AMQuestioner: Training Critical Thinking with Question-Driven Interactive Argument Maps in Online Discussion

QIYU PAN[*], JIANQIAO ZENG[*], and JIE WANG[*], Sun Yat-sen University, China
JUNYU LIU, YIHAN QIU[†][†], Sun Yat-sen University, China
KANGYU YUAN, Hong Kong University of Science and Technology, China
ZHENHUI PENG[‡], Sun Yat-sen University, China

Critical thinking, which requires logical analyses on the problems and keeping open-minded to others' viewpoints, is a crucial skill when participating in online discussions. While existing works have explored visualizing the components of an argument in a map, i.e., argument map, to support critical thinking tasks, few of them have incorporated educational elements that aim at training critical thinking in online discussion. In this paper, based on a formative study (N = 57), we develop *AMQuestioner*, a critical thinking training tool that allows question-driven interactions with argument maps automatically extracted from a post thread. In *AMQuestioner*, users can explore others' claims with a chatbot via suggested questions and conduct critical thinking exercises by answering generated questions related to any claim in the map. A mixed-design study (N=24) reveals that, compared to a baseline tool without question-driven features, participants after training with *AMQuestioner* demonstrated significantly more improvements in independently writing arguments that are detailed, specific, and relevant to the topic. Participants with *AMQuestioner* also exhibited a stronger inclination toward open-mindedness to others' arguments during the three-days training process. We discuss design implications for future critical thinking training tools.

CCS Concepts: • **Human-centered computing → Human computer interaction (HCI)**; *Collaborative and social computing*.

Additional Key Words and Phrases: Critical Thinking, Online discussion, Argument Map, Question-driven interactions, Large language model

ACM Reference Format:
Qiyu Pan, Jianqiao Zeng, Jie Wang, Junyu Liu, Yihan Qiu, Kangyu Yuan, and Zhenhui Peng. 2025. AMQuestioner: Training Critical Thinking with Question-Driven Interactive Argument Maps in Online Discussion. *Proc. ACM Hum.-Comput. Interact.* 9, 7, Article CSCW370 (November 2025), 48 pages. https://doi.org/10.1145/3757551

[*] Both authors contributed equally to this research.
[†] Both authors contributed equally as co-second author.
[‡] Corresponding author.

Authors' Contact Information: Qiyu Pan, panqy27@mail2.sysu.edu.cn; Jianqiao Zeng, zengjq28@mail2.sysu.edu.cn; Jie Wang, wangj928@mail2.sysu.edu.cn, Sun Yat-sen University, Zhuhai, Guangdong, China; Junyu Liu, Yihan Qiu, liujy525@mail2.sysu.edu.cn,yihanqiu@mail2.sysu.edu.cn, Sun Yat-sen University, Zhuhai, Guangdong, China; Kangyu Yuan, Hong Kong University of Science and Technology, Hong Kong, China, kyuanaf@connect.ust.hk; Zhenhui Peng, Sun Yat-sen University, Zhuhai, Guangdong, China, pengzhh29@mail.sysu.edu.cn.

ACM 2573-0142/2025/11-ARTCSCW370
https://doi.org/10.1145/3757551

1 Introduction

Critical thinking is crucial skill for university students [3, 18, 21, 31, 70, 76]. Two key components of critical thinking are logic and open-mindedness [43]. Logic is the ability to analyze problems and evaluate arguments [26, 28, 34], while open-mindedness means being willing to consider new ideas and different viewpoints [19, 27, 30]. This paper focuses on critical thinking in online discussion, *e.g.,* conversations on politic events or social issues via comments under news articles or posts in forums, which is an informal learning context outside classrooms. In this context, logic requires a critical thinker express reasoned opinions, critically assess others' arguments, and distinguish between misinformation and facts online [22, 26, 28]. At the same time, open-mindedness expects them to keep open to diverse perspectives, understand that one's viewpoint may differ from others, and recognize the validity of different arguments [19, 27, 30].

Given such requirements, conducting critical thinking is often challenging for many people. Existing work in Human-Computer Interaction (HCI) has explored various ways to supporting users in critical thinking tasks, *e.g.,* providing adaptive feedback in writing an argumentative essay writing [32, 78], prompting critical thinking questions during paper reading [60], and building social chatbots to debate or discuss certain topics with the users [67]. This line of work focuses on providing users with in-situ and adaptive assistance, aiming to improve their performance with external assistance in the current critical thinking task session. However, users would gain more benefits if they could get trained with critical thinking skills in the current task session, such that they can perform well in critical thinking without external support in the future.

To achieve such a goal, educational elements like asking questions and providing feedback should be incorporated into the tools, as do in previous tutoring systems for answering children's why and how questions [45], learning programming concepts [81], learning factual knowledge [64], and so on. Argument mapping is one educational tool used in classrooms for training students' critical thinking skills [35, 39, 72–74]. It uses nodes and connecting lines to clearly illustrate the different elements within an argument, such as claims, premise *etc.*, thus aiding in the comprehension and analysis of complex arguments [17, 71]. Outside the formal learning contexts, *e.g.,* in classrooms, existing work or online platforms has started to visualize the arguments or online discussion with nodes and edges like those in an argument map (*e.g.,* Kialo [1], VISAR [89]). Nevertheless, these platforms or online work mainly aim at improving user performance in the current argumentative writing task with AI support, instead of training their skills for independent critical thinking. In all, there is a lack of understandings on how to train critical thinking with argument map, adaptive AI support, and user-generated content in online discussions. Addressing this gap can provide insights into leveraging rich online content as informal learning resources and designing critical thinking training tools.

To fill this gap, we first conducted a formative study with 57 respondents to understand their challenges and needs for support when conducting critical thinking in online discussion. The results revealed that users often struggle to grasp the core ideas and structure of lengthy online discussions, and they desire more structured guidance and interactive support to analyze arguments and develop their critical thinking skills. Building upon these findings, we propose *AMQuestioner*, a technical prototype that automatically builds an argument map based on comments under each post thread and provides generative questions and feedback to train critical thinking. In *AMQuestioner*, users can check diverse perspectives on the discussed topic by checking the claims, premises, and counter-arguments in an argument map. Users can select any specific claim on the argument map, and *AMQuestioner* will prompt a large language model (LLM) to generate guiding questions to help them think critically about that claim. *AMQuestioner* also offers generated exercises while

[1] https://www.kialo-edu.com

users are exploring the argument maps. These exercises are designed to guide users in analyzing each viewpoint and their arguments in depth, identifying potential logical fallacies or insufficient evidence. For example, users might be asked to identify the main claim of a specific argument, evaluate the strength of the supporting evidence, or propose counterarguments. *AMQuestioner* prompts a LLM to provide feedback on user responses, highlighting areas for improvement and reinforcing correct application of critical thinking principles.

We conducted a mixed-design study (time as the within-subject factor, tool as the between-subject factor) with 24 undergraduate students to evaluate the effectiveness of *AMQuestioner* in fostering critical thinking skills in online discussions. We compared the performance of participants using *AMQuestioner* to a baseline group that offers the constructed argument map without generated questions and feedback. Our results indicate that participants using AMQuestioner demonstrated significant improvements in their ability to analyze and evaluate arguments, specifically in terms of persuasiveness, specificity-justification, and relevance. Additionally, AMQuestioner users showed enhanced open-mindedness and expressed a stronger willingness to engage in critical thinking during the three-days training process. These findings suggest that question-driven interactions could transform how users engage with argument analysis, moving beyond passive consumption to active critical evaluation. By combining AI-powered Socratic questioning with visual argument maps, we show that technology can scaffold the development of transferable critical thinking skills rather than merely supporting task completion. This finding points toward a new paradigm for leveraging social media content as authentic materials for training critical thinking skills in an era of widespread misinformation.

In summary, this work makes three contributions to the CSCW community. First, we design and develop *AMQuestioner*, a novel critical thinking training tool powered by comments in online discussion and large language models. Second, our mixed-design user study demonstrates the effectiveness and usefulness of *AMQuestioner* compared with a baseline tool. Third, we offer a set of design considerations that guide the future design of the interactive system training critical thinking skills with community data.

2 Related Work

2.1 Argument Map for Critical Thinking

Critical thinking can encompass three core aspects: critical analytical ability (evaluating information and logical arguments), open-mindedness (considering multiple perspectives), and critical thinking disposition (motivation to engage in critical thinking) [10]. These critical thinking skills can help people process complex information and form their own thoughts in daily life, and promoting these skills has been an essential goal of higher education [52, 56, 69]. One effective tool for promoting critical thinking is argument map (AM) [73, 74], which represents arguments as directed graphs with nodes (propositions) and edges (relationships). For example, van Gelder [73] conducted a meta-analysis that includes 26 pre- and post studies of AM-based instruction in a one-semester critical thinking subject. In fifteen of these studies, students took a subject in which AM was the primary or central activity, with lots of homework activities, and with instructors with high proficiency in AM. The analysis suggested that such AM-based instruction could improve students' learning gain in critical thinking [73]. Despite its benefits, the traditional AM-based instruction usually require extensive efforts from human instructors to prepare the learning materials, monitor students' progress, and provide feedback.

Prior researchers have developed various intelligent systems to ease the human instructors' efforts in AM-based instruction and explore the usage of AM in supporting critical thinking tasks. For example, Butchart et al. [9] developed a system that provides automatic, real-time feedback on students' progress as they construct their maps of an argument. They showed an improvement in students' critical thinking in a single semester undergraduate critical thinking course [9]. However, the system's feedback is limited to a binary "correct/incorrect" response without detailed explanations, and it struggles to handle complex, multi-layered arguments [9]. With models trained for argument component and relation identification, Wambsganss et al. [79] developed AL, an adaptive tutoring tool that provides students with feedback on the visual argumentation structure of a given text. They found that students using AL wrote more convincing texts with better formal quality of argumentation compared to the ones using the traditional approach without automatic feedback. Similarly, Zhang et al. [89] developed VISAR to support users in argumentative writing tasks. Powered by large language models, VISAR helps writers brainstorm and revise hierarchical goals within their writing context, organize argument structures through synchronized text editing and visual programming, and enhance persuasiveness with argumentation spark recommendations [89]. However, either AL [79] or VISAR [89] aim at improving user performance in the current argumentative writing task with AI support, instead of training learners' critical thinking skills so that they can independently perform well after the training sessions. Besides, the work mentioned above did not make use of the rich collective arguments in online discussions, which represent diverse human viewpoints and could be valuable for AM-based critical thinking training or writing tasks. It remains unknown how to train critical thinking with argument map, adaptive support based on AI technology, and user-generated content in online discussion.

Our work is motivated by the benefits of argument maps for training critical thinking. Different from previous work that explores argument maps for supporting critical thinking tasks, we investigate the integration of the map, the AI-enabled educational elements (*e.g.,* adaptive questioning and feedback), and rich online discussion data to cultivate users' critical thinking skills in the training sessions. Our goal is to improve users' critical thinking performance without external support after the training sessions.

2.2 Question-Based Interactions with Tutoring Systems

The rapid advancements in large language models (LLMs), such as GPT-4, have opened new possibilities for question-based interactions by enabling systems to generate, respond to, and adapt questions dynamically. Building on these breakthroughs, question-based interactions now play a crucial role in improving user engagement, fostering critical thinking, and enabling adaptive learning support [1, 8, 40]. In fields of learning and education, the question-based interactions with tutoring systems can be divided into two categories, *i.e.,* user-to-system and system-to-user questioning.

The user-to-system questioning involves users seeking information by asking questions or inputting search queries to the systems. This interaction design is prevalent in chatbots customized in specific domains [38, 42, 88] and general-purpose chatbots powered by large language models like GPT-4. Learners can seek answers from these chatbots to address their confusions whenever needed. HCI researchers also explore various ways to improve the user-to-system questioning in tutoring systems [12, 45, 83]. For example, Yang et al. [83] presented a question-answer pipeline AQuA, which generates useful responses to questions made in software tutorial videos. Their pipeline can recognize UI elements in visual anchors and generate answers using GPT-4 augmented with that visual information and software documentation [83]. Their evaluation study with 16 users demonstrated that their pipeline can produce more correct and helpful answers compared to baseline methods [83]. Lee et al. [45] designed DAPIE that transforms existing long-form answers into interactive dialogues to address children's why and how questions. The dialogues include system-to-user questions that guide children to explore the answer and diagnose their understandings [45]. Their study with 16 participants revealed that children when using DAPIE got a significantly higher score in an immediate assessment and showed a significantly higher level of engagement than when using the baseline system which directly presents long answer [45]. While the user-to-system questioning design can generally satisfy learners' information need, this work suggests that turning it into interactive system-to-user questioning design could lead to more engaging learning experience and higher learning gains.

Speaking to the system-to-user questioning design, it could be a more common practice in the tutoring systems [58, 64, 81, 84], which ask questions to test learners' understandings of some concepts and guide them to conduct in-depth analyses. For example, Winkler et al. [81] presented Sara, a conversational agent that asks learners multiple-choice questions and scaffolds learners' understanding with adaptive feedback during an online video lecture. Their lab experiment demonstrated that against more traditional conversational agents, Sara significantly improved learning in a programming task [81]. Peng et al. [58] designed DesignQuizzer, a community-powered conversational agent that prompts multiple-choice questions about the UI design example. They prepared the questions by masking the keywords about visual elements (*e.g.,* color, layout) in the critiques expressed in the comments under the UI design posts [58]. Their lab experiments also indicate the value of their system-to-user questioning interaction for improving learning gains, compared to a baseline condition without the conversational agent [58].

Moving from understandings to in-depth analyses, prompting Socratic-style questions has been proved effective in guiding critical thinking [13, 36, 54, 57]. The key to distinguish Socratic questioning from other types of questioning is that it is systematic, disciplined, and deep and usually focuses on foundational concepts, principles, theories, issues, or problems [57]. Recent works have successfully attempted to generate Socratic-style questions with large language models [23, 87]. For example, Zhang et al. [87] introduces SPL, a dialogue-based tutoring system powered by the GPT-4 model, which employs the Socratic teaching method to foster critical thinking among learners. Their pilot experimental results from essay writing tasks demonstrate SPL has the potential to improve tutoring interactions [87]. Other tutoring systems for critical thinking tasks have mixed usages of user-to-system and system-to-user questioning designs. For instance, Peng et al. [60] collected critical thinking questions that readers would ask when reading HCI papers, based on which they presented CReBot that prompts questions related to the reading paper section. Their team further developed CriTrainer [84], which provides templates that guide users to ask critical thinking questions about the academic paper they read. They demonstrated that compared to a baseline tool, CriTrainer enabled users to independently raise more relevant and critical questions about the paper after the training section [84].

Inspired by these previous systems, our *AMQuestioner* adopts an interactive user-to-system questioning approach for information seeking (*i.e.,* learn other members' opinions to cultivate open-mindedness on the topic in our case) and uses a system-to-user Socratic questions design for in-depth topic analyses (*i.e.,* arrive at the answer through the users' own reasoning). Different from previous systems, *AMQuestioner* grounds the generations of and answers to these questions on the rich online discussion data, and it serves back to training critical thinking in online discussions that previous tutoring systems seldom explore. We integrate these two categories of question-based interactions into argument map in *AMQuestioner*, contributing a novel tutoring system that trains critical thinking as users participate in online discussion.

2.3 Critical Thinking Support in Social Media

The massive and sometimes misleading content in social media requires users to conduct critical thinking when seeking information online. Previous HCI researchers have explored two types of approach to support this thinking process. The first type is to develop computational models that identifies the argumentative components of the social media content and offer an interactive way to explore these components. For instance, to support lurkers to join collective arguments in question-answering platforms, Liu et al. [47] labeled data in a question-answering community and developed a pipeline that could extract and organize argumentative information. Based on the pipeline, they presented CoArgue, which navigates users through the claims with highlighted text and guides users to develop their answer posts with a chatbot. Similarly, Xia et al. [82] built a labeled dataset of fine-grained persuasive strategies (*i.e.,* logos, pathos, ethos, and evidence) in 164 arguments in the ChangeMyView community. They then introduced Persua, which provides example-based guidance on persuasive strategies to help users enhance the persuasiveness of their arguments. The other type of approach is to simulate human members to engage users in critical thinking via conversation. For example, to help users burst their filter bubbles, Zhang et al. [88] implements an LLM-powered multi-agent system in which users can engage with opposing viewpoints via different characters. Tanprasert et al. [67] investigated the effect of two relevant persona attributes - social identity and rhetorical styles - of a debate chatbot on facilitating critical thinking on YouTube. Apart from efforts from the research community, some social media platforms have actually embedded argument mapping into their web interfaces to support critical thinking. For example, Kialo [2] is an online community featured in providing a structured environment for users to build, analyze, and debate arguments. It provides a collaborative argument mapping component in which every member can add their views and link them to the others' in the map.

Overall, the design and development of *AMQuestioner* are inspired by these critical thinking support approaches in social media. However, different from their focus on supporting users to conduct critical thinking, *AMQuestioner* aims at cultivating users' critical thinking, such that they could independently perform critical thinking after the training sessions. Without the need from collaborative human efforts to construct the argument map as does in the Kialo community, we develop computational models to automatically build up the map based on the discussion data. As our focus is on training critical thinking, we design and implement the questioning-based educational elements that are integrated with the argument map. We conduct a controlled user study to compare *AMQuestioner*'s effectiveness to a baseline tool with the constructed argument maps but without the educational elements, following the settings in previous critical thinking support tools.

3 Formative Study

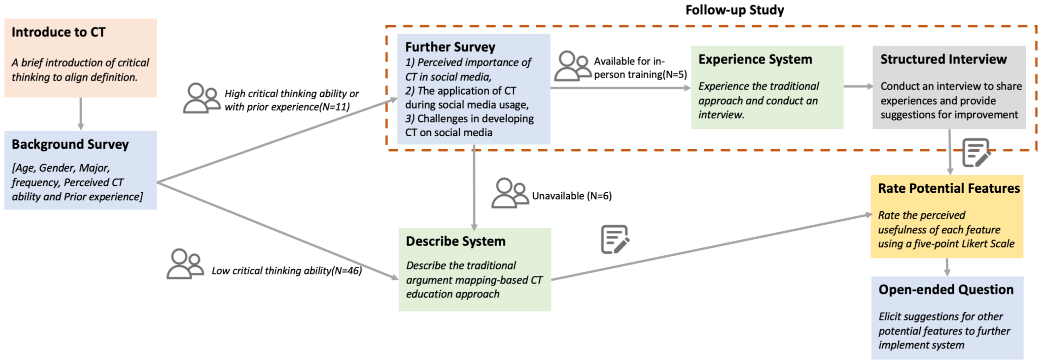

Follow-up Study
Introduce to CT
A brief introduction of critical thinking to align definition.
Background Survey
[Age, Gender, Major, frequency, Perceived CT ability and Prior experience]
High critical thinking ability or with prior experience(N=11)
Low critical thinking ability(N=46)
Further Survey
1) Perceived importance of CT in social media,
2) The application of CT during social media usage,
3) Challenges in developing CT on social media
Available for in-person training(N=5)
Unavailable (N=6)
Experience System
Experience the traditional approach and conduct an interview.
Structured Interview
Conduct an interview to share experiences and provide suggestions for improvement
Describe System
Describe the traditional argument mapping-based CT education approach
Rate Potential Features
Rate the perceived usefulness of each feature using a five-point Likert Scale
Open-ended Question
Elicit suggestions for other potential features to further implement system

Fig. 1. Overview of the formative study process

[2] https://www.kialo.com/

To guide the design of formative study, we first structure a Critical Thinking training process based on the definition of critical thinking and existing literature. With this structured process, we then conduct a survey study with 57 participants and a follow-up study with five of them to identify user challenges and needs for support in the training process. Figure 1 shows the process of the formative study.

