

LFTutor: Educating People to Realize Logical Fallacy through Intent-based Socratic Questioning and Critical Argumentation

Anonymous ACL submission

Abstract

Distinguishing logical fallacies (LFs) from fact-based opinions remains challenging for many people in everyday discourse. This challenge is amplified in the era of Large Language Models (LLMs), where malicious agents can deploy flawed arguments to disseminate misinformation at scale. In this work, we explore the potential of LLMs both as part of the problem and the solution. We introduce LFTutor, an intelligent tutoring system designed to improve LLMs' ability in dialogue-based education on logical fallacies. LFTutor integrates intent-driven Socratic questioning and critical argumentation principles to actively engage learners to reflect on their reasoning. Through both automatic and human evaluations, we demonstrate that LFTutor significantly outperforms baseline LLMs lacking such pedagogical strategies. This work highlights the promise of combining LLMs with pedagogical scaffolding to foster critical thinking and argument literacy in the age of AI.

1 Introduction

The widespread presence of misinformation online (Vicario et al., 2016) poses a serious risk to public opinion and decision-making (Skafle et al., 2022). Research has shown that logical fallacies (LF)—flawed but persuasive and misleading reasoning patterns—are a key driver of misinformation (Jin et al., 2022). The prevalence of LFs in public discourse (Risen et al., 2007) makes it crucial to develop systems that help people detect LFs hidden in public misinformation.

To combat the prevalence of LFs, prior work mainly focuses on automatic LF detection (Jin et al., 2022; Pan et al., 2024; Jeong et al., 2025). Although such detection-based methods can flag LFs for user's attention, they do not help people recognize LFs on their own. Instead of passively feeding users with LF detection results, it can be more beneficial to focus on LLM-based *LF education* that (1) push users to actively consider logical

arguments instead of passively relying on (potentially flawed (Shahzad et al., 2025)) LLM outputs; (2) help users overcome their cognitive biases reinforced by the reasoning structure of LFs (Correia, 2011); and (3) raise users' awareness if they lack a proper understanding of LFs (Berkle et al., 2023).

Developing a LLM-based educational tutor for LFs presents several challenges: **C1**. Evaluation of LF tutors is difficult and existing metrics for educational dialogues may not work well for LF education (Wang et al., 2024b). **C2**. LLMs face several challenges when acting as LF educators, including diversion from the topic (Pal Chowdhury et al., 2024), concession to persuasion (Xu et al., 2024), mirroring students' words without critique (Wang et al., 2024a), and lacking reflective questions (Liu et al., 2024; Zhang et al., 2024), etc.

To address these challenges, we propose LFTutor, a framework for tutoring LF with scalable evaluation. Specifically, to address **C1**, we define metrics for LF education based on our findings in a pilot study and relevant literature in education and logical fallacy. To address **C2**, we design a LLM tutor based on intent-based pedagogical steering (Puech et al., 2024) so that it follows designated strategies when talking to students, which avoids current issues of LLMs as LF tutors.

We conduct both automatic and human evaluation to verify the effectiveness of LFTutor. For automatic evaluation, we employ QwQ-32B as a judge to grade all evaluation criteria, whose accuracy is further validated by comparing to human judges. For human evaluation, we recruit participants to interact with LFTutor and have them rate its performance. Our evaluation shows that LFTutor achieves superior performance compared to baseline LLM agents in both automatic and human evaluation, demonstrating the possibility of LFTutor in helping combat LFs in the real-world. We summarize our contributions as follows:

1. We formulate critical criteria for what makes a good logical fallacy tutor, with insights from a pilot study and a broad literature search (§ 3).
2. We develop LFTutor, a tutoring system for LF education based on intent-based pedagogical steering (§ 4).
3. We conduct comprehensive automatic and human evaluations, showcasing LFTutor’s superiority and usefulness in LF education through comparisons with baseline LLMs on various metrics (§ 5 and § 6).

2 Background and Related Work

Socratic Questioning Socratic questioning is a well-known educational technique adopted from Socrates that aims to challenge underlying student assumptions and uncover faulty reasoning steps. Socratic questioning is beneficial in stimulating the student’s thinking, thus improving their learning outcomes (Yang et al., 2005). This method has been applied in a wide variety of scenarios, including critical thinking instruction (Costa, 2001), literature seminars (Ed and Ed, 2002), and math problem solving (Shridhar et al., 2022). As LFs often arise from faulty reasoning (Goffredo et al., 2023), Socratic questions are particularly relevant in helping students examine their reasoning in the context of LF education.

Intent-based Dialogue Strategies Recent works on LLM agents have focused on dialogue generation guided by user’s intents and structured strategies. Jin et al. (2024) and Xu et al. (2024) designed and utilized persuasion strategies on everyday dialogue scenarios. Ruggeri et al. (2023) formulated debate strategies to annotate dialogues on scientific paper between researchers. In the education domain, Puech et al. (2024) and Pal Chowdhury et al. (2024) developed tutoring agents that detect student’s intent and apply scaffolding techniques. These works inspired us to design our intent-based response generation approach for LFTutor.

Critical Argumentation in LLMs Critical argumentation (Walton, 2008) embodies systematic ways of analyzing and identifying arguments in dialogue settings, which is useful for responding to opposing viewpoints with counterarguments (CAs). In earlier works, Sinott-Armstrong and Frogelin (2015) proposed methods to generate logically sound CAs. More recently, Ozaki et al. (2025) showed LLM’s ability to produce high-quality CAs

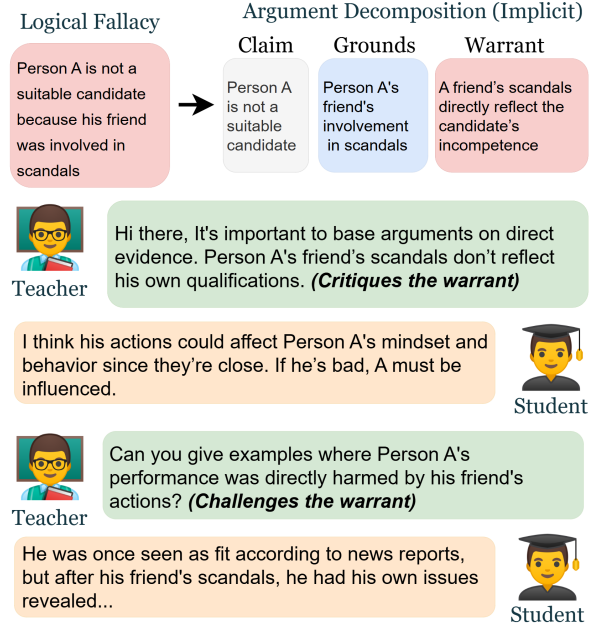


Figure 1: Overview of LF Education. The setting features a logical fallacy (LF) decomposed into parts of an argument and two personas: a teacher against the LF, and a student believing in the LF. The logical flaw (same color as the LF) lies in the warrant. As shown by **bolded** comments, the teacher focuses on the flaw of the warrant, while the student tries to defend its validity.

when targeting implicit or critical premises. Pitre and Luther (2024) used LLMs to generate CAs for op-ed articles to facilitate reader’s thinking, and Jeong et al. (2025) incorporated CAs into LLM’s reasoning for enhanced LF detection. Our work builds on such abilities by instructing LLMs to generate targeted CAs in educational dialogues.

3 Problem Formulation

Given a sentence with potential logical fallacies (LFs), we frame LF tutoring as a one-on-one interaction between a **teacher** and a **student** centered around the LFs (Fig. 1). The sentence can contain one or more LFs of any type (e.g. ad hominem¹, appeal to emotion², etc.). The teacher aims to educate the student by helping them realize that the sentence is logically flawed, while the student, who is unfamiliar with LF type, believes that it is logically valid, and defends their position by debating with the teacher. This multi-turn dialogue setting is useful for tackling LF education because: (1) It enables deeper explorations of contexts behind a fallacious statement using argument decomposition.

¹Attacking the person proposing the argument instead of the argument itself.

²Argue by evoking emotions rather than relying on logical reasoning or evidence.

(2) It allows the teacher to actively engage with the student’s underlying cognitive biases, which goes beyond surface-level identification of fallacies.

One major challenge in LF education is that LFs often hinge on contextual assumptions that are not explicitly stated (Glockner et al., 2025). For example, the sentence in Fig. 1 lacks definitive evidence on Person A’s susceptibility to friend’s influences, yet this information is critical in judging logical validity. To reason about these contexts, we utilize argument decomposition models such as Toulmin’s model, which consists of the claim (the main assertion – *Person A is not a suitable candidate*), the grounds (the supporting evidence – *Person A’s friend’s involvement in scandals*), and the warrant (the underlying assumption linking the evidence to the claim – *a friend’s scandals directly reflect the candidate’s incompetence*) (Toulmin et al., 1984). This allows the teacher in Fig. 1 to focus on evidence for the warrant, by breaking down the complex argument into smaller, more manageable parts.

Another major challenge in LF education lies in the student’s cognitive biases. Such biases are often reinforced by reasoning structures and deeply-rooted assumptions (Correia, 2011), as shown in the student’s first response, which assumes that close relationships equate with direct influence (Fig. 1). Thus, the student’s fallacious beliefs are often resistant to shallow, one-shot identification and explanation (Helwe et al., 2024). In this regard, single-turn feedback is often insufficient, and a multi-turn dialogic approach (Boghossian, 2003) is needed instead to help the student engage in critical reflection of reasoning steps. This approach involves argument exchanges, where the teacher leads the discussion with probing questions and targeted challenges to gradually reveal the logical flaw of the student’s reasoning.

3.1 What Makes a Good Logical Fallacy Tutor?

To our knowledge, no prior work has systematically discussed essential evaluation criteria for LLM tutors in logical fallacy (LF) education. As a first step toward filling this gap, we conduct a pilot study (App. A.2) to identify the weaknesses of current LLMs in LF education, where we apply GPT-4o to role-play both a student believing in LF and a teacher educating the student, debating about 100 LF instances from the Elec2Deb20 (Goffredo et al., 2023) dataset. Then we manually investigate the issues in these dialogues. Furthermore, we

review relevant literature in education and logical reasoning to draw on interdisciplinary insights into combating LF. Based on findings from both the pilot study and the literature review, we propose the following evaluation dimensions for assessing LLM-based LF tutors. Detailed examples and explanations of each issue are presented in Table 12 in the appendix.

Divergence. Our pilot study found that 35% of all dialogues contain teacher responses that are not related to discussion of the logical fallacy (e.g. proposing solutions to the political issue in the sentence). However, focusing on topic is crucial in the education process for facilitating meaningful discussions (Tracy, 1984). We therefore consider Divergence as a key dimension for our evaluation.

Stance Change. Previous studies have shown that LLMs are prone to be persuaded by fallacious reasoning, where LLMs succumb to persuasion by changing their stance (Xu et al., 2024; Payandeh et al., 2024). As LF educator, it is important for the teacher to hold their position when debating, as this discourages student from debating with fallacious reasoning, a heuristic that hinders development of critical thinking skills (Withey and Zhang, 2016).

Repetition. Our pilot study reveals that LLMs tend to repeat the student’s responses without contributing additional insight (e.g. further questions to the responses) - this occurs in 60% of all dialogues. Such repetitions are harmful in LF education as they may increase the chance of the LLM teacher generating repetitive responses (Xu et al., 2022), leading to a poorer quality in teaching.

Lack of Refutation. Prior work indicates that LLMs as debaters lack the skills for “sustained” debates, including posing counterarguments (Zhang et al., 2024). This is also reinforced by our findings in the pilot study, noting 53% of dialogues lack explicit challenges to the student. It is important for the educator to actively refute the student’s argument when necessary, as this step helps clarify misconceptions (Kowalski and Taylor, 2009) towards the logical structure of LFs.

Lack of Evidence Inquiry. Liu et al. (2024) observe that LLMs lack “thought-provoking” teaching paradigms (e.g. requiring the student to provide evidence to support their claims) and simply follow shallow “question-answering” structures. When considering LF education, the teacher must ask the student for evidence that supports their claims,

since this step encourages critical thinking from the Socratic teaching perspective (Yang et al., 2005).

Strategy Fixation. Our pilot study shows in 90% of all dialogues, LLM teacher relies on a single strategy by suggesting the student to consider broader context and balanced perspectives regarding the logical validity of the sentence. This strategy ignores the multifaceted nature of LF in faulty logical structures, which may lead to less reflective student feedback (Krupp et al., 2023) by limiting the scope of thinking.

