Additionally, increasing the basic pension for physical laborers could ensure a comfortable living standard post-retirement.
- **User (type):** What is the rationale behind the delayed retirement policy?
- **Agent 2:** The primary rationale behind the delayed retirement policy is to address the challenges of an aging population. By raising the retirement age, the policy aims to reduce the burden on pension funds and make full use of the labor resources of the elderly. However, I believe this one-size-fits-all policy fails to consider the unique characteristics and needs of different industries and professions, particularly those of physical laborers.

Agent 3:

- **User (choose):** If employees choose to delay their retirement, will their pensions increase?
- **Agent 3:** Yes, if employees choose to delay their retirement, their pensions will likely increase due to the extended period of social security contributions. The amount of pension is influenced by the length of these contributions, although this can vary based on specific circumstances.
- **User (choose):** As a physical laborer, do you believe that professionals in non-physical jobs have more advantages under the delayed retirement policy? If so, could you share your perspective?
- **Agent 3:** Indeed, I believe professionals engaged in non-physical jobs have more advantages under the delayed retirement policy. Their work nature is relatively easier and causes

- less physical strain, making it easier for them to maintain good health and work longer.
- **User (choose):** For those in less physically demanding jobs, do you think they can make a greater contribution to society if they delay retirement?
 - **Agent 3:** I do think that individuals in less physically demanding jobs can make a greater contribution to society if they delay retirement. By extending their working years, they can leverage their vast experience and specialized knowledge to play a more significant role in their professional fields.
 - **User (type):** Do you think delayed retirement reduces job opportunities for young people?
 - **Agent 3:** The delayed retirement policy might have some impact on the employment opportunities for young people, especially in a competitive job market. It could increase the pressure on young people looking for jobs.
 - **User (choose):** If a flexible retirement policy is implemented, what impact do you think it would have on the employment of young people?
 - **Agent 3:** Implementing a flexible retirement policy would have both positive and negative impacts on the employment of young people. On the positive side, early retirement of physical laborers could create more job opportunities for the younger generation. On the negative side, a flexible retirement policy might intensify the competition young people face when seeking jobs. Overall, however, it could help create a more equitable job market.

Agent 4:

- **User (choose):** As you mentioned, early retirement may release certain job opportunities, but a flexible retirement policy might also lead some who wish to retire early to continue working. What is your view on this situation?
- **Agent 4:** I see this situation as having two sides. On one hand, older individuals who continue to work can contribute to society with their rich experience and skills. On the other hand, this might impact the employment opportunities for young people to some extent. A flexible retirement policy should strike a balance, fully utilizing the comprehensive value of older workers while ensuring fair job opportunities for the younger generation.
- **User (choose):** What policies do you think the government could implement to help people adapt to delayed retirement?
- **Agent 4:** I believe the government could consider measures such as providing vocational training, fully utilizing the functions of workers' social insurance, implementing flexible employment policies, and establishing a comprehensive pension service system. These measures would assist people in better adapting to the delayed retirement policy.
- **User (choose):** Could you elaborate on the role of workers' social insurance in helping people adapt to delayed retirement?
- **Agent 4:** In aiding adaptation to delayed retirement, workers' social insurance can play a role by increasing pension subsidy standards, raising the payment ratio for pension insurance, and adjusting subsidy policies. These measures

- could alleviate the economic pressures brought by delayed retirement and strengthen confidence in the policy.
- **User (choose):** Are these adjustments easy to implement in practice? Could they put pressure on the national finances?
 - **Agent 4:** Implementing these adjustments does pose certain challenges and needs to be within the country's fiscal capacity. However, with the right guidance and resource allocation, and by formulating targeted policies, it is possible to achieve them.
 - **User (choose):** Which adjustments do you support to achieve a more equitable delayed retirement policy? Please share your thoughts.
 - **Agent 4:** The adjustments I support include developing flexible retirement age policies tailored to different professions and industries, increasing the basic pension for physical laborers, and providing vocational training and transition opportunities for retirees. These measures would help balance the interests of different generations more equitably under the delayed retirement policy.

Agent 5:

- **User (type):** How does this policy inject vitality into social development?
- **Agent 5:** I believe that the delayed retirement policy can effectively utilize the labor resources of the elderly. It can alleviate the challenges of an aging population to a certain extent, reduce the burden of pension payments, and thus inject vitality into social development.
- **User (type):** But delaying retirement is just postponing the disbursement of pensions; it doesn't mean these people can't work.
- **Agent 5:** You are correct. Delaying retirement indeed only postpones the disbursement of pensions and does not mean these individuals cannot work. This also implies that people can consider transitioning or taking part-time jobs within the retirement age to maintain their health and living standards.
- **User (type):** So, do you think this policy fails to fully utilize the labor resources of the elderly?
- **Agent 5:** I believe this policy does utilize the labor resources of the elderly to a certain extent. Raising the retirement age allows older individuals to continue contributing to society with their rich experience and skills. However, the policy certainly needs improvement, especially in addressing the needs of physical laborers.
- **User (choose):** How do you think this policy should be improved to better cater to the needs of physical laborers?
- **Agent 5:** To better cater to the needs of physical laborers, the government should implement flexible retirement age policies tailored to different professions and industries, especially allowing for early retirement options for physical laborers. Additionally, increasing the basic pension for physical laborers is crucial to ensure their quality of life post-retirement.
- **User (choose):** Besides retirement age policies, what other aspects do you think can be improved to better cater to the needs of physical laborers?

- **Agent 5:** Beyond retirement age policies, improvements can include providing vocational training and transition opportunities, improving the working environment and conditions for laborers, and promoting health and wellness programs. These measures would offer support and benefits for physical laborers.

C EXAMPLES OF ASSESSMENT TASKS

Following are examples of the assessment tasks in the Viewpoints Puzzle, one for each AI agent. The tasks were generated and displayed in Chinese, and were subsequently translated.

Agent 1. What does Wu Xiaofei believe is the rationale behind the delayed retirement policy?

- (A) (Wrong) It is for the sake of the country's economic development.
- (B) (Wrong) It aims to enhance the standard of living for its citizens.
- (C) (Correct) It seeks to address the issues of a declining population and insufficient pension funds.

Agent 2. What are Wang Yanli's primary concerns regarding the delayed retirement policy?

- (A) (Wrong) She is worried about not having enough salary.
- (B) (Correct) She fears her physical strength will not sustain her until the delayed retirement age.
- (C) (Wrong) She is concerned about not having sufficient savings for retirement.

Agent 3. What kind of retirement policy does Zhang Xiaoning believe is more appropriate?

- (A) (Wrong) A one-size-fits-all delayed retirement policy.
- (B) (Wrong) Retirement at a uniform age as stipulated by the state.
- (C) (Correct) A flexible retirement policy, where employees can choose their retirement time based on their physical condition and retirement preparations.

Agent 4. What kind of system does Li Zehan hope the government will introduce to help people adapt to delayed retirement?

- (A) (Wrong) Increase pensions.
- (B) (Correct) Provide psychological support and technical training.
- (C) (Wrong) Reduce working hours.

Agent 5. What benefits does Professor Zhang Hua believe that the delayed retirement policy will bring to societal development?

- (A) (Wrong) Improve the quality of life for individuals.
- (B) (Wrong) Increase government fiscal revenue.
- (C) (Correct) Inject more vitality into social development.