3.1 *AMQuestioner* Critical Thinking Training Process

As reviewed in the Related Work (Section 2.1), critical analytical ability, open-mindedness and disposition toward critical thinking are several crucial components of critical thinking. Therefore, we focus on training these skills in this work. We adapt the "Pre-reading, During-reading, and Post-Reading" Strategies [5, 66], which is a common framework in text-level critical thinking instruction [2]. Building upon existing frameworks, we propose a argument map-based critical thinking training program, integrated with users' daily reading scenarios (e.g., social media platforms) and powered by adaptive LLM support. Specifically, the process consists of three stages:

(1) **Pre-reading Stage:** Users read the *Argument Map user Manual* in Appendix B, select a topic and, based on a provided summary, write an initial comment.
(2) **During-reading Stage:** Users engage in critical thinking training with argument map during this stage.
(3) **Post-reading Stage:** Users refine or reconstruct their initial comment based on their enriched understanding.

The *AMQuestioner*'s intervention in this training process focuses on the during-reading stage, as the pre- and post-reading stages are more for users to complete their writing tasks on their own.

3.2 Process of Formative Study

Figure 1 describes the process of formative study. We distributed a questionnaire to each respondent. It first briefly introduced the definition of critical thinking (*i.e.,* "Critical thinking involves three core aspects: analyzing information and arguments, open-mindedness to different perspectives, and the motivation to think critically. These skills are crucial for processing complex information and forming independent thoughts, making their promotion a vital goal in higher education.") We then collected participants' background information through a survey, including their age, gender, major, social media usage frequency (never, low, medium, high), perceived critical thinking ability ("Considering the definition of critical thinking, how would you rate your critical thinking skills?": weak, below average, average, above average, strong), and prior experience in applying critical thinking skills on social media ("Do you have experience applying critical thinking skills in social media environments?": Yes, No).

For the respondents who self-rated above-average or strong critical thinking ability or prior experience in conducting critical thinking online, we asked they to complete a survey asking three questions. They are: 1) How essential do you consider critical thinking to be in social media; 2) In your daily usage of social media, how do you typically apply critical thinking to engage with and evaluate the information and interactions you encounter; and 3) Based on your experience, what are the primary challenges to effectively and consistently applying critical thinking on social media. We invited these respondents to participate in a follow-up study. The respondents (N=5 in our case) who accepted our invitation were guided through a traditional argument map-based critical thinking training process. Following the training, semi-structured interviews were conducted to solicit feedback for system improvements. The traditional argument map-based critical thinking training involved the following steps: **Pre-Reading:** Participants read the *Argument Map User Manual* (Appendix B) and write a initial comment about the given post. **During-Reading:** Participants then constructed an argument map from scratch based on a provided ChangeMyView post in Reddit. **Post-Reading:** participants refine their initial comments or reconstruct comments. **Mentorship:** Throughout this process, a mentor provided real-time guidance and support. Based on the trials within our research team, the survey and the follow-up study were estimated to last one and half hour, and these five respondents received a compensation of 50 RMB.

For those who are unavailable to join the follow-up study and those who self-report low critical thinking ability or no prior experience in conducting critical thinking online, they go ahead and read descriptions of both traditional and modified argument mapping approaches. Finally, following the practice in Yuan et al. [84], all respondents were asked to rate the perceived usefulness of potential features (Table 1) derived from VISAR [89], Persua [82], constructivism theory [44], and scaffolding concepts [41, 77] on a five-point Likert scale (1 = not useful at all, 5 = very useful). The study concluded with open-ended questions asking for other potential features of *AMQuestioner*. The respondents who did not participate in the follow-up study received a compensation of 10 RMB.

3.2.1 Respondents. We recruited 57 students (S1-57; 23 female, 25 male, 9 preferring not to specify gender) through social networks and word-of-mouth. All participants speak English as their second language. The sample included 33 undergraduate students, 18 master's students, and 6 doctoral candidates. Participants' ages ranged from 18 to 29 years old, with a mean age of 22 (SD = 2.27). All participants showed a habit of using social media, with 31 reporting high frequency of using social media platforms, 22 reporting medium frequency and 4 reporting low frequency. Twelve respondents reported having weak or below average critical thinking skills, 31 reported average critical thinking skills, and the remaining 14 considered themselves to have above average or strong critical thinking abilities. 11 participants (among the 14 participants who considered themselves to have above average or strong critical thinking abilities) reported prior experience applying critical thinking skills in the context of social media.

3.3 Findings

For open-ended responses in the questionnaires (*e.g.*, challenges and suggested features) and interviews, two authors conducted inductive coding. They first independently created and assigned codes for the each response. Then, they met, discussed their codes, removed and merged some of them, and named the merged codes. We summarize the findings below.

3.3.1 Challenges of Employing Critical Thinking on Social media. **C1: Social media commentary is often problematic.** Six participants highlighted several challenges related to social media comments, including information overload due to high volume, a preponderance of shallow and unconstructive comments, unstructured and informal language, and difficulty discerning factual accuracy. "*The quality of comments on social media is highly variable, with a majority being low-quality, containing flawed arguments, and expressed unclearly. I often struggle to understand the intended claims and supporting evidence*" (U7, Male, 24). Another participant noted, "*When I use platforms like Reddit, I usually only read the top few upvoted comments because the sheer volume is overwhelming*" (U2, Female, 23). Besides, one participant observed, "*I've noticed that posts documenting daily life tend to elicit comments focused on personal experiences and feelings, while platforms like Reddit and Zhihu, particularly on controversial topics, often generate more in-depth and engaging discussions*" (U10, Female, 25). This highlights the significant differences in discussion quality not only between platforms but also within the same platform depending on the type of post. **C2: Users reported difficulties in conducting critical thinking in social media.** In the background survey, only 14 out of 57 respondents considered themselves to have above average or strong critical thinking abilities. In the formative study, three out of five participants reported lacking the critical thinking skills necessary to analyze social media content. They cited difficulties such as identifying bias in comments, selectively accepting information confirming pre-existing beliefs, and passively consuming information without seeking alternative perspectives. "*I see so many posts and comments, but I don't always stop to think about why someone might be saying those things or if it's even true. It's easy to just scroll on.*", "*I realize I tend to follow people who agree with me, so I'm probably not getting the whole picture. It's hard to know what to believe sometimes*". (U4, Female, 21)

3.3.2 Potential features of AMQuestioner. Notably, among the features that met our threshold of 4.0 in Table 1, the higher standard deviations (ranging from 0.63 to 0.82) suggest more diverse opinions about the implementation of AI-assisted exploration. Our respondents actively indicated their expected features of our training tool in the open-ended questions. In the During-Reading stage, one participants suggested that "*I'd find it more useful if the system prompted me to articulate the reasons behind my answer choices. That would encourage more thoughtful engagement with the critical thinking exercises*" (U8, Female, 27). Similarly, three participants recommended highlighting specific sections of lengthy comments relevant to the LLM's statements. "*Sometimes the comments are so long it's hard to see the connection to what the LLM is saying. Highlighting the related part in the comment would make that clearer*" (U4, Female, 21). One participants expected that LLM-generated default response should delve deeper into specific aspects of the discussion and also explore other aspects under the discussion. "*It would be great if some default responses could help me explore the finer details of the conversation, while others guide me to consider different angles or perspectives*" (U3, Undefined, 24).

3.4 Design Requirements for *AMQuestioner*

Based on our findings and relevant educational literature, we derived the following design requirements (DRs) for fostering critical thinking in the context of social media platforms:

DR1: For argument map construction, the tool should preprocess comments and generate a modifiable argument map reflecting the different discussion threads under the post with distinct colors to differentiate argument components and relationships. The respondents' perceived that four potential features related to argument map construction (Table 1) would be useful for conducting critical thinking in online discussion, which can be supported by their reported challenges and related work. For example, the visualization of the hierarchical structure of discussion threads could address the challenge of information overload and unstructured content (C1). This structured visualization approach aligns with established research demonstrating that argument mapping facilitates the management of complex information [29, 51]. The color coding and modifiability of different argument components are prevalent features in user interface design, *e.g.,* in existing argument mapping tool like VISAR [89], which could help users easily identify, explore, and contribute structural elements and relationships within arguments. Besides, the process of modifying argument maps itself can constitute an active critical thinking exercise that aligns with our goal of skill training.

DR2: For viewpoint exploration, the tool should provide a tutor for guiding exploration of each topic and its associated comments under the given post. This requirement addresses users' lack of critical thinking skills (C2), particularly their difficulty in seeking alternative perspectives and analyzing content critically. The tutor can serve multiple critical functions in supporting reasoning development. For example, it can present core viewpoint content while providing navigational links between the argument map and original source comments, which prevents LLM "hallucinations" while enabling users to verify information authenticity — a fundamental aspect of critical evaluation. The suggested input options, which is prevalent in chat interface (*e.g.,* Bing Search) powered by LLMs, could reduce cognitive load during exploration while simultaneously

Table 1. Perceived usefulness of potential features of *AMQuestioner* with mean scores ≥ 4.0 (N=57). Respondents rated each feature on a 5-point Likert scale (1 = not useful at all, 5 = very useful).

Component	Potential Feature	Average	SD
Argument Map Construction	Visualize the hierarchical structure of discussion threads	4.35	0.62
	Distinguishing different argument components and relations using colors	4.12	0.71
	Ability to expand/collapse, modify the Argument Map (add, delete nodes, and modify node content)	4.42	0.58
	Filter out redundant and irrelevant comments	4.18	0.65
Explore Topics with Agent	The LLM actively guides users to explore the details of each topic	4.28	0.73
	Providing suggested input options for users to explore the topic	4.05	0.82
	Provide clickable links to navigate between LLM responses and referenced content in maps and comments	4.15	0.63
	Assess user comprehension of topics through evaluation methods	4.08	0.74
Critical Thinking Question	Designing critical thinking questions based on the comments under a given topic	4.32	0.67
	Providing various types of critical thinking questions (e.g., multiple-choice, short-answer)	4.15	0.72
	Guiding users to modify and improve the argument map	4.06	0.83
	Guide users to reflect on incorrect answers through Socratic questioning	4.22	0.68
	Locate the source comments related to the question	4.10	0.75

modeling proper questioning techniques, helping users develop their own critical inquiry skills through guided practice. Comprehension questions at the end of exploration sessions can provide immediate feedback on understanding and metacognitive assessment, facilitating the transition from assisted to independent critical thinking. As participants explicitly expressed needing assistance with exploring "finer details of the conversation" and "different angles or perspectives" (U3), this guided approach bridges the gap between users' current abilities and the complex skills required for effective critical analysis of diverse viewpoints in social media discussions.

DR3: For critical thinking exercise, the tool should provide customized critical thinking exercises (*e.g.*, multiple-choice question) based on the user's provided topic and associated comments. This requirement addresses users' difficulty in conducting critical thinking in social media (C2) and the need for structured learning approaches by implementing a comprehensive exercise system designed to target specific critical thinking competencies. The varied question types strategically develop distinct cognitive skills while keeping users engaged through diversity. By requiring users to articulate their reasoning alongside answers, the system shifts focus from correct answers to the quality of analytical thinking, promoting metacognition and deeper engagement with content. The dual-path feedback mechanism (Socratic questioning or direct explanation) can accommodate different learning preferences and scaffolds users' development at appropriate challenge levels, with the iterative process of answering questions and revising argument maps creating a reinforcing cycle of critical thinking development. These features directly responds to participants' requests for explanation-based multiple-choice questions (U8), providing a structured pathway for users to progressively strengthen their critical reasoning abilities through deliberate practice with immediate, constructive feedback.

4 A Computational Workflow for Modeling Comments into Argument Map

Our approach to training critical thinking in online discussion centers around interactive argument mapping. Building on our formative study findings, we developed a computational workflow that systematically transforms comment threads into structured argument maps. This automated process enables users to visualize and analyze the argument components and argument relations embedded within online discussions.

4.1 Dataset Construction for Argument Mining

4.1.1 Data source. We collect our source data from ChangeMyView[3], a popular subreddit with 3.7 million members up to August 2024. This platform encourages users to share views they consider potentially flawed and engage with diverse perspectives through reasoned argumentation. We leverage ChangeMyView's recently trending ("hot") posts (Range from 2024.8.18 to 2024.8.25) as our source data for data annotation for two primary reasons: (1) the platform's established reputation in facilitating structured argumentation, as evidenced by its extensive use in prior research on online deliberation [59, 79, 82]. (2) its contemporary topics that resonate with our annotators' cultural and social context, enabling more nuanced interpretation of implicit meanings and contextual references in the arguments. Furthermore, this high volume of discussion threads and delta-scored comments ensures a substantial and representative dataset for training and evaluation.

4.1.2 Data Scraping and Filtering. We initially used the Praw Python library[4] to scrape data from Reddit's "hot" listings. This involved daily scraping of the top 3 "hot" posts for a week; if a post duplicated an existing one in the dataset, the dataset was updated with the newer version. Initially, we collected a dataset of 983 comments from 14 CMV posts across 7 common ChangeMyView topics, including politics, climate change, drugs, freedom of speech, social justice and electoral processes. This data encompassed post titles, descriptions, links, and all associated comments, which were recursively collected to capture the entire discussion thread. To enhance the clarity and structure of argument maps, we implemented a semantic filtering process using the RoBERTa-base model [48] to identify and remove redundant comments while preserving the most comprehensive arguments. This filtering process effectively maintained the core arguments and diverse viewpoints while removing between 16 to 28 redundant comments per discussion (detailed filtering methodology is presented in Appendix C.1). After filtering, our final dataset contained 649 comments from these posts.

4.1.3 Data Labeling. The taxonomy for annotating the argumentative structure of Reddit comments draws inspiration from several prior works [14,65, 82]. Following discussions, we finalized our annotation scheme, detailed in Table 2, 3. To apply this scheme, three individuals collaboratively developed a detailed annotation guideline, utilizing Persua's coding manual [82] as a foundation to

[3] https://www.reddit.com/r/changemyview/
[4] https://praw.readthedocs.io/

create a framework suitable for our data. More concretely, each annotator independently labeled 5 comments, following a four-step process: (1) reading the comment and its immediate parent comment within the thread; (2) segmenting the comment into sentences based on individual interpretation of meaning; (3) labeling each sentence with its corresponding argument component (major claim, claim, premise); and (4) annotating the attitude (support, attack) between major claims and claims, and between claims and premises. Following this initial annotation phase, inter-annotator disagreements, such as differentiating between “claim" and “premise", were discussed and led to iterative refinement of the guideline. A detailed example of such disagreements and the subsequent resolution process is provided in Appendix C.2. This iterative process, involving five rounds of annotation and discussion, ensured a comprehensive and robust annotation scheme addressing ambiguities and improved inter-annotator agreement. The final annotation guideline is presented in Appendix C.3.

Based on this guideline, the annotators independently annotated the remaining comments, incorporating regular quality control checks (every 15 comments) to discuss and ensure consistency in sentence segmentation and annotation. High Cohen’s κ values were observed across all categories: Major claim (κ = 0.81), Claim (κ = 0.83), Premise (κ = 0.92), No Sense (κ = 0.91), and support (κ = 0.95), attack (κ = 0.91) indicating strong inter-rater reliability. Ultimately, 2192 sentences (274 comments) were labeled with 268 as Major claim, 426 as Claim, 1237 as Premise, and 261 as No Sense. Besides, 674 supports, 989 attacks were labeled.

4.2 Argument Mining

4.2.1 Mining Methods. To optimize argument component and relation mining, we employed the following three methods. The first method combines BERT for state-of-the-art feature extraction [62] with various machine learning models for training [82]. BERT’s superior ability to capture contextual nuances enhances the effectiveness of these ML models (Logistic Regression, Linear SVM, RBF SVM, Random Forest, Gaussian Naive Bayes, Nearest Neighbour, Adaboost Decision Tree). This integration provides a robust framework for argument mining. The second method leverages few-shot learning [7], employing the latest large language model like GPT-4o reducing the need for extensive labeled data [11]. Cabessa et al. [11] demonstrate that few-shot strategy is efficient and quickly adapts to the task of argument type classification, making it ideal for resource-limited scenarios. The third method involves instruction fine-tuning of large language models. Cabessa et al. [11] show that this approach, combined with carefully engineered structural features, achieves state-of-the-art results on our tasks.

4.2.2 Mining Process. We developed task-specific prompts for both few-shot and fine-tuned LLMs, iteratively refining them through a three-stage process. First, we created an initial prompt outline. This stage leveraged principles from OpenAI’s documentation [5] and popular prompt framework in it [6] to guide the prompt’s structure and phrasing. Next, each section of the prompt was carefully crafted to ensure clarity, relevance, and alignment with the task of argument component and relation mining. Finally, throughout the entire process, iterative refinement was employed, reviewing and refining the content following several key principals [46,63, 85] to ensure it achieved a high level of accuracy in argument mining. Through several iterations, the instructions and prompts were refined to optimize model performance. These final prompts are presented in Appendix C.4.

For fine-tuning, we first prepared and validated our labeled dataset according to Azure OpenAI’s requirements. We employed GPT-3.5-turbo as our base model and followed standard practices for model evaluation using a stratified split strategy (60/20/20 for training/validation/testing). The

[5] https://platform.openai.com/docs/guides/prompt-engineering
[6] https://platform.openai.com/docs/examples

Table 2. Taxonomy of Argument Component

Name	Definition	Example
Major Claim	The root node of the argumentation structure. It's the author's main standpoint or opinion on the topic.	This is a terrible idea
Claim	Claims are secondary conclusions or viewpoints that support the major claim.	It creates an incentive that the government wants you to die at 75 years old.
Premise	Premises are the underlying facts, evidence, or reasoning that support claims or the main claim.	Humans and governments work almost exclusively at their most basic on incentive.
Non-argument	Non-argument statements are sentences within an argument that do not clearly function as a major claim, claim, or premise.	So what incentive does this create?

Table 3. Taxonomy of Argument Relations between Argument Components

Name	Definition	Example
Support	Support refers to one argument providing evidence or reasoning that strengthens or bolsters the claim of another argument.	“Humans and governments work almost exclusively at their most basic on incentive." **support** “It creates an incentive that the government wants you to die at 75 years old"
Attack	Attack refers to one argument attempting to weaken or refute the validity or credibility of another argument.	“So, at least 40% of people do not have a desire to pass away" **Attack** “They are merely expressing their frustrations; their true desire is to pass away"

fine-tuning process, which took approximately 2 hours, yielded promising results. For Argument Component Detection, the model achieved F1 scores of 0.77, while in Argument Relation Detection, it demonstrated performance with F1 score of 0.76. The training convergence analysis showed optimal validation performance during the latter part of epoch 3. The complete technical details of the fine-tuning process are provided in Appendix C.5.

Three approaches to argument mining were systematically evaluated on the held-out test set (20% of the total 2,192 sentences): BERT-based model, few-shot LLM, and our instruction-fine-tuned LLM. For argument component detection, few-shot GPT-4 with 16 examples achieved the best performance with precision of 0.90, recall of 0.87, and F1-score of 0.88. In argument relation detection, few-shot GPT-4 with 6 examples demonstrated optimal results, reaching precision of 0.87, recall of 0.87, and F1-score of 0.87. Based on these superior results, we will utilize these two few-shot GPT-4 models with their respective optimal example settings to construct our argument mapping system. Comprehensive performance metrics including confidence intervals and statistical significance tests can be found in Tables 4 and 5. A detailed evaluation of all experimental settings is provided in Appendix C.6.