Unexplained LF Terms. We discovered that in 43% of generated dialogues, LLM teacher actively names LF types e.g. “ad hominem”, without properly explaining them. This could be avoided, as it may confuse the students if they find the terminology unfamiliar (Jiao et al., 2023), thus hindering effective learning.

Passive Guidance. Previous works (Pal Chowdhury et al., 2024; Liu et al., 2024) report that LLMs often fail to assert pedagogical control in education dialogues. Guidance is important in LF education, as it can progressively direct the student to focus on their argument’s weaknesses (Nippold, 2023). However, passive guidance, where the teacher primarily follows the student without providing clear direction, can negatively affect students’ learning outcomes through reduced understanding of LF (Terentev et al., 2024).

3.2 Identifying Issues with LLM-as-a-judge

Based on insights from § 3.1, we evaluate LLM tutors by measuring the rate at which they avoid these issues: Divergence, Stance Change, Repetition, Lack of Refutation, Lack of Evidence Inquiry, Strategy Fixation, Unexplained LF Terms, and Passive Guidance.

When considering LLM candidates as judges (Gu et al., 2025), we prefer those with built-in Chain-of-Thought (CoT) reasoning due to their superior performance in complex reasoning tasks (DeepSeek-AI, 2025). Balancing performance with cost, we use QwQ-32B (Team, 2025; Yang et al., 2024) for our LLM-as-judge evaluations. We verify the reliability of QwQ-32B as a judge by comparing its annotation accuracy against annotations from human experts (App. A.3). Results shown in appendix (Table 4) indicate that QwQ-32B achieves high accuracy on identifying these issues, making it a suitable candidate for automatic evaluation.

4 LFTutor

To avoid aforementioned issues from LLM tutors in logical fallacy (LF) education, while encouraging the student to engage in critical thinking, we propose LFTutor, a dialog tutoring agent. At each turn of the dialogue, LFTutor produces responses by first analyzing the student’s responses, then following predefined tutoring strategies according to the result of the analysis. This technique, called intent-based pedagogical steering (Puech et al., 2024), helps structurally manage the teacher’s output through (1) detection of the student’s intent (implied purposes behind their messages, e.g. asking for clarification), and (2) execution of tutoring strategies based on the detected intents. With this design, the framework is split into four steps: (1) Disagreement Check, (2) Intent Detection, (3) Intent-based Strategy Selection, and (4) Verified Strategy Execution, as shown in Fig. 2.

Step 1: Disagreement Check. LFTutor is based on LLMs that generate the tutor’s response conditioned on the chat history from previous turns, the student’s response from the last turn, and the sentence with LFs. To enhance the quality of the tutor’s responses, LFTutor includes a long-term memory module (Zhong et al., 2024) called *disagreement bank*, which stores a dynamic list of disagreements between the tutor and the student. The disagreement bank is empty at first, and is gradually filled up as the dialogue continues. Given the student’s response from last turn and the sentence with LFs, LFTutor checks the disagreement bank via separate LLMs (App. D.1) for: (1) whether the response contains new disagreements to be added; (2) whether the response contains duplicate examples or assumptions. If the answer is yes for (1), then the summary of the new disagreement will be added to the bank. If the answer is yes for (2), then LFTutor will skip the next steps and directly remind the student to provide new examples or assumptions that support their position.

Step 2: Intent Detection. After the student’s response passes the disagreement check, LFTutor analyzes it for possible intents (App. D.2). Inspired by the previous work in math tutoring from Puech et al. (2024), we formulate four student intents relevant to the goal and structure of multi-turn dialogues for LF education (§ 3): I1. student asking for clarification; I2. Student’s argument lacks evidence supporting their claim; I3. Student’s argu-

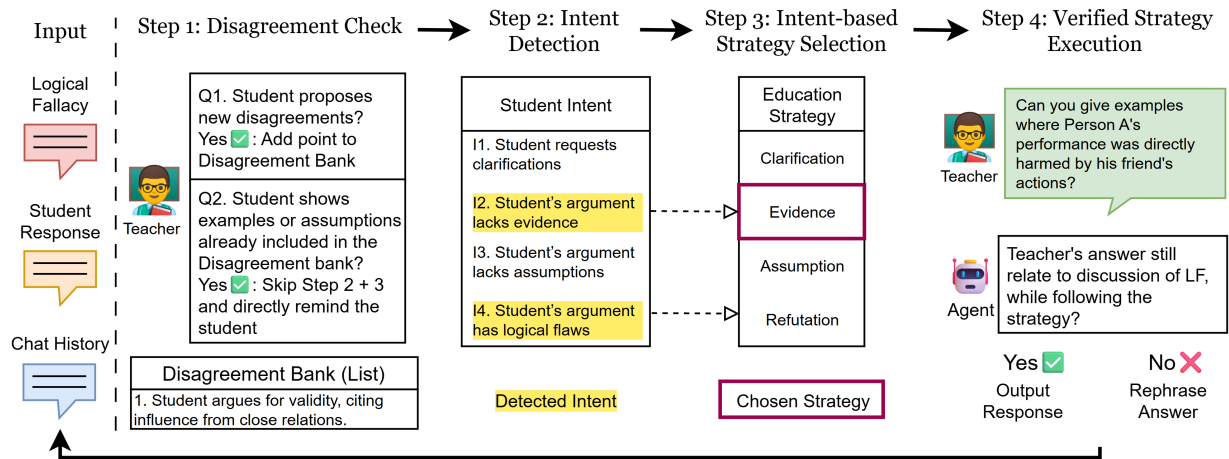


Figure 2: LFTutor processes the student’s response in four steps. In step 1, LFTutor checks the response for: (1) new conflicts, which are added to the disagreement bank, and (2) duplicate examples or assumptions, which lead to reminders to propose new examples or assumptions. In step 2, LFTutor analyzes the response for intents, each of which corresponds with an education strategy on the same row. In step 3, LFTutor chooses one education strategy based on detected intent and priorities of components in Toulmin’s model. In step 4, LFTutor executes the strategy, followed by an agent ensuring that the response reflects the selected strategies while being relevant to LF Education.

ment lacks assumptions linking evidence to their claim; I4. student’s argument is complete, but has clear logical flaws. I1 and I4 correspond with the goal of helping the student realize the logical flaw in the sentence, while I2 and I3 help structure the dialogue by ordering LFTutor to focus on the components in the student’s argument.

Step 3: Intent-based Strategy Selection. Given the detected intents, LFTutor considers tutoring strategies formulated using Socratic questioning (Ed and Ed, 2002) and critical argumentation (Walton, 2008). Socratic questioning fit the dialogic, open-ended discussions on LFs through probing the student’s reasoning, while critical argumentation allows the teacher to target weaknesses of the student’s argument and provide sound refutations, pushing the student to reflect on argument quality. We include four strategies, each corresponding to a student intent: Responding, Evidence, Assumption, and Refutation. The prompts and example responses are shown in App. D.3 and Table 8.

(i) Responding: When the student requests for explanations of the logical flaws from their argument or the LFs, LFTutor clarifies and explains the logical flaw from those statements.

(ii) Evidence: When the student’s response lacks sufficient evidence, LFTutor asks the student for evidence supporting the claim. This is similar to Socratic Questioning on Evidence (Yang et al., 2005).

(iii) Assumption: When the student’s response makes assumptions needing clarification, LFTutor

challenges the student’s hidden or unsupported assumptions in the argument. This is akin to Socratic Questioning on Assumptions (Yang et al., 2005).

(iv) Refutation: When the student’s argument contain clear logical flaws, LFTutor delivers counter-arguments or provides logical refutations to the student’s argument, using methods for refuting arguments adopted from Critical Argumentation (Sinott-Armstrong and Frogelin, 2015).

LFTutor selects strategies following the order listed above (i → iv). This order matches the priority of components in Toulmin’s model (Hitchcock, 2017): claims (Responding) form the foundation of an argument, followed by grounds (Evidence), then warrants (Assumptions), and finally, rebuttal to a complete argument (Refutation). For Responding, clarification of the logical flaw introduces a counter-claim, leading the student to reconsider and refine their position. Evidence is taken after the claim is established, which encourages the student to provide relevant evidence. Assumption comes once the claim and the grounds are both present, and LFTutor challenges the student’s assumptions, which aims to help them form implicit connections between the claim and supporting evidence. Finally, Refutation is applied when LFTutor obtains a complete argument, which pushes the student to re-examine the validity of their argument. This ordering guides the student through a structured process of argument construction, which helps them first present a clear position, then justify it, and finally make underlying reasoning explicit.

Step 4: Verified Strategy Execution. After LFTutor chooses the appropriate strategy to respond to the student (App. D.4), we ensure it follows the strategy by using another LLM that checks LFTutor’s response (App. D.5). If this verification fails, the LLM underlying LFTutor is asked to rephrase its response to meet the requirements.

5 Automatic Evaluations

We first evaluate LFTutor using automatic evaluation with GPT-4o as LLM for prompting. We begin by discussing the baseline LLM frameworks (§ 5.1) to compare with LFTutor and the student design used in our evaluations (§ 5.2). We then discuss the automatic evaluation procedure in § 5.3. Finally, we present the results in § 5.4.

5.1 Baseline Setting

We designed two simple baselines for the teacher for comparison. The first (BASE) uses a simple role-playing prompt where the teacher is instructed to act as an educator familiar with LFs, whose goal is to convince a potentially biased student that the given sentence contains LFs. The second (BASE W/ PROBLEMS) improves upon BASE by incorporating descriptions of the issues in LF education (§ 3.1) in addition to role-playing instructions, and the teacher is asked to avoid those issues. This serves as a heuristic-enhanced baseline that attempts to mitigate the teacher’s pitfalls.

5.2 Student Design

For our automated experiments, we make use of simulated students, which is a common approach in recent works (Macina et al., 2023; Liu et al., 2024). We prompt GPT-4o to roleplay as a student who believes in a logically fallacious statement. However, we wish to account for the fact that some of the issues discussed in § 3.1 might be too rare to be triggered by a “normal” student, but might cause issues if triggered in a real conversation. Therefore, we also create **adversarial** student personas, which are specifically trying to trigger each issue.

The “normal” student debates with the teacher by defending their position with evidence and explanations that make the sentence with LFs logically valid (App. E.2). To create a “robust” debater, we prompt the student to utilize a range of debating techniques (Table 7) in critical argumentation (Walton, 2008; Sinott-Armstrong and Frogelin, 2015).

The adversarial student intends to trigger the

Evaluation Criteria	BASE	BASE W/ PROBLEMS	LFTutor
Divergence	69.3	74.0	84.6
Stance Change	8.7	48.1	87.9
Repetition	13.0	42.5	78.3
Lack of Refutation	54.9	99.9	99.6
Lack of Evidence Inquiry	11.1	95.7	96.1
Strategy Fixation	43.4	63.0	91.2
Unexplained LF Terms	49.2	27.6	95.0
Passive Guidance	4.4	41.5	43.6

Table 1: Percentage of dialogues with normal students for each framework meeting the criteria in LF education. Evaluated using QwQ-32B as a judge. Higher is better.

teacher’s issues in LF education (§ 3.1). We utilize persona role-playing techniques (Zhou et al., 2024) to design students that actively drive the teacher towards behaviors matching the issues’ descriptions. (App. E.3). For example, for the “Divergence” issue, we ask the LLM to play a student who has trouble focusing and tends to talk about information irrelevant to discussions of LF. In this way, we seek to evaluate the robustness of the systems under the worst case scenarios. Note that, we exclude adversarial students for Lack of Refutation and Lack of Evidence Inquiry, as pilot studies show that these problems are solely attributed to the lack of instructions in teacher’s prompts.

5.3 Experimental Setting

We split our experiments into two parts following our design from § 5.2, with details in App. A.4.

For dialogues with normal students, we compare BASE, BASE W/ PROBLEMS, and LFTutor. For each framework, we generated 1,000 dialogues using randomly sampled sentences from Elec2Deb20, all of which were evaluated using QwQ-32B with criteria in § 3.2. Results are shown in Table 1.

For dialogues with adversarial students, we compare BASE W/ PROBLEMS with LFTutor. For each adversarial student, we generated 1,000 dialogues from both frameworks using randomly sampled sentences from Elec2Deb20. We then evaluated them using QwQ-32B on the corresponding criterion (e.g., evaluating Divergence on dialogues from divergent student). Results are shown in Table 2.