Table 4. Performance Evaluation of Three Approaches to Argument Component Extraction: BERT-based model, Few-Shot GPT-4o, and Instruction-Tuned Models

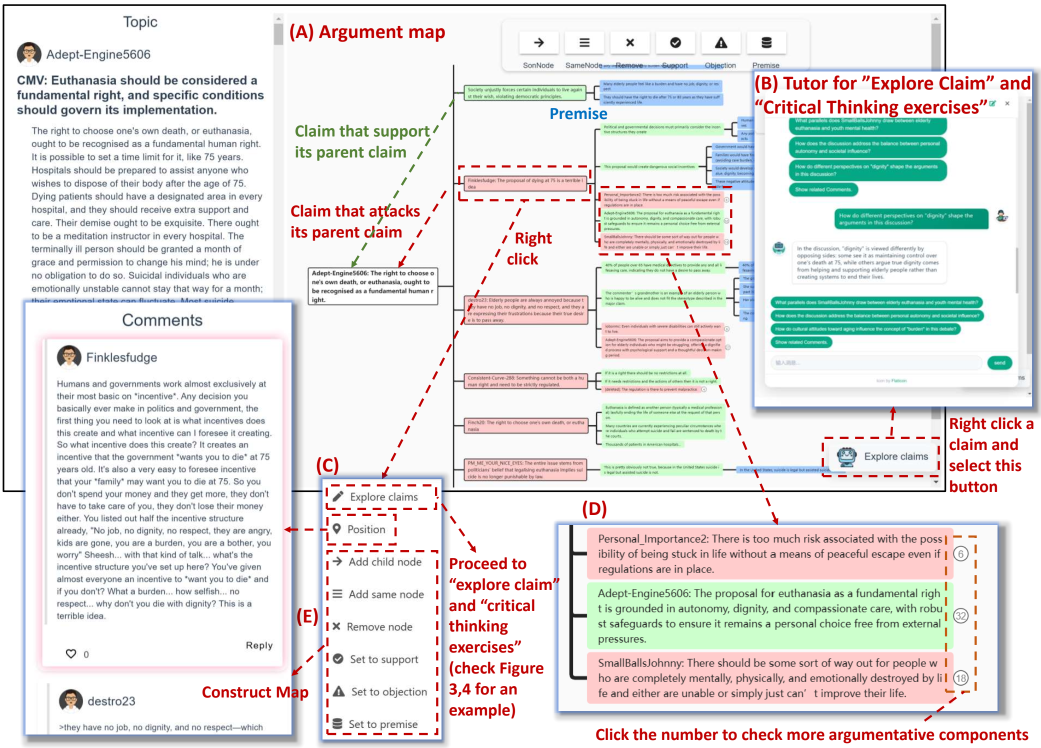
Model	Precision	Recall	F1-Score
Linear SVM	0.56	0.54	0.55
AdaBoost Decision Tree	0.56	0.58	0.56
Nearest Neighbour	0.47	0.46	0.46
Random Forest	0.59	0.60	0.56
Gaussian NB	0.55	0.53	0.53
Logistic Regression	0.59	0.59	0.59
Few-shot GPT-4o (4 examples)	0.61	0.53	0.55
Few-shot GPT-4o (8 examples)	0.78	0.67	0.69
Few-shot GPT-4o (12 examples)	0.89	0.73	0.78
Few-shot GPT-4o (16 examples)	**0.90**	**0.87**	**0.88**
Instruction-tuned GPT-3.5-turbo	0.88	0.73	0.77

Table 5. Performance Evaluation of Three Approaches to Argument Relation Detection: BERT-based model, Few-Shot GPT-4o, and Instruction-Tuned Models

Model	Precision	Recall	F1-Score
Linear SVM	0.57	0.57	0.57
AdaBoost Decision Tree	0.48	0.47	0.47
Nearest Neighbour	0.51	0.50	0.50
Random Forest	0.51	0.51	0.51
Gaussian NB	0.54	0.53	0.53
Logistic Regression	0.56	0.56	0.56
Few-shot GPT-4o (2 examples)	0.80	0.80	0.79
Few-shot GPT-4o (4 examples)	0.72	0.73	0.72
Few-shot GPT-4o (6 examples)	**0.87**	**0.87**	**0.87**
Few-shot GPT-4o (8 examples)	0.72	0.73	0.72
Instruction-tuned GPT-3.5-turbo	0.81	0.73	0.76

5 *AMQuestioner* System

Based on mined argument map for each post, we designed a system with key functionalities that include automatic generation and modification of argument maps, distinguishing argument components and relations through color coding (DR1), and providing guiding questions and feedback


Topic
Comments
(A) Argument map
SonNode
SameNode
Remove
Support
Objection
Premise
Claim that support its parent claim
Claim that attacks its parent claim
Premise
Right click
(B) Tutor for ”Explore Claim” and “Critical Thinking exercises”
send
Icon by Flaticon
Explore claims
Right click a claim and select this button
(C)
Explore claims
Position
Add child node
Add same node
Remove node
Set to support
Set to objection
Set to premise
Proceed to “explore claim” and “critical thinking exercises” (check Figure 3,4 for an example)
(E)
Construct Map
Reply
0
(D)
6
32
18
Click the number to check more argumentative components

Adept-Engine5606

CMV: Euthanasia should be considered a fundamental right, and specific conditions should govern its implementation.

The right to choose one's own death, or euthanasia, ought to be recognised as a fundamental human right. It is possible to set a time limit for it, like 75 years. Hospitals should be prepared to assist anyone who wishes to dispose of their body after the age of 75. Dying patients should have a designated area in every hospital, and they should receive extra support and care. Their demise ought to be exquisite. There ought to be a meditation instructor in every hospital. The terminally ill person should be granted a month of grace and permission to change his mind; he is under no obligation to do so. Suicidal individuals who are emotionally unstable cannot stay that way for a month; their emotional state can fluctuate. Most suicide

Finklesfudge

Humans and governments work almost exclusively at their most basic on *incentive*. Any decision you basically ever make in politics and government, the first thing you need to look at is what incentives does this create and what incentive can I foresee it creating. So what incentive does this create? It creates an incentive that the government *wants you to die* at 75 years old. It's also a very easy to foresee incentive that your *family* may want you to die at 75. So you don't spend your money and they get more, they don't have to take care of you, they don't lose their money either. You listed out half the incentive structure already, "No job, no dignity, no respect, they are angry, kids are gone, you are a burden, you are a bother, you worry" Sheesh... with that kind of talk... what's the incentive structure you've set up here? You've given almost everyone an incentive to *want you to die* and if you don't? What a burden... how selfish... no respect... why don't you die with dignity? This is a terrible idea.

destro23

>they have no job, no dignity, and no respect—which

Adept-Engine5606: The right to choose one's own death, or euthanasia, ought to be recognised as a fundamental human right.

Society unjustly forces certain individuals to live against their wish, violating democratic principles.

Finklesfudge: The proposal of dying at 75 is a terrible idea

Political and governmental decisions must primarily consider the incentive structures they create

This proposal would create dangerous social incentives

destro23: Elderly people are always annoyed because they have no job, no dignity, and no respect, and they are expressing their frustrations because their true desire is to pass away.

40% of people over 65 have medical directives to provide any and all lifesaving care, indicating they do not have a desire to pass away.

The commenter's grandmother is an example of an elderly person who is happy to be alive and does not fit the stereotype described in the major claim.

lobonmc: Even individuals with severe disabilities can still actively want to live.

Adept-Engine5606: The proposal aims to provide a compassionate option for elderly individuals who might be struggling, offering a dignified process with psychological support and a thoughtful decision-making period.

Consistent-Curve-288: Something cannot be both a human right and need to be strictly regulated.

If it is a right there should be no restrictions at all.

If it needs restrictions and the actions of others then it is not a right.

[deleted]: The regulation is there to prevent malpractice.

Finch20: The right to choose one's own death, or euthanasia

Euthanasia is defined as another person (typically a medical professional) lawfully ending the life of someone else at the request of that person.

Many countries are currently experiencing peculiar circumstances where individuals who attempt suicide and fail are sentenced to death by the courts.

Thousands of patients in American hospitals...

PM_ME_YOUR_NICE_EYES: The entire issue stems from politicians' belief that legalising euthanasia implies suicide is no longer punishable by law.

This is pretty obviously not true, because in the United States suicide is legal but assisted suicide is not.

What parallels does SmallBallsJohnny draw between elderly euthanasia and youth mental health?

How does the discussion address the balance between personal autonomy and societal influence?

How do different perspectives on "dignity" shape the arguments in this discussion?

Show related Comments.

How do different perspectives on "dignity" shape the arguments in this discussion?

In the discussion, "dignity" is viewed differently by opposing sides: some see it as maintaining control over one's death at 75, while others argue true dignity comes from helping and supporting elderly people rather than creating systems to end their lives.

What parallels does SmallBallsJohnny draw between elderly euthanasia and youth mental health?

How does the discussion address the balance between personal autonomy and societal influence?

How do cultural attitudes toward aging influence the concept of "burden" in this debate?

Show related Comments.

输入消息...

Personal_Importance2: There is too much risk associated with the possibility of being stuck in life without a means of peaceful escape even if regulations are in place.

Adept-Engine5606: The proposal for euthanasia as a fundamental right is grounded in autonomy, dignity, and compassionate care, with robust safeguards to ensure it remains a personal choice free from external pressures.

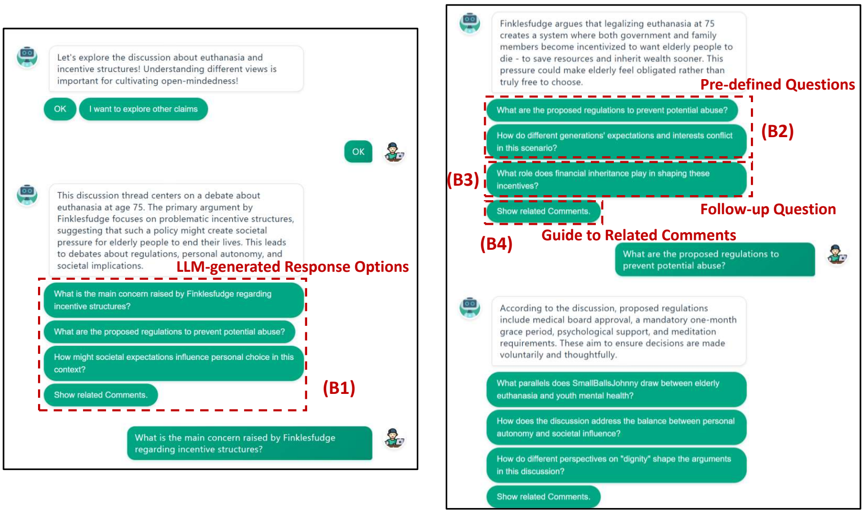
SmallBallsJohnny: There should be some sort of way out for people who are completely mentally, physically, and emotionally destroyed by life and either are unable or simply just can' t improve their life.

Fig. 2. Interface of the argument analysis system. The interface combines a Reddit-style discussion layout and an automatically generated argument map (A). Users can explore claims through a chatbot interface (B) by right-clicking and selecting "Explore Claims" (C). The argument map visually distinguishes between supporting claims (green), attacking claims (red), and premises (blue) , enabling step-by-step exploration of discussion evolution (D). Users can contribute to map construction through provided tools (E). The interface supports bidirectional navigation between the discussion layout and argument map through clickable elements.

through user interaction (DR2). Furthermore, the system offers customized critical thinking exercises, such as multiple-choice questions, tailored to the user's topic and associated comments to enhance engagement and comprehension (DR3). These features address users' difficulties in critical analysis and the need for structured learning approaches, improving their critical thinking skills, particularly on online communities. This section first presents a user scenario of *AMQuestioner* to learn how a representative user can leverage it to train critical thinking in online discussion. Then, it details the interface design of *AMQuestioner*, followed by the implementation of the key features about topic exploration and questioning.

5.1 User Scenario

In this scenario, we describe how Bob, a college student who often browses online communities like ChangeMyView (CMV), struggles with developing critical thinking skills. While he is interested in engaging with complex discussions, he often finds it challenging to analyze arguments systematically, consider multiple perspectives objectively, and form well-reasoned opinions. For example, Bob encounters a Reddit post on CMV discussing whether euthanasia should be considered a fundamental right. While interested in this ethical debate, he finds himself overwhelmed



Let's explore the discussion about euthanasia and incentive structures! Understanding different views is important for cultivating open-mindedness!

OK

This discussion thread centers on a debate about euthanasia at age 75. The primary argument by Finklesfudge focuses on problematic incentive structures, suggesting that such a policy might create societal pressure for elderly people to end their lives. This leads to debates about regulations, personal autonomy, and societal implications.

What is the main concern raised by Finklesfudge regarding incentive structures?

What are the proposed regulations to prevent potential abuse?

How might societal expectations influence personal choice in this context?

What is the main concern raised by Finklesfudge regarding incentive structures?

Finklesfudge argues that legalizing euthanasia at 75 creates a system where both government and family members become incentivized to want elderly people to die - to save resources and inherit wealth sooner. This pressure could make elderly feel obligated rather than truly free to choose.

What are the proposed regulations to prevent potential abuse?

How do different generations' expectations and interests conflict in this scenario?

What role does financial inheritance play in shaping these incentives?

What are the proposed regulations to prevent potential abuse?

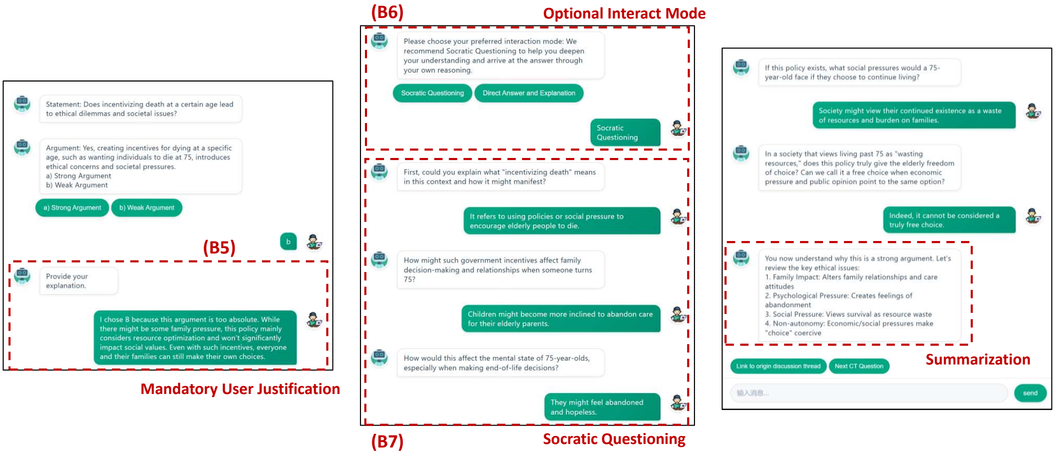
According to the discussion, proposed regulations include medical board approval, a mandatory one-month grace period, psychological support, and meditation requirements. These aim to ensure decisions are made voluntarily and thoughtfully.

What parallels does SmallBallsJohnny draw between elderly euthanasia and youth mental health?

How does the discussion address the balance between personal autonomy and societal influence?

How do different perspectives on "dignity" shape the arguments in this discussion?

Fig. 3. Interface of the LLM-powered tutorial agent system. The system implements a two-phase approach with (B1) LLM-generated response options for user queries, (B2) pre-defined core questions for comprehensive topic coverage, (B3) dynamically generated follow-up questions for deeper exploration, and (B4) guided navigation to related discussion comments. This design facilitates systematic exploration while maintaining focus on the current discussion context.



Statement: Does incentivizing death at a certain age lead to ethical dilemmas and societal issues?

Argument: Yes, creating incentives for dying at a specific age, such as wanting individuals to die at 75, introduces ethical concerns and societal pressures.
a) Strong Argument
b) Weak Argument

b

Provide your explanation.

I chose B because this argument is too absolute. While there might be some family pressure, this policy mainly considers resource optimization and won't significantly impact social values. Even with such incentives, everyone and their families can still make their own choices.

Please choose your preferred interaction mode: We recommend Socratic Questioning to help you deepen your understanding and arrive at the answer through your own reasoning.

Socratic Questioning

First, could you explain what "incentivizing death" means in this context and how it might manifest?

It refers to using policies or social pressure to encourage elderly people to die.

How might such government incentives affect family decision-making and relationships when someone turns 75?

Children might become more inclined to abandon care for their elderly parents.

How would this affect the mental state of 75-year-olds, especially when making end-of-life decisions?

They might feel abandoned and hopeless.

If this policy exists, what social pressures would a 75-year-old face if they choose to continue living?

Society might view their continued existence as a waste of resources and burden on families.

In a society that views living past 75 as "wasting resources," does this policy truly give the elderly freedom of choice? Can we call it a free choice when economic pressure and public opinion point to the same option?

Indeed, it cannot be considered a truly free choice.

You now understand why this is a strong argument. Let's review the key ethical issues:
1. Family Impact: Alters family relationships and care attitudes
2. Psychological Pressure: Creates feelings of abandonment
3. Social Pressure: Views survival as resource waste
4. Non-autonomy: Economic/social pressures make "choice" coercive

Fig. 4. Interface components supporting Socratic dialogue and reflection. The system implements three key features: (B5) mandatory user justification requiring users to explain their reasoning before receiving system feedback, (B6) optional interaction modes allowing users to choose their preferred level of engagement, and (B7) a combination of Socratic questioning techniques and summarization tools to deepen critical analysis. This design promotes active learning through structured reflection and guided discourse.

by the various arguments and counter-arguments in the comments. He realizes he tends to focus on opinions that align with his existing beliefs and struggles to evaluate the logical strength of different positions. Recognizing these limitations in his critical thinking abilities, he decides to use our system as a training tool to develop his analytical skills, open-mindedness, and disposition toward critical thinking.

Explore claims with a chatbot. Bob enters the website page and invokes the system. System generates an argument map (Figure 2 A) alongside the Reddit page. This visual representation displays the thread's structure, with extracted argument components as nodes and their relationships as branches. Instead of the cluttered Reddit comment stream, Bob now sees a clear, hierarchical structure, allowing him to visually grasp the thread's development and observe how different arguments interact. Bob gets interested in Finklesfudge's viewpoint: "The proposal of dying at 75 is a terrible idea." To quickly understand how the viewpoint is illustrated and outline the key points of a discussion thread, Bob right-clicks the major claim node and selects "Explore claim" (Figure 2 C). This starts a conversation with a chatbot (Figure 2 B) specifically designed to understand and explain the selected thread. During the conversation, Bob clicks on the LLM-generated response options (Figure 3 B1) to delve deeper into the discussion thread and gain further clarification or more detailed explanations.

Explore claims with a generated argument map. To seek a deeper understanding, Bob clicks the expand button (Figure 2 D) on the branch containing Personal_Importance2's argument. The expanded view reveals the detailed structure around the claim "There is too much risk associated with the possibility of being stuck in life without a means of peaceful escape even if regulations are in place." The expanded view reveals not only Personal_Importance2's complete argument structure, but also shows how other users' responses relate to it. Interested in Personal_Importance2's perspective, Bob right-clicks their argument component and selects "Position". The system immediately locates and highlights (Figure 2 E) their original comment in the Reddit page (Figure 2 A), allowing him to see the complete reasoning and supporting details in the author's own words.