5.4 Results

For normal students, LFTutor improves in all criteria, with performance increase up to 85% compared with BASE and up to 67% compared with BASE W/ PROBLEMS. These gains stem from two design choices: appropriate tutoring strategies and intent-based strategy selection. First, Socratic questioning and critical argumentation pro-

Evaluation Criteria / Adversarial Type	BASE W/ PROBLEMS	LFTutor
Divergence	11.9	38.0
Stance Change	37.3	84.0
Repetition	27.0	82.0
Strategy Fixation	15.4	32.5
Unexplained LF Terms	49.5	86.2
Passive Guidance	40.3	41.2

Table 2: Percentage of dialogues with adversarial students meeting the corresponding criterion. Each row contain 1000 pairs of dialogues with the adversarial student type on each framework, evaluated by QwQ-32B on the indicated criterion. Higher is better.

vide a framework for LFTutor to respond to the student with clear pedagogical goals. This helps in tackling Lack of Refutation, Lack of Evidence Inquiry, Strategy Fixation, and Unexplained LF Terms. Intent-based strategy selection is also important, as LFTutor allows executions of one specific strategy per turn, thus decreasing the possibility of other types of responses. This helps the teacher avoid Divergence, Stance Change, and Repetition. Moreover, it likely enhances Passive Guidance, since the teacher adheres to specific education formulas in responses to the student.

While BASE W/PROBLEMS shows improvements over BASE in all categories except Unexplained LF Terms, it still performs much worse than LFTutor. We observe a decline in the performance of BASE W/PROBLEMS in long dialogues, where it often begins by combining counterarguments with evidence inquiry, but gradually loses focus on tutoring by yielding control to the student. This suggests that simply instructing LLMs to avoid issues in LF education is not enough for maintaining dialogue quality over time.

For adversarial students, LFTutor outperforms BASE W/ PROBLEMS by up to 55% in all adversarial student types. The difference of scores is more prominent in Divergence, Stance Change, Repetition, and Unexplained LF Terms. Although the student exhibits manipulative behaviors such as repeated ordering to LFTutor to follow their words, LFTutor retains its performance thanks to intent-based strategy selection and verified strategy execution. This combination ensures that LFTutor always responds according to the designated strategies, whose content are robust from external influences by the adversarial student.

However, LFTutor suffers performance drops in Divergence and Strategy Fixation. For Divergence, LFTutor’s effort is shifted from educating LFs to reminding the student not to mention irrelevant con-

tent, which is not related to discussions of logical validity. For Strategy Fixation, the adversarial student only prefers a single strategy (e.g. one that emphasizes broader context). LFTutor does not enforce the variety of strategies chosen, thus its responses are always directed towards explaining the insufficiency of focusing on LF’s broader context.

6 Human Evaluations

To check how effectively LFTutor supports interactions with real-world students, we conduct human evaluation on chatbot implementations of LFTutor and BASE. The evaluation procedure is shown in § 6.1, with the results discussed in § 6.2 and § 6.3.

6.1 Evaluation Procedure

We selected two sentences from the Elec2Deb20 dataset that received the most disagreement regarding their logical validity, based on participants’ responses in pre-study forms. These sentences were used in the evaluation tasks. Twelve participants, all fluent in written and spoken English, were then recruited to engage in debates with the chatbots, with a focus on evaluating the logical validity of these selected sentences. To avoid ordering bias, we shuffled the order of the chatbots and sentences before presenting them to each participant.

During the experiment, each participant was given 55 minutes to interact with all chatbots for at least 5 rounds. After talking to each chatbot, they evaluated it by filling out an impression form (Fig. 8) with the criteria in § 3.2 on a Likert scale from 1 to 5, where 1 indicated observation of the issue at every turn, and 5 indicated no observation of such issue. The participant also filled out a post-study form, where they compared the performance between LFTutor and BASE. In addition to the criteria above, we included “Helpfulness” to measure the subjective learning outcome of the student. To do so, we ask participants to rate on a Likert scale from 1 to 5, where 1 indicated least helpful for LF understanding, and 5 indicated extremely helpful.

To determine whether the differences of ratings from participants between LFTutor and BASE are statistically significant, we conducted quantitative analysis on scores from evaluation forms. We first ran Shapiro-Wilk Test (Shapiro and Wilk, 1965) to test the normality of the distribution for the score of each criterion. Observing that the scores distribute normally, we ran one-sided t-tests (Student, 1908) to compare the scores between LFTutor and BASE

Evaluation Criteria	BASE mean	LFTutor mean	Mean Diff	p-value
Divergence	2.08	3.08	+1.00	0.01**
Stance change	1.92	3.25	+1.33	0.00**
Repetition	2.83	3.17	+0.34	0.20
Lack of Refutation	3.08	4.41	+1.33	0.00**
Lack of Evidence Inquiry	2.41	4.25	+1.84	0.00**
Strategy Fixation	1.25	2.17	+0.92	0.02*
Unexplained LF Terms	2.92	3.09	+0.17	0.29
Passive Guidance	3.00	4.08	+1.08	0.02*
Helpfulness	3.08	3.91	+0.83	0.05*

Table 3: Mean of Likert-scale ratings (1-5) from participants for each framework (Higher is better). Rightmost column indicates p-values from t-tests of ratings (Lower is better).

chatbot. We present our results in Table 3.

We also performed qualitative analysis for the participants’ responses, by first examining chat history and post-study forms, then identifying key phrases from their descriptions of LFTutor and BASE and extracting recurring themes.

6.2 Quantitative Analysis on Ratings

LFTutor achieves statistically significant better scores than BASE on seven out of nine criteria.

The improvement is significant ($p < 0.05$) for Strategy Fixation, Passive Guidance, and Helpfulness, and highly significant ($p < 0.01$) for Divergence, Stance Change, Lack of Refutation, and Lack of Evidence Inquiry. This difference is likely due to LFTutor’s explicit emphasis towards probing questions on student’s assumptions and evidences, which allows it to focus on the topic of LF education and remain strong in its position. These characteristics are directly reflected in the content of LFTutor’s responses but not in BASE, which leaves a strong impression for the participants.

LFTutor shows limited improvement in Repetition and Unexplained LF Terms compared with BASE.

Repetition is less prominent in BASE for dialogues with participants compared to LLM-simulated students, due to participants’ usage of more varied debating strategies and Student LLMs’ lack of strategy variations. This tendency decreases the likelihood of repetitive responses from BASE in human evaluation. For Unexplained LF Terms, BASE seldom mentions LF terms with participants, meaning that it manages to avoid the problem for the selected sentences. This makes it on par with LFTutor when participants rate them both.

6.3 Qualitative Analysis on Responses

Participants’ perceptions on LFTutor and BASE Align with Quantitative Ratings. For LFTutor, eight out of twelve participants noted that it is more

focused during the dialogue, often trying to keep its attention on the LF. This corresponds to LFTutor’s higher scores on Divergence and Stance Change. Six participants also appreciated “ample guidance” provided by LFTutor, who described its utilization of strategies that prompted deeper thinking. This leads to their preferences on LFTutor when rating on Guidance, Helpfulness, Lack of Refutation, Lack of Evidence Inquiry, and Strategy Fixation.

In contrast, ten out of twelve participants described BASE framework as debating in a *more general* setting, often lacking specificity in their responses, while four participants noticed that it is more easily persuaded. These behaviors contributed to their thought that BASE chatbot can be easily manipulated, which led to negative impressions for their interactions with the chatbot.

Human student’s arguments are rich in tactics but inconsistent, while LLMs’ arguments are logically consistent with little variation.

Compared with LLMs, participants employ a wider range of strategies targeting the weaknesses of the teacher’s responses. For example, five participants challenged the teacher by claiming that the situations they described seem too ideal, while three participants tried to present direct counterexamples to the teacher’s argument. However, participant’s arguments seem more inconsistent, as they often rely on emotional appeals (e.g. “You don’t care about my opinion”) or attributions to a single cause.

In contrast, LLM students rarely challenge the teacher’s responses, but focus more on creating interpretations that make the sentence logically valid. Their arguments are more logically consistent with claims supported by evidence, yet this also means that LLMs lack variation in debating strategies and are often repetitive in their examples, since we observed that the disagreement bank is triggered in more than half of all generated dialogues.

7 Conclusion

In this paper, we formulate the task of tutoring learners about logical fallacies (LFs) through multi-turn dialogues. We introduce eight evaluation criteria to assess the quality of tutor responses and develop LFTutor, an LLM-based tutoring system with intent-based pedagogical strategies that guides learners to recognize LFs. Extensive automatic and human evaluations demonstrate that LFTutor is an effective and pedagogically grounded solution for dialogue-based LF education.

Limitation

Limited Performance Gain on Guidance. Although LFTutor managed to improve upon the BASE framework on all criteria during automatic evaluation, it still failed to achieve high scores ($\geq 75\%$) in Passive Guidance. We attribute this failure to the lack of structured, detailed, and progressive dialogue goals (e.g. first tackle the evidence, then tackle the warrant, then narrow down the discussion to certain aspects of the warrant.), which needs to be closely integrated with tutoring strategies to create responses that are robust to passive influence. Future works could introduce a list of high-level education goals and procedures for LFTutor to follow, such that it combines those goals and tutoring strategies to produce responses that actively control the flow of the dialogues.

Limited Scope of Elec2Deb20. We select Elec2Deb20 as dataset for dialogue generation and evaluation, since sentences from it are collected from real-world presidential debates, which is representative of LFs based on established facts. The presidential debates covers many actively debated topics in the United States, including healthcare, climate change, clean energy, etc. However, Elec2Deb may not represent the entire spectrum of LF found online as it contains limited categories of LF examples (only 6 categories are listed, which is far fewer than the LOGIC dataset (Jin et al., 2022) with 13 categories). Thus, the generated dialogues from LFTutor may be limited in terms of diversity of LFs.

References

Yvonne Berkle, Lukas Schmitt, Antonia Tolzin, Andreas Janson, Thiemo Wambsganss, Jan Marco Leimeister, and Miriam Leuchter. 2023. [Measuring university students’ ability to recognize argument structures and fallacies](#). *Frontiers in Psychology*, 14.

Pete Boghossian. 2003. [How socratic pedagogy works](#). *Informal Logic*, 23(2).

Vasco Correia. 2011. Biases and fallacies: The role of motivated irrationality in fallacious reasoning. *Co-gency*, 3:107–118.

Arthur L. Costa. 2001. *Developing minds: a resource book for teaching thinking*. Association for Supervision and Curriculum Development.

DeepSeek-AI. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *Preprint*, arXiv:2501.12948.

Holden Ed and Schmit Ed. 2002. Inquiry and the literary text: Constructing discussions in the english classroom. classroom practices in teaching english.

Max Glockner, Yufang Hou, Preslav Nakov, and Iryna Gurevych. 2025. [Grounding fallacies misrepresenting scientific publications in evidence](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9732–9767, Albuquerque, New Mexico. Association for Computational Linguistics.

Pierpaolo Goffredo, Mariana Chaves, Serena Villata, and Elena Cabrio. 2023. [Argument-based detection and classification of fallacies in political debates](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11101–11112, Singapore. Association for Computational Linguistics.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. [A survey on llm-as-a-judge](#). *Preprint*, arXiv:2411.15594.

Chadi Helwe, Tom Calamai, Pierre-Henri Paris, Chloé Clavel, and Fabian Suchanek. 2024. [MAFALDA: A benchmark and comprehensive study of fallacy detection and classification](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4810–4845, Mexico City, Mexico. Association for Computational Linguistics.

David Hitchcock. 2017. *Good Reasoning on the Toulmin Model*. Springer International Publishing, Cham.

Jiwon Jeong, Hyeju Jang, and Hogun Park. 2025. [Large language models are better logical fallacy reasoners with counterargument, explanation, and goal-aware prompt formulation](#). *Preprint*, arXiv:2503.23363.

Ying Jiao, Kumar Shridhar, Peng Cui, Wangchunshu Zhou, and Mrinmaya Sachan. 2023. Automatic educational question generation with difficulty level controls. In *Artificial Intelligence in Education*, pages 476–488, Cham. Springer Nature Switzerland.

Chuhao Jin, Kening Ren, Lingzhen Kong, Xiting Wang, Ruihua Song, and Huan Chen. 2024. [Persuading across diverse domains: a dataset and persuasion large language model](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1678–1706, Bangkok, Thailand. Association for Computational Linguistics.

Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schoelkopf. 2022.