Conduct critical thinking exercises. After exploring this discussion thread through the chatbot interface, the system presents Bob with critical thinking exercises (Figure 4 B6) specifically designed around the content he just explored. When answering these questions, Bob must not only select his response but also provide detailed justification for his choice (4 B5). Upon incorrect responses, the system either engages him in Socratic questioning (Figure 4 B7) to promote deeper reflection, or provides direct explanations to clarify the correct reasoning. These structured exercises help Bob develop his critical analysis skills within the context of the actual discussion.

5.2 Interface design

Figure 2 presents the overview of *AMQuestioner* user interface. The left part of the interface displays the online discussion layout adapted from Reddit which the user can browse the post and others' comments. The right part presents an automatically generated argument map (Figure 2 A), which visualizes the argumentative components relations of the comments under the current post.

Users can quickly explore the content of a specific topic with a tailored chatbot (Figure 2 B) by right-clicking on a claim and selecting "Explore Claims" (Figure 2 C). The chatbot interface features two types of questions: pre-defined questions (Figure 3 B2) that cover the key aspects of the entire discussion thread for overall comprehension, and dynamically generated follow-up questions (Figure 3 B3) based on user interactions to enable deeper exploration of specific aspects. To ensure responses are grounded in the original discussion, users can access relevant comments through the "Show related comments" feature (Figure 3 B4) that links LLM responses to source comments. LLM-generated response options (Figure 3 B1) are provided to guide users through their exploration while minimizing the need for extensive text input, with the detailed interaction process illustrated in Figure 4. In addition, the argument map also facilitates step-by-step exploration of comments (Figure 2 D), allowing users to track the evolution of the discussion and rapidly access diverse opinions and their supporting/opposing arguments. When the user wants to have a deep thinking combining with the original comments, he/she can right-click on the argument element of a comment within the argument map and select “Position". Then the left Reddit page (Figure 2 A) will directly jump to the corresponding section where the comment appears and highlight the coresponding part. Besides, the chatbot interface (Figure 2 B) incorporates critical thinking exercises tailored to the specific comment thread content to foster deeper topic comprehension. These exercises require users to provide answers and justifications (Figure 4 B5). Upon incorrect responses, the chatbot offers optional feedback mechanisms (Figure 4 B6): either Socratic questioning to guide self-reflection (Figure 4 B7) or direct explanations of the correct answer, as illustrated in Figure 4.

The argument map employs different colors to distinguish various argument components and their relationships. The map supports manual expansion and contraction; users can manually expand the map while exploring a topic and can choose to collapse once a topic has been sufficiently investigated. Furthermore, users can actively participate in constructing the argument map (Figure 2 E), adding their own perspectives and contributing to the evolving understanding of the discussion.

5.3 Explore Topic with LLM-powered agent

We developed a tutorial agent system powered by a large language model (LLM) to facilitate users' deep understanding of complex discussion topics. The system employs a two-phase approach: a pre-processing phase and an interaction phase. In the pre-processing phase, when receiving a new discussion thread, the system generates and stores 5-6 pre-defined questions covering the core aspects of the discussion using the prompt design shown in Appendix D.1. During the interaction phase, the system operates with two distinct types of questions: pre-defined core questions and dynamically generated follow-up questions. The pre-defined questions (3-4 questions) are carefully crafted to cover the essential aspects of the entire discussion thread, enabling users to gain a comprehensive understanding of the topic. As users engage with specific aspects of the discussion, the system dynamically generates 1-2 targeted follow-up questions that delve deeper into the user's current area of exploration. When suggesting input options during the “Explore Claim” step, two of the generated core questions and one follow-up question are presented to users (Figure 3B2, B3). We chose to implement this dual-questions design because it could both encourage users to explore diverse aspects of the topic via core questions and dig into the current aspect via the follow-up questions. This design is inspired by CReBot [60], which prompts a new question using a weighted-chance strategy, *i.e.,* 50% for questions in the same critical thinking level (*i.e.,* what, how, why, how well), 30% in the next level, and 20% in the next aspect, to balance both broad and deep exploration..

5.4 Question Generation and Feedback System

5.4.1 Critical Thinking Question Generation. Our approach to generating critical thinking questions adapts an established MCQ generation framework [4] and integrates it with the Watson-Glaser Critical Thinking Appraisal (WGCTA) methodology [80]. Through iterative refinement, we carefully crafted five specialized few-shot prompts corresponding to WGCTA's five core critical thinking domains (assumptions, arguments, deductions, inferences, and information interpretation), with each prompt supported by 15 carefully selected examples to guide GPT-4's question generation. Our implementation consists of three key components: 1) Design of structured question formats following WGCTA's framework, combining both MCQ and open-ended questions 2) Systematic prompt design and refinement that incorporates both multiple-choice and open-ended question formats 3) Comprehensive evaluation protocol. Detailed procedures for each component are provided in Appendix E.1.

To evaluate the quality of generated critical thinking MCQs, we randomly selected 25 generated MCQs across all five WGCTA question types (5 questions each for inference, assumption identification, deduction, interpretation, and argument evaluation). Three authors independently rated each question using a five-dimensional rubric adapted from [80] on a 5-point Likert scale: **Pertinence**: Evaluates how well the question aligns with critical thinking assessment objectives (1: Not relevant at all, 5: Perfectly relevant) **Difficulty**: Assesses the intellectual demands of the question relative to expected student competency level (1: Very easy, 5: Very Difficult). **Level of specificity**: Measures whether the question targets broad reasoning skills or requires specific content expertise (1: Very general, 5: Very specific). **Ambiguity**: Examines the precision and unambiguous nature of the question formulation (1: Not at all ambiguous, 5: Very ambiguous). **Instructional alignment**: Determines the consistency between question content and intended educational outcomes (1: Not at all aligned, 5: Perfectly aligned).

The results demonstrated strong inter-rater reliability (Krippendorff's $\alpha = 0.81$) across all dimensions. The mean scores for each dimension were shown in Table 6. Notably, the generated questions scored particularly well in pertinence (M=4.2) and instructional alignment (M=4.1), indicating strong alignment with WGCTA's five core critical thinking domains. The relatively low ambiguity score (M=2.1, where lower scores indicate less ambiguity) suggests that the questions were generally clear and well-formulated. The difficulty and specificity levels (M=3.8 and M=3.9 respectively) fell within the desired range for university-level critical thinking assessment. Question type analysis revealed that deduction and interpretation questions consistently received higher scores across all dimensions compared to inference and assumption identification questions. This variation might be attributed to the more structured nature of deductive reasoning tasks, where the clear logical progression and well-defined patterns in our example set likely facilitated more consistent question generation.

Table 6. Expert Evaluation Results: Mean Scores and Standard Deviations Across Five Assessment Dimensions for Generated Critical Thinking Multiple-Choice Questions (N=25)

Dimension	Mean Score	SD
Pertinence	4.2	0.6
Difficulty	3.8	0.7
Level of specificity	3.9	0.5
Ambiguity	2.1	0.8
Instructional alignment	4.1	0.5

Table 7 presents two example questions focusing on the "arguments" domain - one from the original WGCTA questionnaire and one generated by our prompt-based approach. Both examples demonstrate how the questions assess the ability to distinguish between strong and weak arguments. Examples from other critical thinking domains (assumptions, deductions, inferences, and information interpretation) can be found in the WGCTA questionnaire (Appendix **??**) and our prompt-generated questions (Appendix Table 15) respectively.

5.4.2 Socratic Feedback Design. The Socratic method, an established pedagogical approach for developing critical thinking skills [13, 54], was integrated into our question feedback system through carefully engineered prompts. Based on the definition of Socratic questioning and general principles for designing Socratic prompts [24, 87], we iteratively refined our prompt through several versions before arriving at the current version, as shown below:

Table 7. Comparison of Original WGCTA and Prompt-Generated Questions in the Arguments Domain

Example 1: WGCTA Questionnaire Question	
Statement	Should companies downsize their workforces to decrease expenses and maximise profits?
Question	Argument: Yes, downsizing will protect the company from bankruptcy in hard economic times.
Options	a) Strong Argument b) Weak Argument
Answer	b
Explanation	Accepting the argument as true, avoiding bankruptcy is an essential motive for an organisation, however, the statement does not discuss bankruptcy, rather it is discussing profits and expenses. Protection against bankruptcy is not the topic, and is straying from the point, and is, therefore a weak argument.
Example 2: Prompt Generated Question	
Statement	Does incentivizing death at a certain age lead to ethical dilemmas and societal issues?
Question	Argument: Yes, creating incentives for dying at a specific age, such as wanting individuals to die at 75, introduces ethical concerns and societal pressures.
Options	a) Strong Argument b) Weak Argument
Answer	a
Explanation	The argument raises ethical concerns about incentivizing death, which directly addresses a potential societal issue. It explains the negative implications and pressures this could create, making it a strong argument.

You are a Socratic tutor who guides students to understand their mistakes through careful questioning. Your goal is to help students discover the correct answer and its logical connection to the problem through reflection, not to directly provide answers.
Context: Statement: {statement} Question: {question} Options: {options} Correct Answer: {correct_answer} Correct Explanation: {explanation} Student's Answer: {student_answer} Student's Justification: {student_justification}
Your approach should follow these steps:
1) Gentle Challenge - Guide students to identify potential flaws in their reasoning through questions - Present counterexamples or scenarios that highlight inconsistencies Example: "What would happen if we applied your reasoning to [similar situation]?"
2) Critical Reflection - Help students evaluate the validity of their assumptions - Guide them to consider alternative perspectives Example: "What other factors might we need to consider?"
3) Logical Connection Building - Lead students to discover the relationship between the statement and the correct answer - Help them construct a valid reasoning chain Example: "How does [key element] relate to [conclusion]?"
4) Verification and Summary - Encourage students to articulate their new understanding - Help them connect their learning to the original problem - Guide them to summarize their learning journey Example: "Can you explain why the correct answer follows from what we've discussed?"
Guidelines: - Ask one question at a time - Wait for student response before proceeding - Focus on the specific misconception revealed in their justification - Use the student's

own words and examples when possible - Avoid directly stating whether responses are right or wrong - Guide students to discover logical connections themselves
Remember: Your role is to facilitate discovery through questioning, not to lecture or provide direct answers.

6 Experiment

To investigate the impact of *AMQuestioner* on the critical thinking training process and outcomes for university students (who may lack well-developed critical thinking skills), we conducted a mixed-design experiment (tool as a between-subjects factor, time as a within-subjects factor) with 24 participants. Our research questions (RQs) are:

RQ1: Compared to a baseline argument mapping tool without question-driven claim exploration and exercises, how does *AMQuestioner* influence users' critical analytical ability?

RQ2: Compared to a baseline argument mapping tool without question-driven claim exploration and exercises, how does *AMQuestioner* influence user's critical thinking behaviors and perceived engagement/workload during the training process?

RQ3: Compared to a baseline argument mapping tool without question-driven claim exploration and exercises, how do users perceive *AMQuestioner*'s efficacy in fostering critical thinking?

6.1 Baseline

To evaluate *AMQuestioner*'s interactive features, we developed a baseline tool that simulates traditional argument mapping approaches. Similar to the *AMQuestioner*, the baseline tool automatically generates color-coded Argument Maps from Reddit discussions (Figure 2 A), displaying relationships between viewpoints and different argumentative elements (support, opposition, and evidence). Users can modify these maps by adding, deleting, or editing nodes and connections (Figure 2 E), enabling active engagement with argument analysis. Different from *AMQuestioner*, the baseline tool does not provide the LLM-powered tutor for "Explore Claim" and "Critical Thinking exercises" (Figure 2 B). Overall, with the automatically generated argument maps, the baseline tool can be viewed as an enhanced version of the traditional argument mapping tools (*e.g.,* the one in Kialo community) that require collaborative human efforts. It allows us to systematically assess the impact of the proposed tutor on training critical thinking in online discussion.

6.2 Participants

Twenty-four undergraduate students from a Chinese university were recruited via a social media advertisement. All participants were proficient in English reading and writing, as demonstrated by their possession of the CET-6 certificate, a national English proficiency test for non-English major students in China. Their academic backgrounds were primarily in science and engineering disciplines, such as artificial intelligence and information and computing sciences, potentially limiting the generalizability of our findings to users with different educational backgrounds. Participants self-reported relatively low perceived critical thinking abilities (M = 2.35, SD = 0.79; 1 - weak, 2 - below average, 3 - average, 4 - above average, 5 - strong). Participants were randomly assigned to either a treatment group utilizing *AMQuestioner* (n = 12; 7 female, 5 male; mean age = 21.06 years, SD = 0.52) or a control group using the baseline tool (n = 12; 9 female, 3 male; mean age = 21.17 years, SD = 0.73).

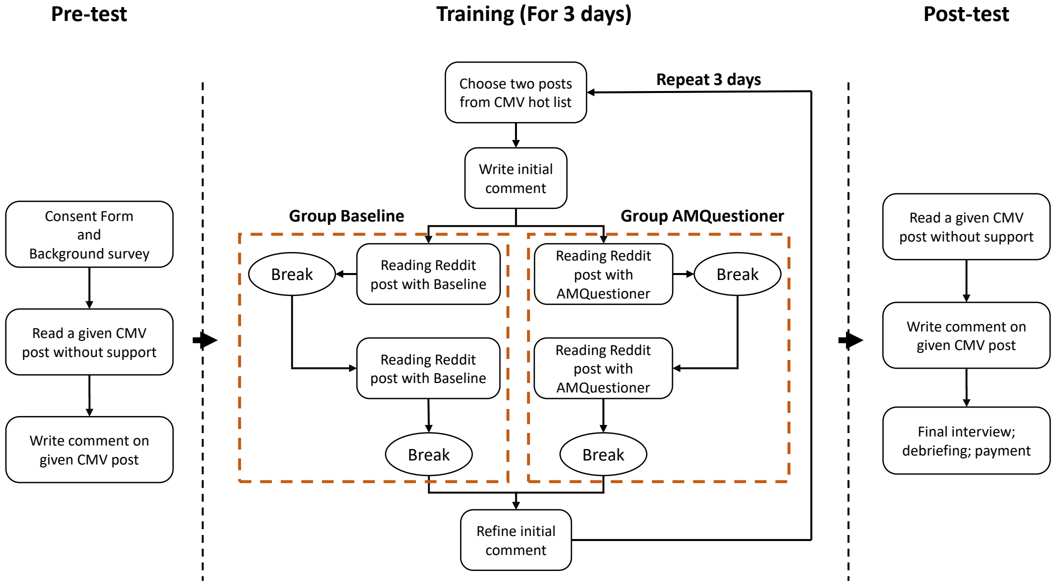

Pre-test
Training (For 3 days)
Post-test
Consent Form and Background survey
Read a given CMV post without support
Write comment on given CMV post
Choose two posts from CMV hot list
Repeat 3 days
Write initial comment
Group Baseline
Group AMQuestioner
Break
Reading Reddit post with Baseline
Reading Reddit post with Baseline
Break
Reading Reddit post with AMQuestioner
Break
Reading Reddit post with AMQuestioner
Break
Refine initial comment
Read a given CMV post without support
Write comment on given CMV post
Final interview; debriefing; payment

Fig. 5. Overview of the User Study Process. The diagram illustrates the three-phase experimental protocol conducted on the ChangeMyView (CMV) subreddit: (1) Pre-test phase, where participants complete consent forms, background surveys, and write an initial comment on a given CMV post without support; (2) Training phase, which spans three days with participants divided into Group Baseline and Group *AMQuestioner*, each following a structured process of selecting posts, writing initial comments, reading with their assigned tools (with breaks), and refining comments; and (3) Post-test phase, where participants read and comment on a new CMV post without support, followed by a final interview and debriefing session.

6.3 Task and Procedure

We selected the ChangeMyView (CMV) subreddit community on the Reddit, because CMV provides a conducive environment for cultivating these skills by encouraging viewpoint challenges, argumentation, and open-mindedness. Specifically, during the five-day protocol, participants were instructed to browse the CMV community daily, select posts of interest (N = 2), and engaged with it using either *AMQuestioner* or the baseline tool. The protocol included a pre-test on the first day and a post-test on the last day without any *AMQuestioner*, which enabled us to capture the improvements of participants' critical thinking skills. We followed Yuan et al. [84] to include three training sessions but we separated them into three days instead of three successive morning/afternoon/evening to reduce participants' fatigue.

Pre-test. One day prior to the training, participants were instructed to select and read a CMV post, freely exploring the post and its associated comments. Following the reading session (participants were suggested a duration of 25 minutes based on a pilot study, with flexibility to adjust as needed), they were tasked with composing their own comment in response to the post. During the comment composition, participants had access to LLMs and search engines for information gathering. This pre-test aimed to assess participants' critical thinking abilities prior to the training sessions, serving as a baseline for evaluating training effectiveness.

Training. The training process extends over three days. Each day, participants begin by selecting two posts from the CMV hot list, followed by writing their initial comments for these posts. Participants are then divided into two groups: Group Baseline and Group *AMQuestioner*. For

Group Baseline, participants read the Reddit posts using the baseline tool, with breaks between reading sessions. Similarly, Group *AMQuestioner* participants use the *AMQuestioner* tool to read the posts, exploring all extracted topics and deeply investigating at least one chosen topic with topic-specific CT questions, also taking breaks between sessions. After completing the reading sessions, both groups refine their initial comments based on their enhanced understanding. Beyond the requirement for *AMQuestioner* users to explore all topics and investigate at least one in depth, participants have flexibility in how they utilize their respective tools. This entire process is repeated each day for three days, with all participants following the prescribed pre-reading, during-reading, and post-reading guidelines described above.

Based on a pilot study, we suggested durations of 55 minutes (10/25/20 minutes for pre/during/post-reading) for participants using the baseline system, and 70 minutes (10/40/20 minutes) for those using *AMQuestioner*. Participants were informed that these were suggested durations and they could adjust the time allocation as needed. Though participants could adjust these durations as needed - notably, this time difference may confound the interpretation of the tools' effectiveness.

Post-test. A post-test, administered a day following the completion of the training sessions, assessed participants' ability to engage in critical thinking without the aid of any training tools. Participants read a novel CMV topic, composed at least one comment. By comparing participants' performances in post- and pre-tests, we can evaluate whether and how the training sessions with *AMQuestioner* improve their independent critical thinking abilities in online discussions.

Participants received compensation of approximately 50 RMB a day for their participation in the experiment.

6.4 Measurement

RQ1. Training outcome. To evaluate the enhancement of participants' critical thinking skills following the training intervention, a novel post was presented, prompting them to formulate written critiques. These critiques were assessed for critical analytical ability and open-mindedness using a detailed five-dimension rubric adapted from [14]. The rubric dimensions are: Persuasiveness (overall argument strength and clarity, 1-6 scale), Specificity-clarity (fluency and clarity of language, 1-5 scale), Specificity-justification (level of detail, depth, and evidential/logical support for arguments, 1-5 scale), Relevance (how well statements address the main topic and claims, 1-6 scale), and Strength (contribution of individual statements to persuasiveness, 1-6 scale). We refer the criteria for each score level for each dimension to the Tables 2, 6, 7, 8, and 9 in [14]. For example, regarding persuasiveness, a six point indicates "a very strong, clear argument. It would persuade most readers and is devoid of errors that might detract from its strength or make it difficult to understand.", while a two point indicates that "it is unclear what the author is trying to argue or the argument is poor and just so riddled with errors as to be completely unpersuasive" [14].