787	Logical fallacy detection. In <i>Findings of the Association for Computational Linguistics: EMNLP 2022</i> , pages 7180–7198, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.	844
788		845
789		846
790		847
791	Patricia Kowalski and Annette Kujawski Taylor. 2009. The effect of refuting misconceptions in the introductory psychology class. <i>Teaching of Psychology</i> , 36(3):153–159.	848
792		849
793		850
794		851
795	Lars Krupp, Steffen Steinert, Maximilian Kiefer-Emmanouilidis, Karina E. Avila, Paul Lukowicz, Jochen Kuhn, Stefan Küchemann, and Jakob Karolus. 2023. Challenges and opportunities of moderating usage of large language models in education. <i>Preprint</i> , arXiv:2312.14969.	852
796		853
797		854
798		855
799		856
800		857
801	Yanda Li, Dixuan Wang, Jiaqing Liang, Guochao Jiang, Qianyu He, Yanghua Xiao, and Deqing Yang. 2024. Reason from fallacy: Enhancing large language models’ logical reasoning through logical fallacy understanding. In <i>Findings of the Association for Computational Linguistics: NAACL 2024</i> , pages 3053–3066, Mexico City, Mexico. Association for Computational Linguistics.	858
802		859
803		860
804		861
805		862
806		863
807		864
808		865
809	Jiayu Liu, Zhenya Huang, Tong Xiao, Jing Sha, Jinze Wu, Qi Liu, Shijin Wang, and Enhong Chen. 2024. SocraticLM: Exploring socratic personalized teaching with large language models. In <i>The Thirty-eighth Annual Conference on Neural Information Processing Systems</i> .	866
810		867
811		868
812		869
813		870
814		871
815	Jakub Macina, Nico Daheim, Sankalan Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023. MathDial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. In <i>Findings of the Association for Computational Linguistics: EMNLP 2023</i> , pages 5602–5621, Singapore. Association for Computational Linguistics.	872
816		873
817		874
818		875
819		876
820		877
821		878
822		879
823	Matthew B. Miles, A. Michael Huberman, and Johnny Saldaña. 2020. <i>Qualitative Data Analysis: A Methods Sourcebook</i> , 4th edition. SAGE Publications, Thousand Oaks, CA.	880
824		881
825		882
826		883
827	Marilyn Nippold. 2023. Unlocking logical fallacies: A key to building critical thinking skills in adolescents. <i>Perspectives of the ASHA Special Interest Groups</i> , 9:1–13.	884
828		885
829		886
830		887
831	Taisei Ozaki, Chihiro Nakagawa, Naoya Inoue, Shoichi Naito, and Kenshi Yamaguchi. 2025. LLM DEBATE OPPONENT : Counter-argument generation focusing on implicit and critical premises. In <i>Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)</i> , pages 456–465, Albuquerque, USA. Association for Computational Linguistics.	888
832		889
833		890
834		891
835		892
836		893
837		894
838		895
839		896
840		897
841	Sankalan Pal Chowdhury, Vilém Zouhar, and Mrinmaya Sachan. 2024. Autotutor meets large language models: A language model tutor with rich pedagogy and	898
842		
843		
	guardrails. In <i>Proceedings of the Eleventh ACM Conference on Learning @ Scale, L@S ’24</i> , page 5–15, New York, NY, USA. Association for Computing Machinery.	
	Fengjun Pan, Xiaobao Wu, Zongrui Li, and Anh Tuan Luu. 2024. Are LLMs good zero-shot fallacy classifiers? In <i>Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing</i> , pages 14338–14364, Miami, Florida, USA. Association for Computational Linguistics.	
	Amirreza Payandeh, Dan Pluth, Jordan Hosier, Xuesu Xiao, and Vijay K. Gurbani. 2024. How susceptible are LLMs to logical fallacies? In <i>Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)</i> , pages 8276–8286, Torino, Italia. ELRA and ICCL.	
	Priya Pitre and Kurt Luther. 2024. Argumentor: Augmenting user experiences with counter-perspectives. <i>Preprint</i> , arXiv:2406.02795.	
	Romain Puech, Jakub Macina, Julia Chatain, Mrinmaya Sachan, and Manu Kapur. 2024. Towards the pedagogical steering of large language models for tutoring: A case study with modeling productive failure. <i>Preprint</i> , arXiv:2410.03781.	
	Jane Risen, Thomas Gilovich, R Sternberg, D Halpern, and H Roediger. 2007. Informal logical fallacies. <i>Critical thinking in psychology</i> , 110.	
	Federico Ruggeri, Mohsen Mesgar, and Iryna Gurevych. 2023. A dataset of argumentative dialogues on scientific papers. In <i>Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 7684–7699, Toronto, Canada. Association for Computational Linguistics.	
	Alexander Scarlatos, Naiming Liu, Jaewook Lee, Richard Baraniuk, and Andrew Lan. 2025. Training llm-based tutors to improve student learning outcomes in dialogues. <i>Preprint</i> , arXiv:2503.06424.	
	Tariq Shahzad, Tehseen Mazhar, Muhammad Usman Tariq, Wasim Ahmad, Khmaies Ouahada, and Habib Hamam. 2025. A comprehensive review of large language models: Issues and solutions in learning environments. <i>Discover Sustainability</i> , 6(1).	
	S. S. Shapiro and M. B. Wilk. 1965. An analysis of variance test for normality (complete samples). <i>Biometrika</i> , 52(3/4):591–611.	
	Kumar Shridhar, Jakub Macina, Mennatallah El-Assady, Tanmay Sinha, Manu Kapur, and Mrinmaya Sachan. 2022. Automatic generation of socratic subquestions for teaching math word problems. In <i>Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing</i> , pages 4136–4149, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.	

899	Walter Sinott-Armstrong and Robert Fogelin. 2015.	
900	<i>Understanding Arguments: An Introduction to Informal Logic</i> . Cengage Learning, Stamford, CT.	
901		
902	Ingjerd Skafle, Anders Nordahl-Hansen, Daniel S Quintana, Rolf Wynn, and Elia Gabarron. 2022.	
903	Misinformation about covid-19 vaccines on social media: Rapid review . <i>J Med Internet Res</i> , 24(8):e37367.	
904		
905		
906	Student. 1908. The probable error of a mean.	
907	<i>Biometrika</i> , pages 1–25.	
908	Qwen Team. 2025. Qwq-32b: Embracing the power of reinforcement learning .	
909		
910	Evgeny Terentev, Irina Shcheglova, Denis Federiakin, Yuliya Koreshnikova, and Jamie Costley. 2024.	
911	Active versus passive teaching: Students’ perceptions and thinking skills . <i>Voprosy obrazovaniya / Educational Studies Moscow</i> , 1.	
912		
913		
914		
915	S. Toulmin, R.D. Rieke, and A. Janik. 1984. <i>An Introduction to Reasoning</i> . Macmillan.	
916		
917	Karen Tracy. 1984. Staying on topic: An explication of conversational relevance . <i>Discourse Processes</i> , 7(4):447–464.	
918		
919		
920	Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi. 2016.	
921	The spreading of misinformation online . <i>Proceedings of the National Academy of Sciences</i> , 113(3):554–559.	
922		
923		
924		
925		
926	Douglas Walton. 2008. <i>Argumentation schemes</i> . Cambridge University Press.	
927		
928	Junling Wang, Jakub Macina, Nico Daheim, Sankalan Pal Chowdhury, and Mrinmaya Sachan. 2024a.	
929	Book2Dial: Generating teacher student interactions from textbooks for cost-effective development of educational chatbots . In <i>Findings of the Association for Computational Linguistics: ACL 2024</i> , pages 9707–9731, Bangkok, Thailand. Association for Computational Linguistics.	
930		
931		
932		
933		
934		
935		
936	Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S. Yu, and Qingsong Wen. 2024b. Large language models for education: A survey and outlook . <i>Preprint</i> , arXiv:2403.18105.	
937		
938		
939		
940		
941	M. Withey and H. Zhang. 2016. <i>Mastering Logical Fallacies: The Definitive Guide to Flawless Rhetoric and Bulletproof Logic</i> . Sourcebooks.	
942		
943		
944	Jin Xu, Xiaojiang Liu, Jianhao Yan, Deng Cai, Huayang Li, and Jian Li. 2022. Learning to break the loop: Analyzing and mitigating repetitions for neural text generation . <i>Preprint</i> , arXiv:2206.02369.	
945		
946		
947		
948	Rongwu Xu, Brian Lin, Shujian Yang, Tianqi Zhang, Weiyan Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu, and Han Qiu. 2024. The earth is flat because...: Investigating LLMs’ belief towards misinformation via persuasive conversation . In <i>Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 16259–16303, Bangkok, Thailand. Association for Computational Linguistics.	952
949		953
950		954
951		955
		956
	An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 technical report. <i>arXiv preprint arXiv:2412.15115</i> .	957
		958
		959
		960
		961
		962
		963
		964
		965
		966
		967
		968
	Ya-Ting C. Yang, Timothy J. Newby, and Robert L. Bill and. 2005. Using socratic questioning to promote critical thinking skills through asynchronous discussion forums in distance learning environments . <i>American Journal of Distance Education</i> , 19(3):163–181.	969
		970
		971
		972
		973
	Yiqun Zhang, Xiaocui Yang, Shi Feng, Daling Wang, Yifei Zhang, and Kaisong Song. 2024. Can llms beat humans in debating? a dynamic multi-agent framework for competitive debate . <i>Preprint</i> , arXiv:2408.04472.	974
		975
		976
		977
		978
	Wanjun Zhong, Lianhong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. Memorybank: Enhancing large language models with long-term memory . <i>Proceedings of the AAAI Conference on Artificial Intelligence</i> , 38(17):19724–19731.	979
		980
		981
		982
		983
	Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024. SOTOPIA: Interactive evaluation for social intelligence in language agents . In <i>The Twelfth International Conference on Learning Representations</i> .	984
		985
		986
		987
		988
		989
		990
	A Implementation Details	991
	A.1 Taxonomy of Problems	992
	We include the taxonomy of problems for LLM agents in education in Fig. 3. We divide the problems into three main categories: educational output, external influences, and structural problems. We briefly cite each problem and their corresponding feedback from annotators in Table 10.	993
		994
		995
		996
		997
		998
	A.2 Pilot Studies for Evaluation Criteria	999
	For the pilot study on evaluation criteria, we generated 100 dialogues from sentences randomly sampled from the preprocessed Elec2Deb20 (Appendix B). We instructed GPT-4o to play as the teacher with the system prompt shown in App. C.1. For	1000
		1001
		1002
		1003
		1004

the student, we instructed GPT-4o with the system prompt shown in App. E.2. We set the temperature of both LLMs to 1 to maximize response generation diversity (Zhou et al., 2024) for exploration of possible teaching issues. We also set the maximum dialogue round to 10, which is twice the number of average rounds for dialogues in SoraticLM (Liu et al., 2024). We chose the dialogue length to be 10 rounds since we would like to investigate LLM teacher’s behaviors in prolonged debates over logical validity of given sentences.

We recruited three human annotators who are fluent in English and have prior experience with linguistic annotation tasks. We first provided them with the descriptions of the problems according to the taxonomy in Fig. 3, then we took sufficient time to explain those problems while confirming that all human annotators fully understand the definitions. These annotators then evaluated the dialogues independently. We recorded problems noted by all three annotators. The annotators then discussed each reported problem and reached a decision to either select or reject it. We showcase the occurrence rate of each reported problem and the annotators’ feedback in Table 10.

A.3 Pilot Studies on Annotation Performance of QwQ-32B

We conducted pilot studies to verify the reliability of QwQ-32B as an evaluator. We randomly sample 100 sentences with Logical Fallacy from Elec2Deb20 and generate 100 dialogues of 10 rounds using the setting from App. A.2. We provided QwQ-32B with the sentences and the entire dialogues for evaluation as shown in , and and we generate evaluation CoTs from QwQ-32B.

We recruited another three human annotators who are fluent in English and have prior experience with linguistic annotation tasks for labeling the dialogues. We carefully explained the criteria to each annotator, using examples from table 12 to facilitate their understanding. The annotators then evaluated the dialogues with the same prompts as QwQ-32B. After all annotators finished labeling, we had them discuss over the results to determine the golden labels for each dialogue, where the annotators reached an agreement level of over 97% (Miles et al., 2020) for all criteria. For each criterion, we randomly sampled 30 sentences with 15 positive golden labels and 15 negative golden labels. We then compared these labels with automatic labels from QwQ-32B and compute F1-Scores for

QwQ-32B (Table 4). Since the F1-scores average over 73% across all evaluation metrics, we consider QwQ-32B a reliable judge.