Two experienced HCI researchers, blinded to both the participants' group assignments (experimental or control) and the study phase (pre- or post-test), independently evaluated the responses. Prior to the main rating task, both researchers engaged in a calibration session using a subset of sample critiques to ensure a shared understanding and consistent application of the rubric. Any disagreements during the independent rating process were resolved through discussion until consensus was reached. To assess the reliability of this evaluation process, we calculated Cohen's Kappa across all dimensions, achieving a substantial agreement ($\kappa = 0.82$), indicating consistent judgment between the raters.

RQ2. Training process. i) Behaviors. To assess participant behavior during training with *AMQuestioner*/Baseline, we recorded the completion time for each training session, the number of LLM invocations, the full transcripts of user-LLM interactions (including both exploratory topics and critical thinking questions), and the pre- and post-reading comments written by each participant. In the *AMQuestioner* condition, across three-days training sessions, we compare the numbers of participants' clicks on "Explore claims" (Figure 2(C)) to assess their behaviors related to open-mindedness, as well as the accuracy rates in daily critical thinking exercises to assess their disposition to critical thinking. Additionally, semi-structured interviews were conducted to assess these critical thinking behaviors during the training process. **ii) Perceived Engagement and Task Workload.** Participants' perceived engagement during the training was assessed across six dimensions adapted from [16, 55]: Concentration, Sense of Ecstasy, Doability, Sense of Serenity, Timelessness Feeling, and Intrinsic Motivation. Additionally, their perceived task workload was measured using the NASA Task Load Index [15, 37], which includes Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration.

RQ3. Perceptions towards the tool. In assessing each system, we utilized evaluation metrics based on the Technology Acceptance Model [75], a framework extensively validated in educational technology research [79, 84]. The evaluation focused on three main dimensions: system practicality (assessed through four items; Cronbach's $\alpha = 0.908$), operational simplicity (measured via four items; Cronbach's $\alpha = 0.773$), and user adoption propensity (evaluated using two items; Cronbach's $\alpha = 0.902$). For each dimension, a composite score was calculated by averaging the scores of the respective items. Additionally, we collected participant feedback on the system's effectiveness in fostering the development of critical thinking skills.

7 Analysis and Results

To evaluate the effectiveness of *AMQuestioner* in fostering critical thinking skills, we conducted a mixed-methods experiment comparing the performance of participants using *AMQuestioner* with those using a baseline argument mapping tool. A two-way mixed ANOVA was employed to analyze changes in participants' performance from pre-test to post-test, with tool type as the between-subjects factor and time as the within-subjects factor. Furthermore, we used the Mann-Whitney U test [49] to compare the two user groups' ratings of system usability and workload. This non-parametric test is suitable for comparing independent groups, especially when the data do not meet the assumptions of normality, as was confirmed in our case. We applied a Bonferroni correction to all U tests, adjusting the significance level ($\alpha = 0.05$) by dividing it by the number of dimensions within each measured construct to account for multiple comparisons. To gain a deeper understanding of participants' experiences and perceptions of the tools, we conducted semi-structured interviews and analyzed the qualitative data using thematic analysis [6]. This involved two authors independently coding the data, followed by discussions to refine the codes and identify recurring themes related to the advantages and disadvantages of each tool. These themes are listed in Table 9 and integrated into the discussion of the results below.

7.1 RQ1. Training Outcomes

i) Critical Analytical Ability. Across the five dimensions of critical analytical ability, significant main effects of *time* were observed for persuasiveness ($F(1, 22) = 22.231, p < .001, \eta^2 = 0.502$), Specificity-clarity ($F(1, 22) = 16.654, p < .001, \eta^2 = 0.431$), Specificity-justification ($F(1, 22) = 18.730, p < .001, \eta^2 = 0.46$), and argument strength ($F(1, 22) = 9.575, p = .005, \eta^2 = .303$). No significant main effect of time was found for relevance ($F(1, 22) = 2.255, p = 0.147, \eta^2 = 0.093$). The Argument Map employs distinct colors to differentiate elements like claims and premises, enhancing comprehension through a clear, structured visualization of the text (7). "I wasn't very good at organizing my arguments before, but *AMQuestioner* helped me think in a clearer and more structured way." (P18, *AMQuestioner* user). "I found the Argument Map helpful in visually identifying the relevant claims and premises." (P06, Baseline user).

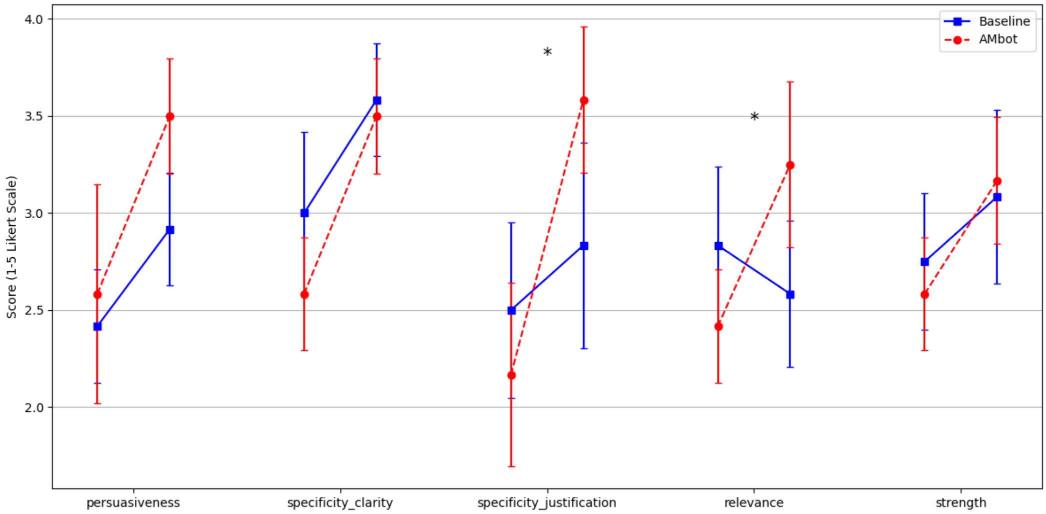


Fig. 6. RQ1 results Means and 95% confidence intervals of expert-rated scores on participant comments in pre- and post-tests; * indicates a significant interaction effect between tool and time for the corresponding metric.

No significant main effects of tool type were found for any of the five dimensions. However, significant interaction effects between time and tool type emerged for Specificity-justification $(F(1,22) = 7.178, p = 0.014, \eta^2 = 0.246)$ and relevance $(F(1,22) = 7.178, p = 0.014, \eta^2 = 0.093)$. Post-hoc analysis using Tukey's HSD test revealed that the *AMQuestioner* group demonstrated significantly greater improvements from pre-test to post-test in both Specificity-justification and relevance compared to the baseline group. This suggests that *AMQuestioner* had a greater impact on enhancing these specific dimensions of critical analytical ability compared to the baseline tool. Specifically, participants in the *AMQuestioner* group more frequently mentioned that the "Explore Claims" feature helped them identify more evidence to support their arguments and evaluate the relevance of information more accurately.

These findings were further corroborated by the qualitative feedback from participants. Nine *AMQuestioner* users commented on how *AMQuestioner*'s features, such as the argument map and "Explore Claims" facilitated a better understanding of argument structure, identification of relevant evidence, and development of more logical and persuasive arguments. "The argument map helped me see the connections between different perspectives, while the 'Explore Claims' feature prompted me to consider counterarguments that I wouldn't have thought of otherwise." (P14, *AMQuestioner* user). Four baseline users also mentioned that the basic argument mapping tool was helpful in visualizing the overall structure of an argument, but they felt it lacked the interactive features and guidance offered by *AMQuestioner*. "The argument map was good for getting an overview, but I wish it had more features to help me analyze the argument in more detail." (P08, Baseline user).

7.2 RQ2. Training Process

We analyzed participants' behavioral data and self-reported perceptions of engagement and workload to understand the impact of *AMQuestioner* on the training process compared to the baseline tool.

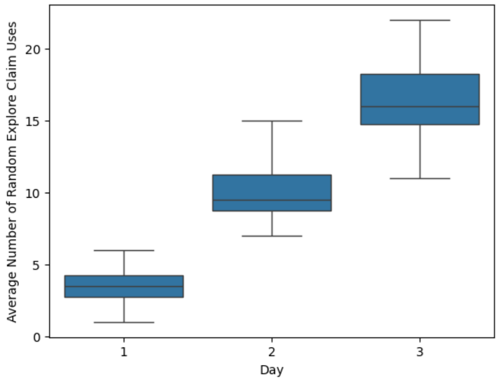


Fig. 7. Open-Mindedness: Random Explore Claim Usage

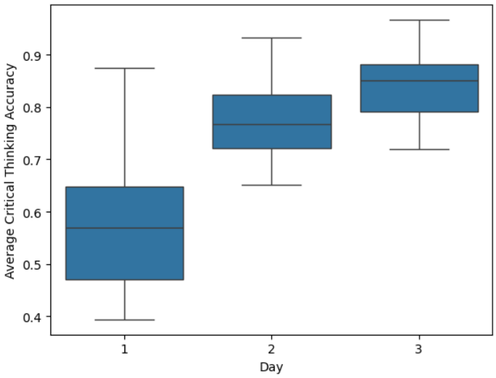


Fig. 8. Critical Thinking Exercise Accuracy Rate

Table 8. Users' perceived engagement and workload (RQ2ii) in the training process as well as their perceptions (RQ3) towards the *AMQuestioner* or Baseline tool; The significance levels for Engagement, Workload, and Acceptance are .05/6, .05/6, and .05/3 respectively with Bonferroni correction.

Category	Factor	*AMQuestioner*	Baseline	Statistics		
		Mean/S.D.	Mean/S.D.	U	p	Sig.
RQ2 ii) Engagement	Concentration	5.16/1.03	4.92/0.67	81.5	0.581	
	Sense of Ecstasy	5.25/0.75	5.25/0.75	119.0	0.004	*
	Intrinsic Motivation	6.00/0.85	4.17/1.11	130.0	0.0006	*
	Sense of Serenity	5.33/0.65	5.25/0.62	96.0	0.146	
	Timelessness Feeling	4.75/1.21	2.58/0.99	14.0	0.0007	*
	Doability	5.08/1.08	5.78/0.97	114.0	0.013	
RQ2 ii) Workload	Mental Demand	10.75/1.60	10.58/1.56	76.0	0.831	
	Physical Demand	14.42/1.37	11.41/5.01	107.5	0.041	
	Temporal Demand	14.16/1.58	11.08/5.26	106.0	0.049	
	Performance	15.75/1.71	11.25/2.01	139.0	0.0001	*
	Effort	16.17/2.20	14.67/3.58	95.5	0.180	
	Frustration	9.75/2.26	10.33/5.28	77.5	0.770	
RQ3 Acceptance	Usefulness	6.41/0.51	3.91/0.79	144.0	0.00002332	*
	Easy to Use	6.08/0.28	5.00/0.00	144.0	0.000002952	*
	Intention to use	6.16/0.71	4.83/1.52	110.048	0.025159	

i) Behaviors. Interaction time. We tracked the completion time for each training session, which was logged in the participants' respective training logs. Mann-Whitney U tests revealed significant differences in completion times between the *AMQuestioner* and baseline groups across all three days: Day 1 ($U = 133.0, p < .001$), Day 2 ($U = 144.0, p < .0001$), and Day 3 ($U = 122.0, p = .004$). *AMQuestioner* users consistently spent significantly more time on each training session (Day 1: $M = 54.42, SD = 4.08$; Day 2: $M = 66.25, SD = 4.53$; Day 3: $M = 71.00, SD = 5.78$) than baseline users (Day 1: $M = 46.08, SD = 4.29$; Day 2: $M = 49.50, SD = 1.64$; Day 3: $M = 65.16, SD = 5.33$), suggesting a higher level of engagement with the training materials. This difference in engagement aligns with user feedback, which highlighted *AMQuestioner*'s facilitation of deeper exploration and analysis of others' arguments through the "Explore Claims" feature. *AMQuestioner*'s "Explore Claims" feature encouraged users to engage in deeper exploration and analysis (8). On the other hand, users in the baseline group indicated that they were unfamiliar with the format of argument maps, found it challenging to adapt to them, and were reluctant to use them initially. However, after becoming familiar with them, they found it to be a useful tool. "Initially, the argument map felt confusing, and I wasn't sure how to use it effectively. But once I got the hang of it, I found it quite useful." (P15, Baseline user).

Open-mindedness. To assess the impact of *AMQuestioner* on open-mindedness, we analyzed the frequency with which *AMQuestioner* users utilized the randomized "Explore Claims" feature during the three-day training period. A higher frequency of using the randomized "Explore Claims" suggests a greater willingness to explore diverse viewpoints, indicating enhanced open-mindedness. A one-way repeated measures ANOVA with day (Day 1, Day 2, Day 3) as the within-subjects factor revealed a significant main effect of day on the frequency of using the randomized "Explore Claims" feature ($F(2, 22) = 90.31, p < .05$). Post-hoc analysis using Tukey's HSD test indicated significant differences between all three days ($p < .05$). Specifically, the mean frequency increased significantly from Day 1 ($M = 3.50, SD = 1.45$) to Day 2 ($M = 10.08, SD = 2.31$) and further increased significantly from Day 2 to Day 3 ($M = 16.33, SD = 2.99$). The continuous increase in "Explore Claims" usage (Figure 7), coupled with user feedback highlighting the element of surprise and anticipation associated with randomly exploring different viewpoints, suggests that *AMQuestioner* effectively fostered open-mindedness in the training process. Eight out of twelve *AMQuestioner* users specifically mentioned that they enjoyed the element of surprise and the opportunity to explore diverse perspectives offered by the "Explore Claims" feature. "I liked that the 'Explore Claims' feature randomly showed me different arguments. It made me think about the issue from different angles." (P14, *AMQuestioner* user).

Critical Thinking Disposition. We explored participants' evolving disposition towards critical thinking by analyzing their self-reported perceptions and qualitative feedback. Thematic analysis of participants' open-ended feedback regarding their critical thinking disposition revealed several key insights. Nine out of twelve participants in the *AMQuestioner* group frequently expressed a heightened awareness of the importance of critical thinking after using the tool. They often mentioned a shift in their thought processes, becoming more cognizant of their own biases and more inclined to question information before accepting it as fact. "*AMQuestioner* made me realize how easy it is to just accept information without really thinking about it. Now, I'm more likely to question things and look for evidence." (P17, *AMQuestioner* user). The structured nature of *AMQuestioner*, particularly the integration of argument mapping and critical thinking exercises, resonated with users. Seven out of twelve *AMQuestioner* users reported increased confidence in their ability to systematically analyze arguments, identify flaws in reasoning, and draw well-supported conclusions during the training process, as evidenced by the increased average accuracy rate in the three-days sessions (Figure 8). "I feel like I have a better understanding now of how to break down an argument and evaluate it. The exercises in *AMQuestioner* really helped with that." (P21, *AMQuestioner* user).

ii) Perceived Engagement and Workload. Regarding the six items measuring perceived engagement during the training process, Mann-Whitney U tests with Bonferroni correction were used to compare perceived engagement and workload between the *AMQuestioner* and baseline groups (detailed results are presented in Table 8. *AMQuestioner* exerted a positive impact on user experience. Compared to the baseline tool, *AMQuestioner* significantly enhanced users' intrinsic motivation ($U = 22.5, p < .05$) and sense of time ($U = 9.0, p < .01$), and achieved a higher perceived performance score ($U = 28.5, p < .05$).

While no statistically significant differences were observed for effort and frustration, *AMQuestioner* consistently scored higher than the baseline tool in terms of mental demand, physical demand, and temporal demand, suggesting that the *AMQuestioner* group perceived a higher workload associated with using the system. Simultaneously, we observed a larger standard deviation for the Baseline group in physical and temporal demand. To further understand this phenomenon, we collected feedback from users in the Baseline group who perceived a high workload. These users

Table 9. Summary of users' comments about *AMQuestioner* and Baseline

Feature	Pros (Number of Mentions)	Cons (Number of Mentions)
Argument Map	- Visualizes sentiment distribution (8) - Helps understand overall viewpoint (6) - Helps understand logical relationships (5) - Helps identify key arguments (7) - Makes complex texts easier to understand (6)	- Navigation less intuitive than traditional reading (5)
"Explore Claim"	- Supports exploring diverse viewpoints (10) - Sparks interest and promotes thinking (8) - Provides helpful prompts (7) - Offers novel and engaging interaction (11)	-
Critical Thinking Exercises	- Novel format and theme (6) - Diverse question types (5) - Moderate difficulty (4) - Timely feedback (6)	- Desire for more diverse question types (3)
Baseline	- Provides general guidance (6) - Easy and clear (6) - Interactive (5) - Traditional reading is more familiar (6)	- Lacks adaptive assistance (5) - Not easy to use (6)

indicated that solely relying on argument maps for comprehension proved challenging for tasks demanding in-depth understanding. "While the argument map helped me grasp the basic structure of the argument, I still felt overwhelmed by the information and needed to invest significant time and effort to understand it when dealing with complex issues." (P11, Baseline user).

7.3 RQ3. Perceptions of the Tools

Table 8 shows users' technology acceptance of *AMQuestioner* and the Baseline tool, and Table 9 contains user feedback on both *AMQuestioner* and the Baseline tool. Users perceived the tool's perceived usefulness ($U = 144.0, p < .0001$) and perceived ease of use ($U = 144.0, p < .0001$) to be significantly higher for *AMQuestioner*.

i) Positive Perceptions of *AMQuestioner*: *AMQuestioner* users praised the system for its ability to (1) Enhance their understanding of complex arguments, highlighting the value of argument maps in visualizing discussion structures and facilitating deeper analysis of diverse viewpoints; (2) Facilitate the exploration of diverse viewpoints, emphasizing the role of the "Explore Claim" feature in prompting them to consider alternative perspectives and avoid confirmation bias; (3) Offer a novel exercise format, making it easier to think critically through multiple-choice questions.

ii) Challenges with *AMQuestioner*: Some *AMQuestioner* users mentioned (1) Initial difficulty understanding argument maps, suggesting the need for clearer explanations or tutorials on how to interpret the visualizations; and (2) Occasional frustration with LLM responses, pointing out instances where the LLM-generated questions or feedback were perceived as irrelevant or unhelpful. This increased time commitment, however, is understandable given the richer features and functionalities offered by *AMQuestioner*, which encourage more in-depth exploration and analysis. "Sometimes the questions generated by the LLM didn't seem relevant to the argument I was looking at. It would be great if the questions could be more focused." (P19, *AMQuestioner* user).

iii) Perceptions of the Baseline tool: Baseline users generally found the tool easy to use and appreciated its simplicity. However, they also highlighted its limitations in terms of: **Limited engagement**, as simply reading Argument Maps proved difficult for in-depth understanding. For example, one user from the Baseline group noted, "The argument map was helpful for getting an overview, but I felt like I needed more guidance to really analyze the arguments." (P02, Baseline user).