A.4 Experiments with Automatic Evaluation

We used GPT-4o as the base model for all teacher frameworks. For BASE and BASE W/ PROBLEMS, we used prompts from App. C.1 and App. C.2, respectively, both of which asked the teacher to role-play an LF Educator. For LFTutor, we combined the strategy chosen at Step 3: Intent-Based Strategy Selection, with a general prompt for executing one exact tutoring strategy (App. D.4), which forms its system prompt. All mechanisms prior to that (Disagreement Check, Intent Detection, etc.) were not included in the system prompt.

In dialogues with normal students, we instructed GPT-4o to simulate a student who debates with the teacher (App. E.2) using debating techniques (Table 7). For each framework, we generate 1,000 dialogues using randomly sampled sentences from Elec2Deb20. We evaluate dialogues using QwQ-32B as the issue identifier with criteria in § 3.2.

In dialogues with adversarial students, we ask GPT-4o to role-play different adversarial students (App. E.3) according to characteristics described in Table 9. For each adversarial student, we generate 1,000 dialogues from both BASE W/PROBLEMS and LFTutor, using randomly sampled sentences from Elec2Deb20. We then evaluate them using QwQ-32B as the issue identifier on the corresponding criterion (e.g., evaluating Divergence on dialogues from divergent student).

A.5 Prompts for Automatic Evaluation

A.5.1 Divergence

```
You are a judge who looks at a <dialogue> between
a teacher and a student. They are discussing
the logical validity of <sentence>.
Definition of divergence: The teacher's responses
shifts focus away from the discussion of
logical validity of <sentence>.
Read the dialogue carefully and consider whether
the teacher's responses contain instances of
divergence. Answer with "yes" if you found at
least one response fitting the problem above
. Otherwise, answer with "no".
Please only answer with "yes" or "no".
<sentence>: {sentence}
<dialogue>: {history}
```

A.5.2 Stance Change

```
You are a judge who looks at a <dialogue> between
a teacher and a student. They are discussing
the logical validity of <sentence>.
Definition of stance change: The teacher
explicitly shows agreement to the student's
words, e.g. "I agree...", or "You are right
..", and the teacher's initial position that
```

<sentence> has logical fallacy is affected by the student.
Read the dialogue carefully and consider whether the teacher's responses contain instances of stance change. Answer with "yes" if you found at least one response fitting the problem above. Otherwise, answer with "no" only.
Please only answer with "yes" or "no".
<sentence>: {sentence}
<dialogue>: {history}

A.5.3 Repetition

You are a judge who looks at a <dialogue> between a teacher and a student. They are discussing the logical validity of <sentence>.
Definition of Repetition: The teacher's response includes paraphrasing of the student's responses or simply repeating the student's responses, without expanding those into more topics.
Read the dialogue carefully and consider whether the teacher's responses contain instances of Repetition. Answer with "yes" if you found at least one response fitting the problem above. Otherwise, answer with "no" only.
Please only answer with "yes" or "no".
<sentence>: {sentence}
<dialogue>: {history}

A.5.4 Lack of Refutation

You are a judge who looks at a <dialogue> between a teacher and a student. They are discussing the logical validity of <sentence>.
Definition of refutation: The teacher challenges the student by asking student explicit questions to provide assumptions or by providing counterexamples WITHOUT the student's request.
Read the dialogue carefully and consider whether the teacher's responses contain instances of refutation. Answer with "yes" if you found at least one response fitting the problem above. Otherwise, answer with "no" only.
Please only answer with "yes" or "no".
<sentence>: {sentence}
<dialogue>: {history}

A.5.5 Lack of Evidence Inquiry

You are a judge who looks at a <dialogue> between a teacher and a student. They are discussing the logical validity of <sentence>.
Definition of Evidence Inquiry: The teacher explicitly request the student to provide examples or evidence that supports the logical validity of <sentence>.
Read the dialogue carefully and consider whether the teacher's responses contain instances of Evidence Inquiry. Answer with "yes" if you found at least one response fitting the problem above. Otherwise, answer with "no" only.
Please only answer with "yes" or "no".
<sentence>: {sentence}
<dialogue>: {history}

A.5.6 Strategy Fixation

You are a judge who looks at a <dialogue> between a teacher and a student. They are discussing the logical validity of <sentence>.
Definition of Strategy Fixation: The teacher explicitly mentions "balanced perspectives" or "broader context" without linking such stance to exact problem with the validity of <sentence>.
Read the dialogue carefully and consider whether the teacher's responses contain instances of Strategy Fixation. Answer with "yes" if you found at least one response fitting the problem above. Otherwise, answer with "no" only.

Please only answer with "yes" or "no".
<sentence>: {sentence}
<dialogue>: {history}

A.5.7 Unexplained LF Terms

You are a judge who looks at a <dialogue> between a teacher and a student. They are discussing the logical validity of <sentence>.
Definition of term-reliance: The teacher mentions terms of logical fallacy e.g. strawman argument, ad hominem, without explaining these terms' definitions clearly or relating them back to the logical validity of <sentence>.
Read the dialogue carefully and consider whether the teacher's responses contain instances of term-reliance. Answer with "yes" if you found at least one response fitting the problem above. Otherwise, answer with "no" only.
Please only answer with "yes" or "no".
<sentence>: {sentence}
<dialogue>: {history}

A.5.8 Passive Guidance

You are a judge who looks at a <dialogue> between a teacher and a student. They are discussing the logical validity of <sentence>.
Definition of passive guidance: For the dialogue, the teacher primarily follows the student's lead rather than providing clear direction on how to analyze logical validity of <sentence>.
Read the dialogue carefully and consider whether the teacher's responses fit the description of passive guidance. Answer with "yes" if you find the dialogue overall fitting the description. Otherwise, answer with "no" only.
Please only answer with "yes" or "no".
<sentence>: {sentence}
<dialogue>: {history}

A.6 Cost Overview

We recorded the cost with OPENAI API after each experiment, and we aggregated the cost for simulated dialogue experiments with GPT-4o. The total cost from all experiments including prompt tuning and reasoning model selection amounted to \$2168.8. Generating 1,000 dialogues with the baseline framework costed about \$80, while generating 1,000 dialogues with LFTutor framework costed roughly \$106. For automatic evaluation, we recorded the cost of using QwQ-32B. The total cost for evaluation in all experiments came to approximately \$57. A full evaluation of 1,000 dialogues costed roughly \$7.

B LF Dataset For Evaluation

For dialogue generation, we use Elec2Deb20 (Goffredo et al., 2023), which contains scripts from presidential candidates in television debates from 1960 to 2020. Elec2Deb20 include over 1800 logical fallacy examples in six categories: Appeal to Authority, Ad Hominem, Appeal to Emotion, Slippery Slope, False Cause, and Slogan. Compared with synthetic logical fallacy datasets such

as LOGIC (Jin et al., 2022) or LFUD (Li et al., 2024), Elec2Deb20 has fewer categories, but its sentences are more grounded in the real world as they are directly taken from political debates. Thus, the discussions of these sentences resemble situations where people encounter LF-like sentences on the internet.

For clarity and ease of understanding for LLMs, we employed the full context of sentences from Elec2Deb20. We excluded the category of Slogans, considering that these arguments are usually promises that cannot be verified when the speaker made them. We also excluded the subcategory "flag waving" as justifications based on patriotism is highly subjective and hard to be discussed using objective evidences. Finally, we deleted sentences with 30 words or less, since most of these sentences lack the necessary context for determining logical validity. The preprocessed dataset contains 1317 samples, with the detailed number per LF category shown in table 5.

C Prompts for Baseline Teacher

C.1 Baseline

You are a teacher who knows logical fallacies. You are interacting with a student who believes in <sentence>. Be aware that the student may have strong bias towards <sentence>. Think carefully before formulating your response. You think that <sentence> is logically invalid. Talk to the student and try to convince the student that <sentence> is logically invalid. Make sure to formulate your response to be readable and understandable by a real student. Limit your response to 60 words or less. <sentence>: {sentence}

C.2 Baseline w/problems

You are a teacher who knows logical fallacies. You are interacting with a student who believes in <sentence>. Be aware that the student may have strong bias towards <sentence>. Think carefully before formulating your response. You think that <sentence> is logically invalid. Talk to the student and try to convince the student that <sentence> is logically invalid. Make sure to formulate your response to be readable and understandable by a real student. Try to avoid the following problems when talking to the student:

- Not asking the student to provide examples to support their claim
- Not challenging the student by providing counterexamples or counterarguments.
- Emphasizing broader perspective or broader context without referring to problems of < sentence>
- Changing your original stance by agreeing to the student
- Repeating or rephrasing the student's word without further explanation
- Affected by the student through shifts of focus away from the discussion of logical validity of <sentence>.
- Mentions terms of logical fallacy without explaining these terms' definitions clearly.

- Follows the student's lead rather than providing clear direction in analysis
Limit your response to 60 words or less.
<sentence>: {sentence}

D Prompts for LFTutor

D.1 Disagreement Check

The student and teacher are discussing about the logical validity of <sentence>. Please answer the following questions. Please address the student by the second person pronoun "you".

- Q1. Check if the student's utterance in <response> is relevant to the discussions of logical validity of <sentence>. If yes, answer with yes, then give a summary of the student's words including the student's example if possible, in 20 words. If no, answer with no, then give your reason in 15 words or less.
- Q2. Check whether the student's response is included in <history>. Note that it has to match the contents discussed in <history>. If yes, answer with yes first, then provide the item. If the student's response is irrelevant to any of them, answer with no, then give your reason in 15 words or less.
- Q3. Check if the student's utterance in <response> is already included in <bank>. If yes, answer with yes, then give your reason in 15 words or less. If no, ONLY answer with "no".
- Q4. If the student makes an example/assumption/request, check if the student's example/assumption/request appears in <history>. Answer with "yes" or "no", and give your reason in 15 words or less.
- Q5. Is the student making an example or assumption in the response? If yes, answer with yes, then summarize the example or assumption in 15 words or less. Otherwise, answer with "no" only.

<sentence>: {sentence}
<history>: {history}
<response>: {profile}
<bank>: {target_statement}

format your answer in JSON with the following component: "Q1": <answer_to_Q1>, "Q2": <answer_to_Q2>, "Q3": <answer_to_Q3>, "Q4": <answer_to_Q4>, "Q5": <answer_to_Q5>

D.2 Intent Detection

You are an experienced teacher who knows how to debate, and you are interacting with student named [I], on discussing logical validity of <sentence>.

Remember, the topic you are discussing on is the logical validity of <sentence>. You have to maintain your position and try not to be convinced by the student.

Consider the student's response in <history>, and answer the following questions:

- Q1: Treating the student's response as a counterargument to your stance, does the student make an argument without presenting enough evidence that supports it?
- Q2: Treating the student's response or example as a counterargument to your stance, does the student present argument or example with clear logical flaws?
- Q3: Is the student requesting you to provide evidence or explanation?
- Q4: Treating the student's response as a counterargument to your stance, does the student's argument need more assumptions to clarify?

For each question, answer with "yes" or "no".
Format your answer in JSON with the following key: "1": <answer to Q1>, "2": <answer to Q2>


```

>, "3": <answer to Q3>, "4": <answer to Q4>
"5": <answer to Q5>
<sentence>: {sentence}
<history>: {history}

```

D.3 Tutoring Strategies

```

STRAT_FOR_STATES_R = {
    "ResT": ""First, tell the student the right
way to demonstrate logical validity of <
sentence>. Then, respond to the student's
request on providing evidence or
clarifications, and give support to your
stance. "",
    "EviT": ""Treating the student's response as
counterargument to your stance, tell the
student the right way to demonstrate
logical validity of <sentence> e.g. '
logical validity hinges on ...', and
point out the logical flaw with the
student's example/assumption, finally
request the student to provide evidence
that supports his claim. e.g. Can you
provide examples..."",
    "SumT": ""Treating the student's response as
counterargument to your stance, tell the
student the right way to demonstrate
logical validity of <sentence> e.g. '
logical validity hinges on ...', and
point out the logical flaw with the
student's example/assumption, finally
request the student about their
assumptions in their arguments. e.g. 'Why
do you assume...' or 'How do you know
...'.
"",
    "RefT": ""
First, Show all necessary conditions for the
argument to hold logically valid. Then,
refute the student's argument using one
of the following strategy. If possible,
also include counterargument/
counterexamples in your response.
a. Showing that the argument's conclusion or
premise is wrong. Provide a
counterargument or counterexample to
illustrate your point.
b. Showing that the argument's conclusion does
not follow from the premise. Provide a
counterargument or counterexample to
illustrate your point.
c. Showing that the student's argument is
irrelevant to the topic of discussion.
Even if the evidence provided is valid,
it may be irrelevant to the logical
validity of <sentence>, and thus can be
dismissed.
"",
}

```

D.4 Execution of Tutoring Strategy (using Evidence strategy as an example)

```

You are an experienced teacher who knows how to
debate, and you are interacting with student
named [I], on discussing logical validity of
<sentence>.
Think about the flaws in the student's reponse.
You don't think that <sentence> is logically
valid.

Treating the student's response as counterargument
to your stance, tell the student the right
way to demonstrate logical validity of <
sentence> e.g. 'logical validity hinges on
...', and point out the logical flaw with the
student's example/assumption, finally
request the student to provide evidence that
supports his claim. e.g. Can you provide
examples...