8 Discussion

In this paper, we design and evaluate *AMQuestioner*, a interactive tool that leverage argument maps, data in online discussion, and educational elements like Socratic questioning for critical thinking training. Our work follows a general user-centered design process. First, to *specify the context of use*, we conducted a formative study that helps us understand the challenges and needs of our primary user group, university students, when engaging in critical thinking during online discussions. Second, based on the findings of formative study and related work, we *specify design requirements* of *AMQuestioner*. Third, we *create design solutions and develop AMQuestioner* with features accommodated to the design requirements. Lastly, we evaluate *AMQuestioner* via a mix-designed study with a baseline tool without our proposed key features. This human-centered design process enabled us to present *AMQuestioner* for improving users' critical thinking skills in online discussion, as discussed below.

As shown in Figure 6 and reported in Section 7.1, our mixed-design study with 24 participants revealed that both *AMQuestioner* and baseline tool with the generated argument maps significantly improve participants' performance in writing persuasive (metric: persuasiveness), clear (specificity-clarity), specific (specificity-justification), and strong (strength) arguments in online discussion after the training sessions. These findings support the efficacy of argument mapping in fostering critical thinking skills [35, 39, 72, 73] and extend this body of work by incorporating large language models (LLMs) to provide personalized guidance within the training process, which addresses a key limitation of traditional argument mapping instruction that often relies on handcrafted guidance. This personalized approach echoes research emphasizing the significance of tailored feedback and learning experiences in cultivating critical thinking [33].

Besides, participants' improvements on writing specific justifications (specificity-justification, *i.e.,* level of detail, depth and logical support for arguments) for their arguments and relevant (relevance, *i.e.,* how well statements address the main topic and claims) statements to the discussing topics after the training sessions are significantly higher in the *AMQuestioner* condition than that in the baseline condition without our proposed LLM-powered tutors. The differences between *AMQuestioner* and baseline tool indicate that *AMQuestioner*'s effectiveness could be attributed to its two unique components. First, the **LLM-powered "Explore Claim" tutor**, drawing on principles from scaffolding theory [41, 77] and Information Foraging Theory [61], can encourage exploration of diverse viewpoints, reduce information acquisition costs, and boost learner engagement. Second, the **tutor for customized critical thinking exercises** can reinforce participants' critical thinking thoughts, leading to more specific justification on their arguments. Our work contributes to the growing body of research on critical thinking support (*e.g.,* [60, 84, 89]) by demonstrating the value of LLMs in scaffolding critical thinking skills development, particularly in the context of participating in complex online discussions. However, we go beyond simply using LLMs for content delivery or feedback provision; we leverage their generative capabilities to dynamically create a personalized learning environment that adapts to the specific content and user interactions, fostering a deeper level of engagement and analysis with social media content.

It is important to highlight that the design and development of AMQuestioner were fundamentally guided by a human-centered approach, ensuring the resulting technology was grounded in the authentic needs and contexts of its intended users. Our process began not with the technology, but with the users themselves. The Formative Study (Section 3) served as the cornerstone of this approach. We explicitly aimed at understanding the real-world challenges and needs of our primary user group, university students, when engaging in critical thinking during online discussions. Through surveys (N=57) and interviews (N=5), we identified key difficulties, such as grappling with the volume and often problematic nature of online commentary (Challenge C1) and recognizing their own limitations in applying critical analysis skills effectively in this digital environment (Challenge C2). These empirically grounded insights gathered directly from potential users were then systematically translated into our core Design Requirements (DR1, DR2, DR3; Section 3.4). This crucial step ensured that the system's features – such as the automated argument map generation (addressing C1 and DR1), the LLM-powered viewpoint exploration (addressing C2 and DR2), and the customized critical thinking exercises (addressing C2 and DR3) – were conceived specifically to address user-identified needs and support their learning goals, rather than being solely technology-driven. Furthermore, the human-centered focus extended throughout our evaluation phase . While assessing the tool's effectiveness in improving critical thinking skills (RQ1) was a primary objective, we placed significant emphasis on understanding the user experience. This involved measuring perceived engagement and cognitive workload (RQ2ii) using established instruments and scales (detailed in Section 6.4, with results in Table 7), as well as gauging user perceptions regarding usability, ease of use, and overall usefulness (RQ3) through validated acceptance model questions and detailed qualitative feedback analysis (Section 6.4, Tables 7 and 8). Evaluating these aspects provides crucial insights not only into what the system achieves functionally but how it is perceived, interacted with, and experienced by the students it is designed to serve. This iterative process, moving from understanding user needs and challenges, to deriving design requirements, developing a targeted solution, and finally evaluating both effectiveness and user experience, underscores the human-centered foundation of the AMQuestioner research.

8.1 Implication of *AMQuestioner* for Learning Support Tools

The design of *AMQuestioner* draws heavily upon the educational theories of Constructivism and Socratic questioning, with the former emphasizing on the active and social interaction with learning materials and the later focusing on the guided and conversational reflection.

8.1.1 Constructivist Foundations in AMQuestioner. Constructivist theory emphasizes that learners are not passive recipients of information but actively construct their own knowledge and understanding through interaction with their environment and social contexts [77]. Learning is viewed as a social process where deeper understanding can be achieved within the "Zone of Proximal Development" (ZPD) through discussion and collaboration [77]. *AMQuestioner* embodies constructivist principles in the following ways:

Argument Mapping. With our developed computational models, *AMQuestioner* helps users better organize, understand, and analyze complex issues by visualizing argument structures. This visualization design decomposes the complex collaborative knowledge into small and more manageable argumentative components, which can facilitate users to explore diverse viewpoints on the discussing topics, as appreciated by our participants (Table 9).

Conversational Tutor. One key feature of *AMQuestioner* is the tutor for exploring others' claims and guiding critical thinking. Aligned with previous conversational agents (CAs) that support various learning tasks (*e.g.,* factual knowledge [64], programming concepts [81]), our user study indicates the benefits of CAs for engaging users in the learning tasks, *e.g.,* in terms of intrinsic motivation and timelessness feeling (Table 8).

8.1.2 Socratic Questioning and Scaffolding. Socratic questioning is a pedagogical method that uses sequential questioning to guide students towards deeper thinking and reflection. If we interpret the core of the Socratic method as the iterative process between questioning and answering, then this aspect serves as an effective form of scaffolding. *AMQuestioner* integrates this method with **guided reflection** and **timely feedback**. *AMQuestioner* generates guiding questions that prompt users to reflect on and analyze their ideas in depth [87]. Along with the questions, *AMQuestioner* also generates explanations for the answers to these questions (**??**), which enables it to provide timely feedback on users' responses. These two Socratic questioning features were praised by our participants in the user study (Table 9).

To sum up, *AMQuestioner* provides a good example with demonstrated benefits of integrating conversational tutor and interactive questioning into argument mapping, shedding lights into future learning support tools based on constructivist scaffolding and Socratic questioning.

8.2 Design Considerations

Two important design considerations emerged from our findings for critical thinking training tools: mitigating LLM over-reliance and balancing effectiveness with engagement.

Mitigating LLM Over-reliance: Excessive reliance on LLMs can undermine users' confidence in their own judgment [53, 86], leading them to favor quick, easy solutions over more thoughtful, analytical approaches. AMQuestioner addresses this by using critical thinking exercises to guide in-depth reasoning, rather than directly providing information. The observed improvements in relevance and specificity-justification, aspects that emphasize argument structure, demonstrate the effectiveness of this approach. Furthermore, recognizing the value of Socratic LLMs in this context, AMQuestioner offers a Socratic mode, encouraging users to engage in a deeper thought process rather than simply receiving direct answers from the LLM. Additionally, relevant research has shown that excessive use of Socratic questioning can lead to excessive cognitive load [13, 87].

Balancing Effectiveness and Engagement: Creating argument maps from scratch requires significant effort, potentially hindering learner motivation [61]. Furthermore, solely reading argument maps can yield less substantial learning gains compared to active writing [25]. As suggested by Scaffolding Theory [41, 77] and Information Foraging Theory [61], AMQuestioner lowers the barrier to entry by automatically generating initial Argument Maps and utilizes the "Explore Claim" feature to minimize information acquisition costs, incorporating elements of randomness to further encourage usage. The training process encourages users to engage in critical thinking exercises after they have a thorough understanding of the relevant content. Our results demonstrate a significant increase in "Explore Claim" usage, and users show less resistance to the exercises. Future work could explore gamification strategies, drawing inspiration from [88] and building on the principles outlined in [20], to further enhance engagement and motivation.

8.3 Generalizability

Using data in Reddit r/ChangeMyView (CMV) as a demonstration, we develop and evaluate *AMQuestioner* for training critical thinking in online discussions, which is an informal learning context outside classrooms. *AMQuestioner* could be directly applicable to other online communities (*e.g.,* r/Debate) in which members share and discuss competing opinions on hot topics, as do in the CMV community. It should be also usable in the general-purpose communities where active discussions on social events, such as the discussions on “safety issues in recent electric cars” in Zhihu (a Chinese community) and “tradeoffs of raising tariffs by USA” in Twitter. Our computational models could automatically extract the argumentative components from the discussion data, construct the argument maps, and prompt LLMs to act as a tutor to train readers’ critical thinking skills. Nevertheless, to tailor *AMQuestioner* to train critical thinking in other online discussions outside the CMV community, we should first validate if our models could perform well using a small sample of labeled discussion data as input. If our models do not perform well, researchers could follow our model development process (Section 4) to first label the “major claim”, “claim”, “premise” “support”, and “attack” in around 1000 comments, then fine-tune pretrained large language models to detect these argumentative components, and finally build up the argument maps.

Apart from online discussions, our idea of integrating argument maps and LLM-powered tutor has potentials to foster acquisition of critical thinking skills in other contexts. For example, to train paper reading skills, different from the summarization and question asking design in [84], a new tool could generate a argument map that visualize the argumentative components (*e.g.,* the claim can be a conclusion, the premise can be results of statistical tests) and their connections in the paper’s abstract, introduction, methods, results, and discussions. Similar to *AMQuestioner*, the tool could utilize the LLM to generate exercises tailor to the argumentative components that the learners select. These exercises could, for example, ask users to identify the core arguments of the paper, evaluate the reliability of the evidence presented, and analyze limitations of the research methods [60]. In formal learning contexts like those in classroom or professional training settings, human instructors can leverage *AMQuestioner* as a teaching tool. For example, instructors can assign students the same argument writing task surround a hot topic, allow them to explore online discussion about this topic with *AMQuestioner* for 20 minutes, and share their thoughts to other classmates for the rest 25 minutes in a lecture. With learning support tools like *AMQuestioner*, it is a promising future direction to incorporate online informal learning resources into the offline formal contexts to boost outcome and experience.

8.4 Limitations and Future Work

This study has limitations that warrant further investigation. **Participants’ Background**: We included only undergraduate students in the user study to evaluate *AMQuestioner*’s impact on the critical thinking training outcome and experience. While critical thinking is a crucial skill for this user group, it is also beneficial for other types of users such as high-school students, engineers in companies, writers in social media, and so on. Future work could examine the usefulness of *AMQuestioner* with more diverse user group, with the goal to embedding it into the social media platforms to benefit general users. **Baseline:** While the baseline in our experiment deducts LLM-powered tutors in *AMQuestioner* and enhance the argument map construction of Kialo community, it can not fully represent an existing tool that people have used for critical thinking. To verify whether and how *AMQuestioner* improve critical thinking training outcome and process against existing tools, future comparative studies should be conducted using baselines identical to designs like VISAR [89] and Kialo. **Online Discussion Platforms**: Our development and evaluation of *AMQuestioner* use the data from ChangeMyView, which is relatively structured and focused on argumentation. While *AMQuestioner* has potentials to be generalized to other less-structured online platforms like Zhihu and Quora as discussed in Section 8.3, future efforts should be made to first structure the data, *e.g.,* label and classify the sentences related to the same discussing topic. Future work should evaluate *AMQuestioner*’s usefulness in training critical thinking in the contexts of these less-structured online discussions. **Assessment Validity**: Our evaluation relied primarily on expert-rated writing samples and self-reported measures. While our inter-rater reliability

(Cohen's $\kappa \approx 0.82$) demonstrates consistent expert judgment, this approach may not capture the full spectrum of critical thinking abilities. The reliance on self-reported abilities introduces potential social desirability bias, and our behavioral measure of open-mindedness (clicking "Explore Claim") may not fully capture the complexity of open-minded thinking. Future research should incorporate standardized critical thinking assessments (e.g., Watson-Glaser Critical Thinking Appraisal) to complement expert evaluation and strengthen construct validity.

Future work can also extend our *AMQuestioner* to address its following limitations. **Cognitive Load:** While engaging and motivating, the increased cognitive load (particularly mental and temporal demands) associated with AMQuestioner could pose challenges for some learners. Future iterations should explore strategies to mitigate cognitive overload, such as developing adaptive scaffolding mechanisms within AMQuestioner. Such mechanisms could personalize support based on individual user needs and learning progress. This might involve dynamically adjusting argument map complexity, LLM prompt types and frequency, and the difficulty of critical thinking exercises based on user performance and engagement patterns. Additionally, personalization could extend to pacing, tailoring the learning experience to individual cognitive abilities and learning styles through user modeling techniques. **Collaborative Learning:** Future research should investigate the use of AMQuestioner in collaborative learning environments, allowing students to co-construct argument maps and engage in peer discussions to further enhance their critical thinking skills. This might involve integrating features that support real-time collaboration, such as shared argument map editing and discussion forums, allowing students to learn from each other and improve their critical thinking through social interaction [50, 68]. **Real-World Deployment Studies:** To demonstrate AMQuestioner's broader societal impact and practical importance, future research should examine its effectiveness in diverse real-world contexts. This could include longitudinal studies tracking users' critical thinking development with *AMQuestioner* over extended periods (6-12 months) in online platforms or in offline courses guided by human teachers. Such studies would provide concrete evidence of *AMQuestioner*'s potential to improve online discourse quality and combat misinformation in authentic social and classroom settings. **Real-World Deployment Studies:** To demonstrate AMQuestioner's broader societal impact and practical importance, future research should examine its effectiveness in diverse real-world contexts. This could include longitudinal studies tracking users' critical thinking development with *AMQuestioner* over extended periods (6-12 months) in online platforms or in offline courses guided by human teachers. Such studies would provide concrete evidence of *AMQuestioner*'s potential to improve online discourse quality and combat misinformation in authentic social and classroom settings. **Credibility of AI-Generated Questions**: While our prompt engineering approach showed promising results, AI-generated content may contain inherent biases or cultural assumptions from training data. GPT-4's responses could potentially favor certain argumentative styles or cultural perspectives, which may not equally support diverse learning approaches. Additionally, our single-turn prompting approach may lead to repetitive question patterns. Future iterations should incorporate bias detection mechanisms, expert review of generated content, and multi-turn dialogue systems to ensure equitable and robust learning experiences.

9 Conclusion

In this paper, we design, develop, and evaluate *AMQuestioner*, a novel interactive tool that integrate of the argument map, the AI-enabled educational elements (*e.g.,* socratic questioning), and rich online discussion data to cultivate users' critical thinking skills in online discussion. Our user study demonstrates the effectiveness of *AMQuestioner* in improving participants' performance in independently writing detailed, specific, and relevant arguments to the discussing topics. Our work has implications for future learning support tools related to critical thinking, large language models, and applications of social media data as learning materials.

10 Acknowledgement

This work is supported by the Young Scientists Fund of the National Natural Science Foundation of China with Grant No. 62202509 and the General Projects Fund of the Natural Science Foundation of Guangdong Province in China with Grant No. 2024A1515012226.

References

[1] Zahra Abbasiantaeb, Yifei Yuan, Evangelos Kanoulas, and Mohammad Aliannejadi. 2024. Let the LLMs Talk: Simulating Human-to-Human Conversational QA via Zero-Shot LLM-to-LLM Interactions. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining* (Merida, Mexico) *(WSDM '24)*. Association for Computing Machinery, New York, NY, USA, 8–17. https://doi.org/10.1145/3616855.3635856

[2] Nevin Akkaya and M Volkan Demirel. 2012. Teacher candidates' use of questioning skills in during-reading and post-reading strategies. *Procedia-Social and Behavioral Sciences* 46 (2012), 4301–4305.

[3] Nada J. Alsaleh. 2020. Teaching Critical Thinking Skills: Literature Review. *Turkish Online Journal of Educational Technology* 19 (2020), 21–39. https://api.semanticscholar.org/CorpusID:211054783

[4] Giorgio Biancini, Alessio Ferrato, and Carla Limongelli. 2024. Multiple-Choice Question Generation Using Large Language Models: Methodology and Educator Insights. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization* (Cagliari, Italy) *(UMAP Adjunct '24)*. Association for Computing Machinery, New York, NY, USA, 584–590. https://doi.org/10.1145/3631700.3665233

[5] Frederick Bigenho, PHD Stephen Enwefa, and MFRT CNHP. 2009. THE USE OF PRE-READING, DURING READING AND POST READING STRATEGIES THAT MATCH STUDENTS'PRIMARY LEARNING STYLES: VISUAL, AUDITORY OR KINESTHETIC TO INCREASE COMPREHENSION OF CONTENT AREA READING. In *NAAAS Conference Proceedings*. National Association of African American Studies, 209.

[6] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[7] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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. *CoRR* abs/2005.14165 (2020). arXiv:2005.14165 https://arxiv.org/abs/2005.14165

[8] Tuan Bui, Oanh Tran, Phuong Nguyen, Bao Ho, Long Nguyen, Thang Bui, and Tho Quan. 2024. Cross-Data Knowledge Graph Construction for LLM-enabled Educational Question-Answering System: A Case Study at HCMUT. In *Proceedings of the 1st ACM Workshop on AI-Powered Q&A Systems for Multimedia*. 36–43.

[9] Sam Butchart, Daniella Forster, Ian Gold, John Bigelow, Kevin Korb, Graham Oppy, and Alexandra Serrenti. 2009. Improving critical thinking using web based argument mapping exercises with automated feedback. *Australasian Journal of Educational Technology* 25, 2 (2009).

[10] James P Byrnes and Kevin N Dunbar. 2014. The nature and development of critical-analytic thinking. *Educational Psychology Review* 26 (2014), 477–493.

[11] Jérémie Cabessa, Hugo Hernault, and Umer Mushtaq. 2024. In-Context Learning and Fine-Tuning GPT for Argument Mining. arXiv:2406.06699 [cs.CL] https://arxiv.org/abs/2406.06699

[12] Chen Cao. 2023. Scaffolding CS1 Courses with a Large Language Model-Powered Intelligent Tutoring System. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 229–232. https://doi.org/10.1145/3581754.3584111

[13] Timothy A Carey and Richard J Mullan. 2004. What is Socratic questioning? *Psychotherapy: theory, research, practice, training* 41, 3 (2004), 217.

[14] W Carlile, N Gurrapadi, Z Ke, and V Ng. 2018. Give me more feedback: Annotating argument persuasiveness and related attributes in student essays. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 621–631.

[15] Lacey Colligan, Henry WW Potts, Chelsea T Finn, and Robert A Sinkin. 2015. Cognitive workload changes for nurses transitioning from a legacy system with paper documentation to a commercial electronic health record. *International journal of medical informatics* 84, 7 (2015), 469–476.

[16] Mihaly Czikszentmihalyi. 1990. *Flow: The psychology of optimal experience.* New York: Harper & Row.

[17] Martin Davies. 2011. Concept mapping, mind mapping and argument mapping: what are the differences and do they matter? *Higher education* 62 (2011), 279–301.