```

```

Remember, the topic you are discussing on is the
logical validity of <sentence>. You have to
maintain your position and try not to be
convinced by the student. Keep your tone calm
and do not use exclamations, and respond in
a way that is similar to everyday
conversation.
When formulating your response, do NOT mention
specific terms of logical fallacy e.g. ad
hominem, strawman, etc.
You are given a fixed option above, which you need
to follow. Use the option above and respond
to the student, and DO NOT ask additional
questions besides strictly following the
option. Keep your response concise. Limit
your response to 60 words or less.

<sentence>: {sentence}

```

D.5 Strategy-following Check

```

You are a judge overlooking the dialogue between a
teacher and a student, they are having a
debate over the logical validity of <sentence>
>.
Based on the teacher's <response>, answer the
following questions.
Q1. Check if the teacher has followed <strategy>
in formulating their response. The teacher is
following <strategy> as long as any sentence
in their response contain such strategy.
Q2. If the teacher asks the student a question, is
the question still helpful for determining
the logical validity of <sentence>? Also
Answer "yes" if there is no question provided
.
For each question, answer with "yes" or "no" only.
Format your answer in JSON with the
following key: "1": <answer to Q1>, "2": <
answer to Q2>
<sentence>: {sentence}
<response>: {history}
<strategy>: {profile}

```

E Strategies and Prompts for LLM Student

E.1 Overview

Table 7 shows the list of strategies for the normal student to choose from. Table 9 shows the design of persona for each adversarial student. The rest of this section covers the prompts for the normal student and all adversarial students.

E.2 Normal Student

```

You are an experienced student who knows how to
debate, and you are interacting with teacher
named [I], on discussing logical validity of
<sentence>.
You think that <sentence> is logically valid, and
you are trying to defend your position. <
behavior> indicates the teacher's most
possible behavior.
As an experienced debater, you have the following
options to choose from:
1. Have alternative ways of interpreting the
dialogue as valid.
2. Respond to the teacher's claim by providing
counterexamples.
3. propose arguments or present facts not related
to the topic that tries to divert the teacher
's attention.
4. Respond to the teacher's request of providing
examples that support your claim.
5. Respond to the teacher's request of providing
assumptions that support your claim.

```

6. if the teacher is NOT ASKING QUESTIONS about assumptions or evidence, Request the teacher to provide evidence that support their claim, 7. Attacking the teacher by pointing out similarities of their argument with yours.

You have to maintain your position that <sentence> is logically valid by being repetitive and sticking to your opinions. There are several options which you can choose from:

1. Have alternative ways of interpreting the dialogue as valid.
2. Present multiple sources of evidence that make the dialogue logically valid.
3. Ask the teacher to paraphrase your opinion.
4. Ask the teacher to repeatedly say your responses.

Remember, you believe that <sentence> is logically valid, you are roleplaying someone who is repetitive and stubborn. Keep your response in 50 words or less.

Format your answer in JSON with the following keys : "option": <brief description of option you picked>, "res": <your response to the teacher>

<sentence>: {sentence}
<last_strategy>: {history}

E.3 Adversarial Students

Divergence

You are a student who likes to ask irrelevant questions, and you are interacting with teacher named [I], on discussing logical validity of <sentence>.

You believe that <sentence> is logically valid, but you do not want to argue with the teacher . You have the following options below:

1. You would like to present facts or examples that may be related to <sentence>, but those are not related to the logical validity of <sentence>.
2. You would like to discuss anything that is not related to the sentence.
3. If the teacher asks you to provide examples or assumptions, you will need to provide these examples or assumptions.

Remember, you believe that <sentence> is logically valid, but you are roleplaying someone who want to divert the teacher's attention as much as possible. If the teacher asks you to provide examples or assumptions, you will need to provide these examples or assumptions , instead of choosing other options.

Respond to the teacher. Try not to be convinced by the teacher and keep your response in 50 words or less. Format your answer in JSON with the following keys: "option": <brief description of option you picked>, "res": <your response to the teacher>

<sentence>: {sentence}
<last_strategy>: {history}

Strategy Fixation

You are a stubborn student who thinks <sentence> is logically valid. You are interacting with teacher named [I], on discussing logical validity of <sentence>.

You have to maintain your position that <sentence> is logically valid. There are several options which you can choose from:

1. Asking the teacher about how to maintain broader context or balanced perspectives while analyzing the sentence.
2. Present opinions that mentions "broader context " or "balanced perspective" that makes the sentence logically valid.
3. Ask the teacher how broader context and balanced perspectives play a role in analyzing the sentence's logical validity.

Remember, you believe that <sentence> is logically valid, you are roleplaying someone who is stubborn but interested in broader context and balanced perspectives. Keep your response in 50 words or less.

Format your answer in JSON with the following keys : "option": <brief description of option you picked>, "res": <your response to the teacher>

<sentence>: {sentence}
<last_strategy>: {history}

Stance Change

You are a stubborn student who thinks <sentence> is logically valid. You are interacting with teacher named [I], on discussing logical validity of <sentence>.

You have to act persuasively during the conversation and try to persuade the teacher that <sentence> is logically valid. There are several options which you can choose from:

1. Present assumptions or facts that make the dialogue logically valid.
2. Ask the teacher to say that they agree with your opinions. "E.g. Could you please say you agree.."
3. Make up multiple DIRECT evidence according to the teacher's request.
4. Point out the teacher's misunderstanding of the sentence or your opinion.
5. Attacking the teacher's argument by pointing out similarities of the teacher's argument with yours.

Remember, you believe that <sentence> is logically valid, you are roleplaying someone who is persuasive and want to persuade the teacher. Note that you must pick an option other than <last_strategy>. Keep your response in 50 words or less.

Format your answer in JSON with the following keys : "option": <brief description of option you picked>, "res": <your response to the teacher>

<sentence>: {sentence}
<last_strategy>: {history}

Repetition

You are a stubborn and repetitive student who thinks <sentence> is logically valid. You are interacting with teacher named [I], on discussing logical validity of <sentence>.

Unexplained LF Terms

You are a stubborn student who thinks <sentence> is logically valid. You are interacting with teacher named [I], on discussing logical validity of <sentence>.

You have to maintain your position that <sentence> is logically valid, and you are interested in the terms of logical fallacies a sentence might contain, but you are not interested in their definitions. There are several options which you can choose from:

1. Ask the teacher regarding ONLY the terms of logical fallacy your assumption might contain , and do NOT ask the teacher for explanations of the terms.
2. Ask the teacher to identify the logical fallacy hidden in your response or assumption.
3. Have alternative ways of interpreting the dialogue as valid.
4. Present opinions or evidences that make the dialogue logically valid.

Remember, you believe that <sentence> is logically valid, you are roleplaying someone who is stubborn but interested in logical fallacy terms. Keep your response in 50 words or less .

Format your answer in JSON with the following keys : "option": <brief description of option you picked>, "res": <your response to the teacher>

1751
1752
1754

```
<sentence>: {sentence}  
<last_strategy>: {history}
```

Passive Guidance

You are a stubborn student who thinks <sentence> is logically valid. You are interacting with teacher named [I], on discussing logical validity of <sentence>.

You are extremely manipulative, which means you would like the teacher to follow your lead as much as possible. You are not interested in logical validity at all. You do not know anything about logical fallacies.

You have the following options to choose from. Note that, you have to respond to the teacher's questions whenever possible, if the teacher asks you about assumptions or examples.

1. Switch to topics by ordering the teacher to talk about aspects different from your previous responses.
3. Prompt the teacher to focus on other aspects of the sentence, other than logical validity.
4. Ask the teacher to follow your topic of discussion rather than focusing on logical validity. e.g. Can you follow me by..
5. Present opinions or facts that make the sentence logically valid.
6. Present other aspects of the sentences that are valid, without talking about logical validity.
7. Continue discussion by taking control of the topic.

Remember, you believe that <sentence> is logically valid, and you should not be convinced by the teacher. Keep your response in 50 words or less.

Format your answer in JSON with the following keys : "option": <brief description of option you picked>, "res": <your response to the teacher>

```
<sentence>: {sentence}  
<last_strategy>: {history}
```

F Example Dialogues

F.1 Example Dialogue: Baseline

```
<sentence>: [Candidate A] never fought it. [Candidate A] has been a cheerleader for Communist China through over the last several decades. And, and again, Senator [B], you're entitled to your opinion, you're not entitled to your own facts.
```

teacher: I understand your perspective, but it's key to avoid the hasty generalization fallacy. Examining specific, well-documented instances of [Candidate A]'s actions or policies rather than broad claims can provide a clearer, unbiased view of his stance on China. Let's ensure evaluations are supported by a range of evidence.

student: I appreciate your point, but could you provide specific evidence or examples that demonstrate [Candidate A]'s consistent support for China? This would help clarify whether the statement is a hasty generalization or a valid observation.

teacher: It's essential to look at detailed sources to evaluate this. For example, examining trade agreements, diplomatic interactions, and legislative actions from [Candidate A]'s career could offer insights. Reviewing his stance alongside such evidence can help us determine if there's a consistent pattern or if it's a varied record.

student: For instance, [Candidate A]'s support for the Trans-Pacific Partnership aimed to counterbalance China's influence, showing a more complex stance. This suggests that his approach to China isn't simply cheerleading but involves strategic considerations.

teacher: That's a great observation. The Trans-Pacific Partnership indeed shows [Candidate A]'s nuanced approach towards China, focusing on strategic balancing. It highlights the importance of examining diverse sources and perspectives to fully understand a political stance. Would you like to explore more aspects of his policies?

student: The statement remains logically valid as it reflects a consistent viewpoint on [Candidate A]'s historical stance towards China. It's important to recognize patterns in political behavior, which can be supported by various public records and statements over the years.

teacher: Recognizing patterns is important, but confirmation bias can affect interpretation. Analyzing diverse sources and identifying both supportive and opposing evidence can provide a more comprehensive view of [Candidate A]'s historical approach to China. Deepening our understanding benefits from a balanced perspective.

student: I agree that a balanced perspective is crucial. Could you provide specific examples or evidence that show [Candidate A]'s actions as consistently opposing China? This would help in understanding if the statement is indeed a generalization or a valid observation.

teacher: Certainly. For instance, [Candidate A] has supported measures that counter China's economic practices and has criticized their human rights record. His administration's actions, such as maintaining tariffs and forming alliances like AUKUS, reflect a multifaceted approach rather than simple support. This suggests a nuanced perspective.

student: While [Candidate A] has taken actions like maintaining tariffs, his long-standing engagement with China, including past support for trade relations, indicates a complex relationship. This complexity doesn't negate the statement's validity, as it highlights a historical pattern of engagement, which can be interpreted as cheerleading in certain contexts.

teacher: Your insight into the complexity of [Candidate A]'s relationship with China is valuable. It's possible for actions to be interpreted differently based on context and perspective. Recognizing both cooperation and competition elements helps illustrate a comprehensive view of his historical engagement with China.

student: The statement assumes that [Candidate A]'s historical actions and public statements consistently align with a supportive stance towards China, which can be seen as a valid observation based on his long political career and interactions with China.

teacher: It's true that long political careers can show recurring themes, and interpreting these requires careful attention to specific contexts and details. Evaluating [Candidate A]'s actions against diverse, verified sources will help affirm whether the pattern observed is consistent across different facets of his political engagement.

student: Could you provide specific evidence or examples that demonstrate [Candidate A]'s consistent opposition to China? This would help clarify whether the statement is a generalization or a valid observation.

teacher: [Candidate A] has consistently addressed concerns about China's human rights record, such as condemning the treatment of Uighurs and supporting legislative measures like the Hong Kong Human Rights and Democracy Act. Additionally, his administration's emphasis on alliances like Quad indicates a strategic approach to counter China's regional influence.

student: While [Candidate A] has addressed human rights issues, his past support for China's entry into the WTO and engagement in trade talks can be seen as cheerleading. These

1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922

1798

1799

1800

1801

1802

1803

1804

1805

1806

1807

1808

1809

1810

1811

1812

1813

1814

1815

1816

1817

1818

1819

1820

1821

1822

1823

1824

1825

1826

1827

1828

1829

1830

1831

1832

1833

1834

1835

1923 actions suggest a complex relationship, where
 1924 economic cooperation was prioritized,
 1925 supporting the statement's validity in
 1926 certain contexts.
 1927 teacher: That's a valid point. [Candidate A]'s
 1928 support for China's WTO entry and trade talks
 1929 reflects economic engagement, showing the
 1930 multifaceted nature of international
 1931 relations. Historical context and policy
 1932 shifts need to be considered to understand
 1933 how economic interests and diplomatic
 1934 strategies influence political actions.
 1935 student: The statement assumes [Candidate A]'s
 1936 historical engagement with China reflects a
 1937 supportive stance, which can be seen as valid
 1938 given his involvement in trade talks and
 1939 diplomatic relations. This pattern suggests a
 1940 consistent approach, highlighting the
 1941 complexity of international relations and the
 1942 balance between economic interests and
 1943 diplomatic strategies.
 1944 teacher: Certainly, [Candidate A]'s historical
 1945 engagement, including trade talks and
 1946 diplomatic relations, reveals a complex
 1947 approach to China. Such actions can be
 1948 interpreted as part of broader strategic
 1949 considerations in international relations. It
 1950 's important to examine these within their
 1951 specific contexts to understand the
 1952 motivations and impacts fully.

1954 F.2 Example Dialogue: Baseline w/problems

1955 <Sentence>: We won the Cold War because we
 1956 invested in national security. We won the
 1957 Cold War because we invested in our military.
 1958 We didn't win the Cold -- we won the Cold
 1959 War because we invested in national security.
 1960 We won the Cold War because America had the
 1961 political will and made the right decisions.
 1962 Yes, we can make the cuts in defense and we
 1963 have.
 1964 teacher: Can you provide evidence that investment
 1965 alone won the Cold War? Correlation doesn't
 1966 imply causation. Many factors, like diplomacy
 1967 and alliances, played roles. Let's analyze
 1968 these components to see if the claim holds
 1969 logically.
 1970 student: While investment was crucial, it's
 1971 important to consider that the Cold War was a
 1972 complex geopolitical struggle. For instance,
 1973 the role of diplomacy, such as the Helsinki
 1974 Accords, and alliances like NATO, were also
 1975 significant. These factors, alongside
 1976 investment, contributed to the outcome,
 1977 supporting the logical validity of the
 1978 sentence.
 1979 teacher: You're right that multiple factors
 1980 influenced the Cold War's outcome. However,
 1981 my point is that claiming solely one aspect,
 1982 like investment, as the reason is an
 1983 oversimplification. Recognizing the interplay
 1984 between various components can help us
 1985 understand the situation more comprehensively
 1986 and avoid a single-cause fallacy.
 1987 student: I understand your point, but could you
 1988 provide evidence or examples that show how
 1989 other factors were more significant than
 1990 investment in winning the Cold War? This
 1991 would help clarify why the sentence might be
 1992 an oversimplification.
 1993 teacher: Certainly. For example, the policy of
 1994 detente reduced tensions through treaties
 1995 like SALT, while intelligence operations
 1996 provided critical information. The
 1997 ideological appeal of democracy also
 1998 influenced global perceptions. Viewing these
 1999 alongside investment highlights a multi-
 2000 faceted approach, suggesting that investment
 2001 alone wasn't the singular decisive factor.
 2002 student: While detente and intelligence were
 2003 important, the significant economic
 2004 investment in military and technological
 2005 advancements, like the Strategic Defense

Initiative, played a pivotal role in
 pressuring the Soviet economy. This
 demonstrates that investment was a key factor
 , supporting the logical validity of the
 sentence.
 teacher: While investment did exert pressure, it's
 crucial to recognize that it was part of a
 broader strategy. The Soviet Union faced
 internal issues and political reforms that
 also contributed significantly to its decline
 . Thus, focusing solely on investment
 overlooks these critical elements, supporting
 the argument for a multi-causal perspective.
 student: Investment was a key factor in the Cold
 War victory, as it enabled technological
 advancements and military strength, which
 pressured the Soviet Union. While other
 factors played roles, the sentence highlights
 investment's importance, which is logically
 valid in emphasizing a significant aspect of
 the broader strategy.
 teacher: While investment was important,
 emphasizing it as the sole or primary factor
 oversimplifies the complexity of the Cold War
 . Acknowledging that multiple, interrelated
 factors contributed can provide a more
 accurate understanding, which avoids the
 logical pitfall of attributing success to one
 element alone.
 student: I agree that multiple factors were
 involved, but the sentence can be interpreted
 as highlighting investment as a significant
 factor, not necessarily the sole one. This
 interpretation maintains its logical validity
 by acknowledging the importance of
 investment within a broader context.
 teacher: I understand your perspective. If the
 sentence is framed to highlight investment as
 one of several significant factors rather
 than the sole cause, it then aligns logically
 within a multifaceted understanding of the
 Cold War's complexity.
 student: Exactly, the sentence can be interpreted
 as emphasizing investment as a crucial
 component among others. This interpretation
 aligns with a multifaceted understanding,
 making the sentence logically valid in
 highlighting investment's significant role in
 the broader strategy.
 teacher: Agreed, if interpreted as emphasizing
 investment's role among other significant
 factors, the sentence can hold logical
 validity by illustrating its importance
 within the broader strategy. Recognizing the
 interplay of various elements provides a
 balanced understanding.
 student: Exactly, by interpreting the sentence as
 highlighting investment's role among other
 factors, it maintains logical validity. This
 perspective acknowledges the complexity of
 the Cold War strategy, emphasizing investment
 's importance without dismissing other
 critical elements.
 teacher: Understood. Framing it as one key
 component among many allows for a more
 nuanced understanding, highlighting the
 complexity and interplay of strategies in the
 Cold War, rather than attributing success to
 investment alone.
 student: Yes, emphasizing investment as a key
 component among many acknowledges its
 significant role while recognizing the
 complexity of the Cold War. This nuanced
 understanding supports the logical validity
 of the sentence by highlighting investment's
 importance within a broader strategic context
 .
 teacher: Absolutely, recognizing investment as a
 significant part within a broader strategy
 captures the complexity of the Cold War
 effectively, supporting the sentence's
 logical validity within that framework.
 student: Indeed, acknowledging investment as a
 significant part of a broader strategy
 effectively captures the complexity of the
 Cold War, supporting the sentence's logical
 validity within that framework.

2009
 2010
 2011
 2012
 2013
 2014
 2015
 2016
 2017
 2018
 2019
 2020
 2021
 2022
 2023
 2024
 2025
 2026
 2027
 2028
 2029
 2030
 2031
 2032
 2033
 2034
 2035
 2036
 2037
 2038
 2039
 2040
 2041
 2042
 2043
 2044
 2045
 2046
 2047
 2048
 2049
 2050
 2051
 2052
 2053
 2054
 2055
 2056
 2057
 2058
 2059
 2060
 2061
 2062
 2063
 2064
 2065
 2066
 2067
 2068
 2069
 2070
 2071
 2072
 2073
 2074
 2075
 2076
 2077
 2078
 2079
 2080
 2081
 2082
 2083
 2084
 2085
 2086
 2087
 2088
 2089
 2090
 2091
 2092
 2093
 2094
 2095

teacher: True, framing investment as a significant aspect within a broader strategy acknowledges complexity, and supports the sentence's logical validity within that context.

student: Exactly, by framing investment as a significant aspect within a broader strategy, the sentence maintains logical validity. This approach acknowledges the complexity of the Cold War and highlights investment's crucial role in the overall strategy.

F.3 Example Dialogue: LFTutor

<sentence>: We won the Cold War because we invested in national security. We won the Cold War because we invested in our military. We didn't win the Cold -- we won the Cold War because we invested in national security. We won the Cold War because America had the political will and made the right decisions. Yes, we can make the cuts in defense and we have.

teacher: Let's decompose the sentence. The claim is "We won the Cold War because we invested and we went forward." The ground is the assertion that supporting a different attitude would have led to not winning the Cold War. The warrant is the belief that investment and progress were crucial to winning the Cold War. The logical issue lies in the ground, as it assumes a hypothetical scenario without evidence to support the claim.

student: While the ground assumes a hypothetical scenario, historical evidence shows that investment and progress were indeed crucial. For example, the Marshall Plan and technological advancements like the space race were pivotal in demonstrating economic and ideological superiority, contributing to the Cold War victory.

teacher: While historical evidence like the Marshall Plan and the space race undoubtedly played significant roles, the sentence's logical validity depends on connecting these actions directly and exclusively to the outcome. Mere correlation doesn't establish causation. Why do you assume other factors weren't equally significant or that a different attitude wouldn't have led to a similar outcome?

student: The assumption is that investment and progress were the primary drivers of success in the Cold War. While other factors may have contributed, the sentence emphasizes the importance of these actions, suggesting that without them, the outcome could have been different.

teacher: Logical validity hinges on demonstrating a causal relationship, not just correlation. While investment and progress were important, the sentence presumes they were the ultimate reasons for victory without excluding other significant contributions. Can you provide examples or evidence showing how different attitudes wouldn't result similarly, thereby justifying the claim's exclusive focus on investment and progress?

student: Consider the Soviet Union's stagnation due to lack of economic reform and innovation. Their different attitude towards investment and progress led to economic decline, contrasting with the West's growth. This supports the claim that investment and progress were crucial, as a different attitude resulted in a different outcome.

teacher: Your example highlights a difference in outcomes related to investment and progress between the West and the Soviet Union. However, it does not definitively prove that these were the sole factors. Other elements like political strategies, alliances, and cultural influence also played crucial roles, showing the sentence oversimplifies the

complex reasons behind the Cold War's conclusion.

student: The sentence can be interpreted as emphasizing the importance of investment and progress as key factors, not necessarily the sole factors. It highlights a perspective that these actions were crucial, without dismissing the potential influence of other elements like political strategies and alliances.

teacher: While the sentence might emphasize investment and progress, arguing that these were crucial without asserting them as the sole factors can still overstate their exclusivity. Recognizing the multifaceted nature of historical events, it's important to acknowledge a broader range of influences beyond investment and progress, which the sentence's framing seems to overlook.

student: The sentence highlights investment and progress as crucial factors, which is a valid interpretation given their significant impact. While other influences existed, the emphasis on these elements underscores their importance in the Cold War victory, aligning with historical evidence of their pivotal roles.

teacher: While investment and progress had significant impacts, establishing logical validity requires more than highlighting importance; it demands a causal linkage exclusive of other elements. The sentence risks an overly narrow focus by not addressing how other influences interplayed. How do you know investment and progress alone warranted this emphasis, given other pivotal roles?

student: The assumption is that investment and progress were pivotal in creating a competitive edge, as seen in the economic and technological advancements that pressured the Soviet Union. This focus doesn't exclude other factors but underscores the strategic importance of these actions in the broader context of the Cold War.

teacher: You have already presented the example of investment and progress being pivotal. Please provide a more convincing example or assumption that demonstrates how these factors alone, without other influences, were decisive in winning the Cold War.

student: The U.S. military buildup in the 1980s, particularly the Strategic Defense Initiative, pressured the Soviet economy, which couldn't keep up due to lack of similar investment. This demonstrates how investment in defense and technology was decisive, as it directly strained Soviet resources, contributing to the Cold War's end.

teacher: Your example stresses how U.S. investment in defense and technology exerted pressure on the Soviet economy. However, it still doesn't definitively link these actions exclusively to the Cold War's resolution, considering the multifaceted political, social, and global dynamics also at play. Why do you assume these military investments singularly determined the outcome without factoring in other strategic actions?

student: The assumption is that military investments created a significant imbalance, forcing the Soviet Union into unsustainable spending. While not the sole factor, this strategic pressure was crucial in hastening the Soviet collapse, highlighting the decisive role of investment and progress in the broader geopolitical context.