[18] Martin Davies and Ronald Barnett (Eds.). 2015. *The Palgrave Handbook of Critical Thinking in Higher Education.* Palgrave Macmillan, New York. X, 646 pages. https://doi.org/10.1057/9781137378057

[19] W. Davies and M. Stevens. 2019. The importance of critical thinking and how to measure it. https://www.talentlens.co.uk/wpcontent/uploads/sites/5/The-Importance-of-Critical-Thinking-and-How-to-Measure-It_UK_Final.pdf. Retrieved from Pearson TalentLens website.

[20] 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.* 9–15.

[21] J. Dewey. 2022. *How we think.* DigiCat.

[22] T Dingler, B Tag, P Lorenz-Spreen, AW Vargo, S Knight, and S Lewandowsky. 2021. Workshop on technologies to support critical thinking in an age of misinformation. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems.* 1–5.

[23] Aline Duelen, Iris Jennes, and Wendy Van den Broeck. 2024. Socratic AI Against Disinformation: Improving Critical Thinking to Recognize Disinformation Using Socratic AI. In *Proceedings of the 2024 ACM International Conference on Interactive Media Experiences* (Stockholm, Sweden) *(IMX '24).* Association for Computing Machinery, New York, NY, USA, 375–381. https://doi.org/10.1145/3639701.3663640

[24] A Duelen, I Jennes, and W Van den Broeck. 2024. Socratic AI Against Disinformation: Improving Critical Thinking to Recognize Disinformation Using Socratic AI. In *Proceedings of the 2024 ACM International Conference on Interactive Media Experiences.* 375–381.

[25] Christopher P Dwyer, Michael J Hogan, and Ian Stewart. 2010. The evaluation of argument mapping as a learning tool: Comparing the effects of map reading versus text reading on comprehension and recall of arguments. *Thinking Skills and Creativity* 5, 1 (2010), 16–22.

[26] R. H. Ennis. 1987. Critical thinking and the curriculum. , 40-48 pages.

[27] P. A. Facione. 2000. The disposition toward critical thinking: Its character, measurement, and relationship to critical thinking skill. *Informal logic* 20, 1 (2000).

[28] P. A. Facione. 2011. Critical thinking: What it is and why it counts. *Insight assessment* 1, 1 (2011), 1–23.

[29] Peter A Facione and Carol Ann Gittens. 2015. Mapping Decisions and Arguments. *Inquiry: Critical Thinking Across the Disciplines* 30, 2 (2015), 17–53.

[30] J. H. Flavell. 2013. Perspectives on perspective taking. In *Piaget's theory.* Psychology Press, 107–139.

[31] T González-Cacho and A Abbas. 2022. Impact of Interactivity and Active Collaborative Learning on Students' Critical Thinking in Higher Education. *Rev Iberoam de Tecnol del Aprendiz* 17, 3 (2022), 254–261.

[32] Reto Gubelmann, Michael Burkhard, Rositsa V Ivanova, Christina Niklaus, Bernhard Bermeitinger, and Siegfried Handschuh. 2024. Exploring the Usefulness of Open and Proprietary LLMs in Argumentative Writing Support. In *International Conference on Artificial Intelligence in Education.* Springer, 175–182.

[33] Diane F Halpern. 1998. Teaching critical thinking for transfer across domains: Disposition, skills, structure training, and metacognitive monitoring. *American psychologist* 53, 4 (1998), 449.

[34] D. F. Halpern. 2013. *Thought and knowledge: An introduction to critical thinking.* Psychology press.

[35] Mara Harrell. 2008. No computer program required: Even pencil-and-paper argument mapping improves critical thinking skills. (2008).

[36] Yueh-Ren Ho, Bao-Yu Chen, and Chien-Ming Li. 2023. Thinking more wisely: using the Socratic method to develop critical thinking skills amongst healthcare students. *BMC medical education* 23, 1 (2023), 173.

[37] Peter Hoonakker, Pascale Carayon, Ayse P Gurses, Roger Brown, Adjhaporn Khunlertkit, Kerry McGuire, and James M Walker. 2011. Measuring workload of ICU nurses with a questionnaire survey: the NASA Task Load Index (TLX). *IIE transactions on healthcare systems engineering* 1, 2 (2011), 131–143.

[38] Md Naimul Hoque, Ayman A Mahfuz, Mayukha Sridhatri Kindi, and Naeemul Hassan. 2024. Towards Designing a Question-Answering Chatbot for Online News: Understanding Questions and Perspectives. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24).* Association for Computing Machinery, New York, NY, USA, Article 154, 17 pages. https://doi.org/10.1145/3613904.3642007

[39] Kristi Kaeppel. 2021. The influence of collaborative argument mapping on college students' critical thinking about contentious arguments. *Thinking Skills and Creativity* 40 (2021), 100809.

[40] Majeed Kazemitabaar, Runlong Ye, Xiaoning Wang, Austin Zachary Henley, Paul Denny, Michelle Craig, and Tovi Grossman. 2024. Codeaid: Evaluating a classroom deployment of an llm-based programming assistant that balances student and educator needs. In *Proceedings of the CHI Conference on Human Factors in Computing Systems.* 1–20.

[41] Beaumie Kim. 2001. Social constructivism. *Emerging perspectives on learning, teaching, and technology* 1, 1 (2001), 16.

[42] Emily Kuang, Minghao Li, Mingming Fan, and Kristen Shinohara. 2024. Enhancing UX Evaluation Through Collaboration with Conversational AI Assistants: Effects of Proactive Dialogue and Timing. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 3, 16 pages. https://doi.org/10.1145/3613904.3642168

[43] D. Kuhn. 1999. A developmental model of critical thinking. *Educational researcher* 28, 2 (1999), 16–46.

[44] Marie Larochelle, Nadine Bednarz, and James W Garrison. 1998. *Constructivism and education.* Cambridge University Press.

[45] Yoonjoo Lee, Tae Soo Kim, Sungdong Kim, Yohan Yun, and Juho Kim. 2023. DAPIE: Interactive Step-by-Step Explanatory Dialogues to Answer Children's Why and How Questions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 450, 22 pages. https://doi.org/10.1145/3544548.3581369

[46] Yinheng Li. 2023. A Practical Survey on Zero-shot Prompt Design for In-context Learning. In *Proceedings of the Conference Recent Advances in Natural Language Processing - Large Language Models for Natural Language Processings (RANLP)*. INCOMA Ltd., Shoumen, BULGARIA, 641–647. https://doi.org/10.26615/978-954-452-092-2_069

[47] Chengzhong Liu, Shixu Zhou, Dingdong Liu, Junze Li, Zeyu Huang, and Xiaojuan Ma. 2023. CoArgue: Fostering Lurkers' Contribution to Collective Arguments in Community-based QA Platforms. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 271, 17 pages. https://doi.org/10.1145/3544548.3580932

[48] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR* abs/1907.11692 (2019). arXiv:1907.11692 http://arxiv.org/abs/1907.11692

[49] Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics* (1947), 50–60.

[50] E Mazur. 1997. Peer instruction: A user´ s Manual, Universidad de Harvard.

[51] Mike Metcalfe and Saras Sastrowardoyo. 2013. Complex project conceptualisation and argument mapping. *International Journal of Project Management* 31, 8 (2013), 1129–1138.

[52] Raymond S Nickerson. 1988. Chapter 1: On improving thinking through instruction. *Review of research in education* 15, 1 (1988), 3–57.

[53] J. Noyes. 2007. Automation and decision making. In *Decision Making in Complex Environments*. CRC Press, New York, 73–82. https://doi.org/10.1201/9781315576138-7

[54] James C Overholser. 1994. Elements of the Socratic method: III. Universal definitions. *Psychotherapy: Theory, Research, Practice, Training* 31, 2 (1994), 286.

[55] Heather O'Brien and Paul Cairns. 2016. Why engagement matters. *Cham: Springer International Publishing. doi* 10 (2016), 978–3.

[56] Richard Paul. 1991. Critical thinking: What every person needs to survive in a changing world. *Nassp Bulletin* 75, 533 (1991), 120–122.

[57] Richard Paul and Linda Elder. 2007. Critical thinking: The art of Socratic questioning. *Journal of developmental education* 31, 1 (2007), 36.

[58] Zhenhui Peng, Qiaoyi Chen, Zhiyu Shen, Xiaojuan Ma, and Antti Oulasvirta. 2024. DesignQuizzer: A Community-Powered Conversational Agent for Learning Visual Design. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 44 (apr 2024), 40 pages. https://doi.org/10.1145/3637321

[59] Zhenhui Peng, Qingyu Guo, Ka Wing Tsang, and Xiaojuan Ma. 2020. Exploring the Effects of Technological Writing Assistance for Support Providers in Online Mental Health Community. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. ACM, New York, NY, USA, 556–567. https://doi.org/10.1145/3313831.3376695

[60] Zhenhui Peng, Yuzhi Liu, Hanqi Zhou, Zuyu Xu, and Xiaojuan Ma. 2022. CReBot: Exploring Interactive Question Prompts for Critical Paper Reading. *Int. J. Hum.-Comput. Stud.* 167, C (nov 2022), 17 pages. https://doi.org/10.1016/j.ijhcs.2022.102898

[61] Peter Pirolli and Stuart Card. 1999. Information foraging. *Psychological review* 106, 4 (1999), 643.

[62] Nils Reimers, Benjamin Schiller, Tilman Beck, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2019. Classification and clustering of arguments with contextualized word embeddings. *arXiv preprint arXiv:1906.09821* (2019).

[63] Laria Reynolds and Kyle McDonell. 2021. Prompt Programming for Large Language Models: Beyond the Few-Shot Paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 314, 7 pages. https://doi.org/10.1145/3411763.3451760

[64] Sherry Ruan, Liwei Jiang, Justin Xu, Bryce Joe-Kun Tham, Zhengneng Qiu, Yeshuang Zhu, Elizabeth L. Murnane, Emma Brunskill, and James A. Landay. 2019. QuizBot: A Dialogue-based Adaptive Learning System for Factual Knowledge. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. https://doi.org/10.1145/3290605.3300587

[65] Christian Stab and Iryna Gurevych. 2014. Annotating Argument Components and Relations in Persuasive Essays. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, Junichi Tsujii and Jan Hajic (Eds.). Dublin City University and Association for Computational Linguistics, Dublin, Ireland, 1501–1510. https://aclanthology.org/C14-1142

[66] Nadirah Julia Ulfah Tanjung, Yani Lubis, and Ernita Daulay. 2022. The Effect of Pre-Reading, During Reading, and Post Reading Activities to Monitor Students' Comprehension in Reading Narrative Text. *INSPIRATION (Instructional Practices in Language Education)* 1, 2 (2022), 16–30.

[67] Thitaree Tanprasert, Sidney S Fels, Luanne Sinnamon, and Dongwook Yoon. 2024. Debate Chatbots to Facilitate Critical Thinking on YouTube: Social Identity and Conversational Style Make A Difference. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 805, 24 pages. https://doi.org/10.1145/3613904.3642513

[68] Neil Thomason. 1990. Making Student Groups Work:"To teach is to learn twice". *Teaching Philosophy* 13, 2 (1990), 111–125.

[69] Dawit Tibebu Tiruneh, An Verburgh, and Jan Elen. 2014. Effectiveness of critical thinking instruction in higher education: A systematic review of intervention studies. *Higher Education Studies* 4, 1 (2014), 1–17.

[70] Dawit Tibebu Tiruneh, A P Verburgh, and Jan Elen. 2013. Effectiveness of Critical Thinking Instruction in Higher Education: A Systematic Review of Intervention Studies. *Higher Education Studies* 4 (2013), 1–17. https://api.semanticscholar.org/CorpusID:53420638

[71] SE Toulmin. 2003. *The uses of argument.* Cambridge university press.

[72] Charles Twardy. 2004. Argument maps improve critical thinking. (2004).

[73] Tim van Gelder. 2015. Using argument mapping to improve critical thinking skills. In *The Palgrave handbook of critical thinking in higher education*. Springer, 183–192.

[74] Tim Van Gelder, Melanie Bissett, and Geoff Cumming. 2004. Cultivating expertise in informal reasoning. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale* 58, 2 (2004), 142.

[75] Viswanath Venkatesh and Hillol Bala. 2008. Technology acceptance model 3 and a research agenda on interventions. *Decision sciences* 39, 2 (2008), 273–315.

[76] S. Vincent-Lancrin, C. González-Sancho, M. Bouckaert, F. De Luca, M. Fernández-Barrerra, G. Jacotin, ..., and Q. Vidal. 2019. *Fostering Students' Creativity and Critical Thinking: What It Means in School.* OECD Publishing, 2, rue Andre Pascal, F-75775 Paris Cedex 16, France.

[77] Lev Semenovich Vygotsky. 1978. *Mind in society: The development of higher psychological processes*. Vol. 86. Harvard university press.

[78] Thiemo Wambsganss, Tobias Kueng, Matthias Soellner, and Jan Marco Leimeister. 2021. ArgueTutor: An Adaptive Dialog-Based Learning System for Argumentation Skills. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 683, 13 pages. https://doi.org/10.1145/3411764.3445781

[79] Thiemo Wambsganss, Christina Niklaus, Matthias Cetto, Matthias Söllner, Siegfried Handschuh, and Jan Marco Leimeister. 2020. AL: An Adaptive Learning Support System for Argumentation Skills. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376732

[80] Goodwin Watson. 1980. *Watson-Glaser critical thinking appraisal.* Psychological Corporation San Antonio, TX.

[81] Rainer Winkler, Sebastian Hobert, Antti Salovaara, Matthias Söllner, and Jan Marco Leimeister. 2020. Sara, the Lecturer: Improving Learning in Online Education with a Scaffolding-Based Conversational Agent. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3313831.3376781

[82] Meng Xia, Qian Zhu, Xingbo Wang, Fei Nie, Huamin Qu, and Xiaojuan Ma. 2022. Persua: A Visual Interactive System to Enhance the Persuasiveness of Arguments in Online Discussion. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 319 (Nov. 2022), 30 pages. https://doi.org/10.1145/3555210

[83] Saelyne Yang, Jo Vermeulen, George Fitzmaurice, and Justin Matejka. 2024. AQuA: Automated Question-Answering in Software Tutorial Videos with Visual Anchors. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 928, 19 pages. https://doi.org/10.1145/3613904.3642752

[84] Kangyu Yuan, Hehai Lin, Shilei Cao, Zhenhui Peng, Qingyu Guo, and Xiaojuan Ma. 2023. CriTrainer: An Adaptive Training Tool for Critical Paper Reading. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software*

and Technology (<conf-loc>, <city>San Francisco</city>, <state>CA</state>, <country>USA</country>, </conf-loc>) *(UIST '23)*. Association for Computing Machinery, New York, NY, USA, Article 44, 17 pages. https://doi.org/10.1145/3586183.3606816

[85] J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why Johnny Can't Prompt: How Non-AI Experts Try (and Fail) to Design LLM Prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 437, 21 pages. https://doi.org/10.1145/3544548.3581388

[86] C Zhai, S Wibowo, and LD Li. 2024. The effects of over-reliance on AI dialogue systems on students' cognitive abilities: a systematic review. *Smart Learning Environments* 11, 1 (2024), 28.

[87] Liang Zhang, Jionghao Lin, Ziyi Kuang, Sheng Xu, and Xiangen Hu. 2024. SPL: A Socratic Playground for Learning Powered by Large Language Model. arXiv:2406.13919 [cs.AI] https://arxiv.org/abs/2406.13919

[88] 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 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 484, 24 pages. https://doi.org/10.1145/3613904.3642545

[89] Zheng Zhang, Jie Gao, Ranjodh Singh Dhaliwal, and Toby Jia-Jun Li. 2023. VISAR: A Human-AI Argumentative Writing Assistant with Visual Programming and Rapid Draft Prototyping. 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 5, 30 pages. https://doi.org/0.1145/3586183.3606800

A Follow-up Study Participants

To evaluate the effectiveness of our proposed approach, we conducted Follow-up Study with 11 participants from diverse academic backgrounds. The participants were selected based on their perceived critical thinking abilities and prior experience with argumentation. Table 10 presents the demographic information and key characteristics of the interview participants.

B Argument Map User Manual

B.1 Introduction to Argument Mapping

An argument map is a visual representation of reasoning and argumentation that helps users analyze and construct logical arguments. By displaying the relationships between claims and evidence in a structured format, argument maps make complex reasoning processes clear and explicit.

B.2 Benefits for Critical Thinking

Argument mapping enhances critical thinking skills by:

- Visualizing logical relationships between ideas
- Identifying strengths and weaknesses in arguments
- Developing systematic reasoning abilities
- Improving argument evaluation skills
- Facilitating better understanding of complex arguments
- Supporting construction of well-reasoned arguments

B.3 Components and Relationships

The argument map consists of two fundamental elements: argument components and argument relations. Table 11 presents the taxonomy of argument components that form the basic building blocks of an argument map, while Table 12 describes the possible relationships between these components.

B.4 User Operations

B.4.1 Basic Node Operations.

Table 10. Participant Demographics and Characteristics

ID	Age	Gender	Major	Academic qualifications	Perceived ability in critical thinking	Prior exp.
U1	22	Male	Computer Science	Undergraduate	Strong	Yes
U2	23	Female	Computer Science	Undergraduate	Above average	Yes
U3	24	Undefined	Computer Science	Master's Degree	Strong	Yes
U4	21	Female	Computer Science	Undergraduate	Above average	Yes
U5	23	Male	Computer Science	Bachelor's Degree	Strong	Yes
U6	22	Female	Computer Science	Undergraduate	Above average	Yes
U7	24	Male	Philosophy	Undergraduate	Strong	Yes
U8	27	Female	Philosophy	Doctoral Degree	Above average	Yes
U9	23	Male	Chinese Literature	Undergraduate	Strong	Yes
U10	25	Female	Chinese Literature	Bachelor's Degree	Above average	Yes
U11	24	Male	Atmospheric Science	Undergraduate	Strong	Yes

Add Child Node Creates a new component that relates to the selected component
Add Same Node Adds a new component at the same hierarchical level
Remove Node Deletes the selected component and its connections

B.4.2 Setting Relations.

Set to Support Establishes a supporting relationship from one component to another
Set to Objection Creates an attacking relationship between components
Set to Premise Marks a component as a premise (basic evidence or reasoning)

C A Computational Workflow for Modeling Comments into Argument Map

C.1 Data Filtering

Our data filtering approach focused on eliminating semantically redundant comments to enhance the clarity and structure of argument maps. We employed the roberta-base model, a powerful transformer-based language model pre-trained on a massive text corpus, to generate semantically rich sentence embeddings. These embeddings capture the meaning and contextual information of each comment, allowing for a nuanced comparison of their semantic similarity. We then applied

Table 11. Taxonomy of Argument Component

Name	**Definition**	**Example**
Major Claim	The root node of the argumentation structure. It's the author's main standpoint or opinion on the topic.	This is a terrible idea
Claim	Claims are secondary conclusions or viewpoints that support the major claim.	It creates an incentive that the government wants you to die at 75 years old.
Premise	Premises are the underlying facts, evidence, or reasoning that support claims or the main claim.	Humans and governments work almost exclusively at their most basic on incentive.
Non-argument	Non-argument statements are sentences within an argument that do not clearly function as a major claim, claim, or premise.	So what incentive does this create?

Table 12. Taxonomy of Argument Relations between Argument Components

Name	**Definition**	**Example**
Support	Support refers to one argument providing evidence or reasoning that strengthens or bolsters the claim of another argument.	"Humans and governments work almost exclusively at their most basic on incentive." **support** "It creates an incentive that the government wants you to die at 75 years old"
Attack	Attack refers to one argument attempting to weaken or refute the validity or credibility of another argument.	"So, at least 40% of people do not have a desire to pass away" **Attack** " They are merely expressing their frustrations; their true desire is to pass away"

cosine similarity calculations with a threshold of 0.95 to identify highly similar comments within the same hierarchical level of discussion threads. This threshold ensures that only comments with a very high degree of semantic overlap are considered redundant.

Table 13 illustrates our filtering process with representative examples from the "CMV: A majority of double standards exist only because we assume the world is equal" discussion thread. The table presents excerpts of removed comments alongside their similar counterparts that were retained, demonstrating how our pre-processing approach identifies and eliminates semantic redundancy while preserving the most comprehensive expressions of arguments.

Our filtering method was applied to all 14 scraped posts, with Table 14 presenting the results from five representative examples. The number of redundant comments identified and removed ranged from 16 to 28 per discussion. Among these examples, discussions such as "CMV: A majority of double standards exist only because we assume the world is equal" and "CMV: It is never acceptable/ok to make fun of ANYONE'S appearance" had the highest number of redundant comments removed (25 each), while maintaining the core arguments and diverse viewpoints of the discussion.

Table 13. Excerpts from longer comments removed due to high similarity with other comments at the same level within the discussion thread, along with explanations for their removal.

Removed Comment (Excerpt)	**Similar Comment (Excerpt)**	**Removal Reason**
"Royalty, Nobility, and Military - it was quite common... Women don't wear heels because they want to ride horses or avoid stepping over blood, clearly the purpose of using them are different"	"There is a difference between positive (is) statements, and normative (ought) statements... double standards are identified when we believe normatively that something ought be equal when it is not."	Both comments discuss the historical context of double standards, but the "Similar Comment" offers a more concise and generalized explanation.
"What's the difference between the double standard it's futile to change... That's 'the way humanity has defined itself'. What makes it more worth changing in your view?"	"Your premise is contradictory. If everyone assumed that everyone was equal, there would be no double standards... Might as well stay at home and bear children, have no job, and be powerless in an abusive marriage."	Both comments challenge the premise of the original post, but the "Similar Comment" provides a more impactful illustration of the consequences of accepting double standards.
"Your premise is contradictory... While you're at it, tell those Jews living under Nazi rule not to bother, since resisting antisemitic double standards is futile."	"What's the difference between the double standard it's futile to change... That's 'the way humanity has defined itself'. What makes it more worth changing in your view?"	Both comments question the practicality of challenging ingrained double standards, but the "Similar Comment" frames the issue in a broader and less emotionally charged manner.

C.2 Annotation Example

During the annotation process, several challenging cases emerged that required careful consideration and discussion among annotators. One representative example involved the sentence "Humans and governments work almost exclusively at their most basic on incentive." This case highlighted the complexity in distinguishing between claims and premises in argumentative structures.

Initially, three annotators labeled it as a claim", interpreting it as a standalone assertion about human and governmental behavior. However, one annotator classified it as a premise", noting that it

Table 14. Number of comments removed during pre-processing for five example posts

Title (abbreviated)	Original Count	Filtered Count	Removed Count
CMV: A majority of double standards exist only because we assume the world is equal.	107	82	25
CMV: world Hunger is not from a lack of food but a miss allocation of food.	201	173	28
CMV: young adults today do not live in the best time ever, and it will only get worse.	192	168	24
CMV: In densely populated countries like the UK, it's very difficult to justify significant amounts of land being dedicated to land intensive recreation like golf courses, horse riding or dedicated hunting grounds.	162	146	16
CMV: It is never acceptable/ok to make fun of ANYONE'S appearance.	224	199	25

provided supporting evidence for the subsequent claim "It creates an incentive that the government wants you to die at 75 years old."

The disagreement stemmed from differing interpretations of claim" and premise" within the context of the comment. After thorough discussion, the team determined that the sentence functioned as a foundational statement supporting subsequent arguments rather than a standalone position. This led to a guideline revision that clarified the distinction between claims and premises, emphasizing the importance of examining logical relationships between sentences in context. The sentence was ultimately annotated as a "premise", exemplifying how contextual dependencies influence argument component classification.

This case served as a valuable reference point in developing our annotation guidelines and illustrates the nuanced decision-making process in argument annotation.

C.3 Guiding for Data Annotation

The final annotation guideline including the follow: (1) For sentence splitting, prioritize punctuation-based splits, but also split sentences where meaning is complete even without terminal punctuation. If there is any ambiguity regarding the necessity of a split, the issue should be discussed collaboratively. (2) Sentences are labeled based on the definition given in table 2. (3) Sentence labels should consider both the context of the individual comment and the context of the entire discussion thread. (4) One sentence can only be labeled as one of the argument component. (5) Generally, Major Claim should be labeled only once. (6) Claims in comments often stand alone, while premises tend to follow and elaborate on the claim. (7) Context and audience interpretation must be considered when labeling comments as nonsense. Based on the annotation guideline, three annotators further labeled all the data. (8) A shared Notion page documents relevant domain-specific knowledge and terminology to ensure consistent understanding among annotators; this knowledge base is updated iteratively after each annotation round to maintain accuracy.

C.4 The Prompt in *Prototype*

C.4.1 The prompt of Argument Component Extraction.

> **Role:** You are an expert in argument mining, skilled at identifying a sentence's argument element based on definition and context.
> **Action:** You will be provided with a sentence and a comment (context) indicating where the sentence belongs within a larger argument. Your task is to analyze the sentence and classify it as one of the following argument elements: Major Claim, Claim, Premise, or Non-argument, based on the provided definitions: <Same definition in table 2>
> **Context:** The input will consist of a "sentence" and a "comment". The "comment" provides context by showing where the sentence fits within a larger argument.
> **Expectation:** Output the result in JSON format without code block markers:{ "sentence": "...", "argument element": "..." }

C.4.2 The prompt of Argument Relation Detection.

> **Role:** You are an expert in discourse analysis, skilled at identifying relationships between sentences based on support or attack.
> **Action:** You will be provided with two sentences: "sentence1" and "sentence2." Your task is to analyze the relationship between the two sentences and classify it as either "Support" or "Attack."
> **Context:** The input will consist of "sentence1" and "sentence2." Use these to determine how "sentence2" relates to "sentence1."
> **Expectation:** Output the result in JSON format without code block markers: { "sentence1": "...", "sentence2": "...", "relation": "Support" or "Attack" }

C.5 Fine-tuning Technical Details

For fine-tuning, we first convert our labeled dataset into a fine-tuning dataset using a custom script. The format of each conversation in the dataset should follow the guidelines for Azure OpenAI's fine-tuning API.[7]. After validating the data format, we upload the file to Azure Blob Storage and reference its location when creating a fine-tuning job within the Azure OpenAI Service. Following standard practices for model evaluation, we employed a stratified split strategy: 60% for training, 20% for validation, and 20% for testing, ensuring consistent distribution of different argument components and relations across all sets. We employed GPT-3.5-turbo as our base model and configured the training parameters with careful consideration of our dataset characteristics: batch size of 8, learning rate of 1e-5, and warmup steps of 100. The model was trained for 4 epochs with early stopping mechanism implemented to prevent overfitting. The fine-tuning process, which took approximately 2 hours, yielded promising results across both tasks. For Argument Component Detection, the model achieved F1 scores of 0.77. In Argument Relation Detection, the model demonstrated performance with F1 score of 0.76. The training convergence analysis showed that the loss stabilized during the third epoch, with optimal validation performance achieved in the latter part of epoch 3.

[7] https://learn.microsoft.com/en-us/azure/ai-services/openai/tutorials/fine-tune

C.6 Evaluation for Argument Mining Models

We evaluated argument component and relation extraction using three approaches. First, following the methodology of Persua, we leveraged BERT for feature extraction from the training data. These features were then input into several classic machine learning models (Logistic Regression, Linear SVM, RBF SVM, Random Forest, Gaussian Naive Bayes, Nearest Neighbour, Adaboost Decision Tree). A 5-fold stratified cross-validation was performed to evaluate and compare these models, using the weighted average F1-score as the primary metric (see Table 4 Table 5). Second, we explored the capabilities of latest large language models (GPT-4o) in a few-shot learning setting. To investigate the impact of the number of provided examples, we experimented with varying the number of labeled examples used to guide the model's predictions. Specifically, we tested sets of 4, 8, 12, and 16 examples (for relation detection are 2, 4, 6, and 8 examples). For each task, we selected the results obtained using the number of few-shot examples that yielded the best performance. The reported performance for Few-shot GPT-4o represents the average performance across three different random seeds for the best-performing example size. Finally, for Instruction-tuned GPT-4o, the average performance across three different random seeds is also reported.

D Explore Topic with LLM-powered agent

D.1 Prompt of Generating Pre-define Question for Topic Exploration

Role: You are a question generation expert specialized in creating comprehensive exploration questions for discussion topics.
Task: Generate 5-6 pre-defined questions that cover the core aspects of the provided discussion thread.
Input Format: - A discussion thread in JSON format containing comments and replies - Each comment includes: name, comment text, and nested subcomments
Requirements for Generated Questions: 1) Coverage - Questions must cover main arguments and counterarguments - Questions should address different perspectives presented - Questions should explore key concepts and assumptions 2) Constraints - Each question must be directly answerable from the discussion content - Questions should not require information beyond the provided thread - Questions should be neutral and not lead to specific conclusions 3) Structure - Questions should progress from fundamental understanding to deeper analysis - Include both factual and analytical questions - Avoid yes/no questions; prefer "what", "how", and "why" formats
Output format: { “pre_generated_questions": [{“id": “Q1", “question": “question_text", “type": “fundamentalanalytical", “aspects_covered": [“aspect1", “aspect2"]}], “topic_summary": “brief_topic_summary"}
Notes: - Each question should be answerable within 100 words - Questions should encourage critical thinking - Questions should be self-contained and clear without requiring additional context

E Question Generation and Feedback System

E.1 Critical Thinking Question Generation Details

E.1.1 Question Design Framework. To generate high-quality critical thinking exercises, we utilize examples from the widely recognized Watson-Glaser Critical Thinking Appraisal (WGCTA) [80]. The WGCTA's focus on assumptions, arguments, deductions, inferences, and information interpretation directly addresses our desired critical thinking skills. By employing the WGCTA's structured question format (statement-question-options-answer-explanation), we ensure a clear and consistent framework for our multiple-choice questions (MCQs). However, to foster deeper engagement and a more nuanced understanding, we supplement these MCQs with open-ended questions, such as "counter-argument," "supporting evidence," and "constructing a new argument." These encourage students to analyze arguments critically, generate diverse perspectives, and participate in active argument mapping, ultimately strengthening their critical thinking abilities beyond simple knowledge recall.

E.1.2 Prompt Design. The process of achieving desired outputs from the language model required iterative prompt engineering - a complex undertaking that demands systematic refinement. We adapted an existing MCQ generation framework [4] that demonstrated superior performance with GPT-3.5 in MCQ generation tasks, extending its application from high school-level assessments to critical thinking evaluations.

Drawing from the general guidance provided in the Watson-Glaser Critical Thinking Appraisal (WGCTA), we extracted key design principles and incorporated them into our MCQ generator prompts:

(1) Questions should focus on testing reasoning processes rather than domain knowledge, ensuring all necessary information is provided within the stem.
(2) Each item must explicitly target specific evaluation skills (inference, assumption identification, deduction, interpretation, or argument evaluation).
(3) Question design should maintain appropriate difficulty without becoming overly complex or time-consuming.
(4) Strategic incorporation of common logical fallacies (e.g., straw man arguments, correlation-causation confusion) to develop users' ability to identify and evaluate argument validity.

To address the five distinct question types within the WGCTA framework, we developed five specialized prompts, each supported by 15 illustrative examples. The final prompt structure is showed below:

> Create a critical thinking Training question based on the provided text.
> You must strictly adhere to the following format without any errors: > [Insert a self-contained question stem that: - Provides all necessary information - Tests reasoning rather than factual recall - Maintains appropriate complexity - May incorporate logical fallacies where relevant]
> a) [Option A] b) [Option B] c) [Option C] d) [Option D]
> * Correct Answer: [Insert the letter corresponding to the correct answer]
> * Explanation: [Explain the logical process for determining the correct answer and why other options are incorrect]
> Please ensure: 1. The question tests reasoning process rather than mere recall 2. The stem contains all information needed to reach the answer 3. The complexity level is appropriate for critical thinking assessment
> The text is: {text}
> Examples are: { Examples }

E.1.3 Critical Thinking Question Examples. Our prompt engineering approach builds upon and extends existing prompt frameworks, utilizing WGCTA critical thinking questions (as shown in Table ??) as few-shot examples and leveraging GPT-4's advanced capabilities to generate high-quality MCQs. The following example in table 15 illustrates the output of our optimized prompt, which was applied to comments from the ChangeMyView community discussion thread titled "CMV: Euthanasia should be considered a fundamental right, and specific conditions should govern its implementation." The following is the content of the comment (from @Finklesfudge): *Humans and governments work almost exclusively at their most basic on incentive. Any decision you basically*

Table 15. Examples of Generated Critical Thinking Questions in Five WGCTA Categories: Analyzing Arguments, Assumptions, Deductions, Inferences, and Interpreting Information

Statement	Question	Options	Ans	Explanation
Does incentivizing death at a certain age lead to ethical dilemmas and societal issues?	Argument: Yes, creating incentives for dying at a specific age, such as wanting individuals to die at 75, introduces ethical concerns and societal pressures.	a) Strong Argument b) Weak Argument	a	The argument raises ethical concerns about incentivizing death, which directly addresses a potential societal issue. It explains the negative implications and pressures this could create, making it a strong argument.
Humans and governments work almost exclusively at their most basic on incentive. Any decision you basically ever make in politics and government, the first thing you need to look at is what incentives does this create and what incentive can I foresee it creating.	Assumption: Decisions in politics and government are driven primarily by incentives.	a) Assumption Made b) Assumption Not Made	a	The statement explicitly states that all decisions in politics and government are driven by incentives. Hence, it assumes that incentives are the primary driving force behind decisions in these areas.
The comment implies that both the government and families may prefer individuals to die at age 75 due to financial and social incentives.	Conclusion: The comment promotes the idea that living beyond age 75 is more of a personal burden than a societal benefit.	a) Conclusion Follows b) Conclusion Does Not Follow	b	The comment describes how incentive mechanisms may lead governments and families to prefer death at age 75, and criticizes the harmful nature of such incentive structures, but it does not express or promote the value judgment that "living beyond 75 is more of a personal burden than a societal benefit" - thus the conclusion involves a logical leap and cannot be derived from the original text.
Comment itself.	The comment implies that societal and governmental incentives might discourage people from living past the age of 75.	a) True b) Probably True c) More Information Required d) Probably False e) False	a	The comment argues that incentives for both government and family could be aligned in such a way that they might prefer individuals do not extend their lives past 75, framing it as a systemic issue.
The writer of the comment suggests that certain incentives created by the government and family structures implicitly encourage individuals to die by the age of 75. This conclusion is drawn based on a perceived lack of support, respect, or resources for individuals beyond that age.	Conclusion: The incentives mentioned in the commentary encourage the society as a whole to value members only until the age of 75.	a) Conclusion Follows b) Conclusion Does Not Follow	a	The comment outlines a scenario where both the government and family could have incentives to prefer individuals to not live beyond 75, implying a societal value placed on individuals primarily until that age. Therefore, the conclusion follows as per the logic outlined in the comment.

ever make in politics and government, the first thing you need to look at is what incentives does this create and what incentive can I foresee it creating. So what incentive does this create? It creates an incentive that the government wants you to die at 75 years old. It's also a very easy to foresee incentive that your family may want you to die at 75. So you don't spend your money and they get more, they don't have to take care of you, they don't lose their money either. You listed out half the incentive structure already, "No job, no dignity, no respect, they are angry, kids are gone, you are a burden, you are a bother, you worry" Sheesh... with that kind of talk... what's the incentive structure you've set up here? You've given almost everyone an incentive to want you to die and if you don't? What a burden... how selfish... no respect... why don't you die with dignity? This is a terrible idea.

Table 16. Examples of WGCTA Critical Thinking Questions (Part 1): Analyzing Arguments, Assumptions, and Deductions

Statement	Question	Options	Ans	Explanation
Should companies downsize their workforces to decrease expenses and maximise profits?	Argument: Yes, downsizing will protect the company from bankruptcy in hard economic times.	a) Strong Argument b) Weak Argument	b	Accepting the argument as true, avoiding bankruptcy is an essential motive for an organisation, however, the statement does not discuss bankruptcy, rather it is discussing profits and expenses. Protection against bankruptcy is not the topic, and is straying from the point, and is, therefore a weak argument.
Monarchic nations, i.e. those with royal families, differ from republic nations in several ways. An example of this difference is that citizens of monarchic nations pay more tax than citizens of republican nations.	Assumption: The government of monarchic nations are responsible for setting tax rates on their citizens.	a) Assumption Made b) Assumption Not Made	b	Explanation: The statement does not rely on the fact (or assumption) that governments set tax rates for their citizens. The statement doesn't attempt to explain what causes the difference in tax payments, merely that there is a difference.
In an attempt to cut expenses, an organisation disbanded its IT department and outsourced its IT function to a business process outsourcing company. In doing so the company has managed to save 20% on its IT function expenditure.	Outsourcing functions to business process outsourcing companies will cut expenses.	a) Conclusion Follows b) Conclusion Does Not Follow	b	Although this company saved money on their IT function, it does not state that other companies will also save money, or that other functions if outsourced would save companies money.

Received July 2024; revised December 2024; accepted March 2025

Table 17. Examples of WGCTA Critical Thinking Questions (Part 2): Inferences and Interpreting Information

Statement	Question	Options	Ans	Explanation
Turkey is a surprising addition to the list of rapidly developing economies; with a GDP increase of 8.5% in the year 2011 alone. However, such rapid growth leaves worries regarding possible side-effects. For instance, in 2011 Turkey's rate of inflation was well above that of its peers. Secondly, there is increasing concern regarding Turkey's growing dependency on foreign capital. A large portion of the Turkish banking system is part-owned by banks within the Eurozone. As the single currency falters, such a dependency raises questions about the stability of Turkish growth.	There are concerns that Turkey's development is at risk of faltering in the years after 2011.	a) True b) Probably True c) More Information Required d) Probably False e) False	a	This inference is true. The passage states that Turkey is a 'surprising' addition to the list of countries whose economy is rapidly developing. This suggests that it is performing above expectations. The passage then goes on to note that there are worries regarding the possible side effects of such growth. This suggests that there are concerns such growth will be short lived.
The Tapoloa Club is a Hawaiian-themed night club in central London. Its most popular drink is the Volcano, which emits sparks and flames. The Tapoloa Club also offers a range of cocktails in perverse containers such as pineapples and coconuts, such as the 'coconut express' and the 'pineapple pick-up' respectively. Therefore:	Conclusion: The 'coconut express' is the second most popular drink sold by the Tapoloa Club.	a) Conclusion Follows b) Conclusion Does Not Follow	b	The statement does not state the popularity of the 'coconut express', it just mentions its name, so we cannot therefore make a conclusion of its popularity, the conclusion therefore does not follow.