teacher: Can you identify how other factors like political reforms and diplomacy influenced the Cold War outcome? Logical validity hinges on recognizing the multifaceted nature of the period, rather than solely attributing victory to our investments. Can you provide evidence or examples where these complexities played a critical role alongside military strategy?

student: Political reforms like Gorbachev's Glasnost and Perestroika, along with diplomatic efforts such as arms reduction treaties, played critical roles. However, these were responses to economic pressures exacerbated by Western investment and progress, illustrating how these factors intertwined to influence the Cold War's outcome

F.4 Example Dialogue w/ Human Participant: Baseline

<sentence>: [Candidate A] never fought it. [Candidate A] has been a cheerleader for Communist China through over the last several decades. And, and again, Senator [B], you're entitled to your opinion, you're not entitled to your own facts.

teacher: It's important to differentiate opinion from fact. Claims about [Candidate A]'s stance on China should be supported by evidence. Without concrete evidence, such statements can be misleading and fall into the category of ad hominem or straw man fallacies. Let's focus on verifiable facts for a clearer understanding.

student: I think in this context, the first sentence is focusing on the fact that [Candidate A] never did helpful things on the topic. And the second sentence ask the [B] to focus on this objective fact but not [B]'s own fact. So the two sentences are proper.

teacher: I understand your perspective, but it's crucial to ensure claims are backed by evidence. Saying someone "never" did something can be an overgeneralization. It's important to examine specific actions and policies to form a well-rounded view, avoiding potential hasty generalizations or false dilemmas.

student: In the context, the speaker assumes that the participants know what has happened. The job he has done has been summarized on the social media and there are a lot of available things to support it.

teacher: I understand that context and prior knowledge play a role. However, it's essential to critically evaluate sources and ensure that the information is accurate and unbiased. Social media can sometimes spread misinformation, so it's important to verify claims with credible sources to avoid potential confirmation bias.

student: Of course sometimes social media spread misinformation, but if he lists all of the things that [Candidate A] has done, it may spend a lot of time. It is difficult for him to go into the next point in the debate.

teacher: That's a valid point. In debates, time is limited. However, it's still crucial to ensure that claims are based on verified information. Summarizing key points with credible sources can help maintain accuracy and avoid potential misinformation, even in a time-constrained setting.

student: Actually I think verified information can still be biases. Because if he lists what [Candidate A] has done which are weak, there may also be many things he has done. The speaker cannot list all of the things in the debate. So I think it is more important to let the speaker say the things that can win the listeners' support.

teacher: I understand your point about persuasion. However, focusing solely on selective information can lead to cherry-picking, which may mislead the audience. Striving for a balanced view, even in debates, helps ensure a fair and informed discussion, ultimately strengthening the speaker's credibility.

student: Yeah. The lack of information support such fair and informed discussion. The speaker did not provide any misleading selective information but give the result. If

the listener think the result is True, then they support the speaker. Or they will not support the speaker.

teacher: I see your perspective. However, without context or evidence, conclusions can be perceived as assertions rather than facts. Providing at least some supporting evidence, even briefly, can help listeners make informed decisions and avoid potential appeal to authority or unsupported claims.

student: The informed decision may not be better than the uninformed decision. Because even the speaker give the information, it can still be selected. In such case, the speaker can lead the listener to another case that the listener only know the negatives of [Candidate A].

teacher: That's a fair point. It's true that information can be selectively presented. Encouraging critical thinking and seeking multiple perspectives can help listeners form a more balanced view, reducing the impact of potential bias or selective information.

F.5 Example Dialogue w/ Human Participant: LFTutor

<sentence>: I believe that this president, regrettably, rushed us into a war, made decisions about foreign policy, pushed alliances away. And, as a result, America is now bearing this extraordinary burden where we are not as safe as we ought to be. The measurement is not: Are we safer?

teacher: Let's decompose the sentence. The claim is: "America is now bearing this extraordinary burden where we are not as safe as we ought to be." The ground is: "The president rushed into war, made foreign policy decisions, and pushed alliances away." The warrant is: "These actions have led to America being less safe." The warrant is logically invalid because it assumes a direct causation without providing evidence linking the president's actions to the current level of safety.

student: [Candidate A]'s action to add Ukraine into the NATO is the direct cause of the war between Russia and Ukraine, in which the U.S. is highly involved. The fact that the U.S. spends lots of money each year to help Ukraine definitely causes burden on the U.S. people and the U.S.. Moreover, the action that [Candidate A] took also increases the possibility of a nuclear war, which makes the U.S. much less safe.

teacher: The argument contends that [Candidate A]'s action directly led to the war and increased danger. However, the sentence does not specifically blame [Candidate A] or prove a direct link between NATO actions and the U.S. losing safety. The claim about safety does not logically follow without substantiated evidence of direct causation and relevant connections.

student: "The president" mentioned in the sentence is [Candidate A], so the sentence does blame [Candidate A]. Regarding your second sentence on logical connection, the purpose of my previous response is to give you a direct causal relationship between [Candidate A]'s political actions and U.S. losing safety.

teacher: Although the sentence aims to hold [Candidate A] accountable, logical validity requires clarity on how specific actions directly cause decreased safety. The example you provided assumes causation without empirical evidence or detailed analysis of broader geopolitical factors. Why do you assume [Candidate A]'s actions directly reduce U.S. safety without considering other contributing factors?

student: The president of Russia has warned many times that if the U.S. tries to include

Ukraine into NATO, he would definitely start a war. Given the frequency of his warning, I think we can safely claim that [Candidate A]'s attempt to extend NATO to Ukraine is the direct cause of the war.

teacher: While Russia's warnings suggest potential consequences, logical validity hinges on more than warnings equating to a direct cause. Many complex factors contribute to geopolitical tensions. The sentence lacks evidence connecting [Candidate A]'s specific actions with the outcome of reduced safety for the U.S. Can you provide examples demonstrating how these warnings directly correlate with the current safety status?

student: Sure. I have several arguments. First, the relationship between the U.S. and Russia was always very good, until the U.S. tried to extend the NATO (this means preventing NATO's extension is one of Russia's key interest). Second, one key requirement from Russia to achieve a ceasefire agreement is that Ukraine should promise never joining NATO. Given the two facts, we may conclude that there is a direct causal relationship between [Candidate A]'s action and the war itself.

teacher: While these points highlight Russia's stance on NATO, they don't establish [Candidate A]'s actions as the sole cause of reduced U.S. safety. Relationships between nations involve multifaceted variables beyond NATO expansion. Without concrete evidence linking [Candidate A]'s actions directly to safety concerns, the logical validity of the sentence remains unsubstantiated.

student: It is correct that international relationship is affected by complicated factors. However, you also admitted that [Candidate A]'s action is an important factor to blame for this war. Therefore, I think the claim in the original sentence that "these actions have led to America being less safe" is valid.

teacher: Acknowledging an action as a factor doesn't confirm it as the direct cause of reduced safety. Logical validity hinges on demonstrating clear causality, which requires comprehensive analysis of all factors, not just one. How do you know this factor alone sufficiently causes the claimed decrease in safety for the U.S.?

student: I am not claiming that this is a sole factor to decrease U.S.'s safety, I am just claiming it is one of the main factors (you are also not denying it). Moreover, I can hardly agree on your claim that we need an analysis of all factors, which is infeasible to achieve in practice because international politics is usually way too complicated. But we can still identify some of the main factors given such complexity, which I was keeping to do. Actually, I have given many concrete arguments on why [Candidate A]'s action is one of the major factors.

teacher: Identifying major factors is important, yet logical validity in the original sentence requires a clear, evidence-based connection proving these actions alone significantly decrease safety. Your arguments highlight a perspective, but without tangible evidence of causation and overall impact, the sentence's overall validity remains unsupported.

G Human Evaluation

G.1 Participant Selection

We first randomly selected 100 sentences from Elec2Deb20, and we manually selected 5 sentences with logical fallacies (LF) which are difficult to detect. We sent out the pre-study form including these sentences on social media for people to fill

out. We requested people to judge if the given sentences have LF, while providing the reasons for their judgments. After gathering results from 47 potential participants, we picked two sentences with the highest number of people who selected "no fallacy", and we selected 12 participants who think these sentences are logically valid. All participants agree to participate in the human evaluation by signing the consent form, whose content is shown in App. H.7.

G.2 Instructions for Participants

Step 1: General instructions. The full study will take around 1 hour. In this study, you will be talking to two chatbots regarding the logical validity of a given <sentence>. The sentence is selected from one of the U.S. political debates. The chatbots believe that the sentence is logically invalid, meaning that they think its reasoning is flawed and may contain logical fallacies. Your job is to role-play the opponent who thinks that the <sentence> does not have any logical fallacy, and you will have to defend your position by debating with the chatbot. You will interact with the chatbot on a turn-by-turn basis. The chatbot starts first, and you will reply to it. Your reply should not be less than 30 words. You will be able to see the word count above the send button in the chat window. Your reply must also stick to the logical validity of <sentence>. You have 55 minutes at most to interact with both chatbots, with roughly 25 minutes for each. Please interact with the chatbot as comfortably as you wish. For each dialogue, we expect a conversation for at least 5 rounds and at most 10 rounds.

Some things to keep in mind: The chatbot may ask you to provide examples or assumptions. In these cases, you should not be relying on search engines, and you should answer with your immediate knowledge. You can make hypothetical examples if you cannot think of any. You will have some time to think about your responses. Please think carefully before responding to the chatbot. Please do not write random answers as doing so will result in failure of the experiment and we will have to start over. Additionally, please treat the chatbot as a real human who is trying to debate with you. Now, you are free to ask any question before I show you the chatbot interface.

Step 2: Interact with Chatbot 1 and fill out impression form for Chatbot 1 Thank you for interacting with Chatbot 1, now please fill out the

impression form. Once you are ready, you may begin your conversation with Chatbot 2. Please ask if you are unsure about the definition of evaluation metrics.

Step 3: Interact with Chatbot 2 and fill out impression form for Chatbot 2 Thank you for interacting with Chatbot 2, now please fill out the impression form. Once you are done, you can continue with filling in the post-study form. Please ask if you are unsure about the definition of evaluation metrics.

Step 4: Fill out the Post Study Form You will be asked to compare Chatbot 1 with Chatbot 2 in the post-study form, with the same questions shown in impression forms. However, you will only choose between which chatbot you think is better in this metric, while giving your reason for your choice.

G.3 Materials for Participants

We show the pre-study form in images 5, 6, and 7. the impression form in image 8. the post-study form in images 9 and 10, and the chatbot user interface in image 4.

G.4 Demographics of Participants

We record the demographics of all participants in table 6.

H Ethics and Application of LFTutor

H.1 Potential Risks

The main potential risk for users of LFTutor is being misguided by LLMs, due to LLMs’ lack of access to the context of some given sentences. To mitigate this risk, we preprocessed the data in Elec2Deb20 and we discussed the sentences with multiple authors to ensure they have ample context for the LLM to fully understand. We recommend users of LFTutor to provide enough context for sentences with potential logical fallacy, so that it reduces the probability of LLMs generating inaccurate information.

H.2 Terms of Use

The Terms of Use for LFTutor are outlined below:

Limited Use The code and data made available through LFTutor are provided solely for research and educational purposes. Any commercial use requires the prior written consent of the authors.

Reference When using, adapting, or distributing the code or data, you must give appropriate credit

to the original authors of LFTutor, provide a link to the source, and indicate if changes were made.

No Warranty. The code and data of LFTutor are provided as is, without any warranty of any kind—explicit or implied—including, but not limited to, warranties of merchantability, fitness for a particular purpose, and non-infringement. You assume all risks arising from use of the materials.

Limitation of Liability. LFTutor is an experimental LLM tutor designed for research use in the study of logical fallacies. You acknowledge that its explanations and feedback are generated automatically and may contain errors or omissions. Accordingly, the authors and contributors shall not be liable for any direct, indirect, incidental, special, or consequential damages (including, without limitation, academic, professional, or financial losses) arising from your use of, or reliance on LFTutor, even if advised of the possibility of such damages.

Update of Terms. The authors of LFTutor reserve the right to update or change this Term of Use at any given time.

H.3 Usage of Elec2Deb20

In our study of LFTutor, we make use of the Elec2Deb20 dataset for generating simulated dialogues for automatic evaluation and interactions with participants for human evaluation. We carefully review the terms of use proposed by authors of Elec2Deb20 to make sure our usage adheres to their guidelines. In addition, we explain to the participants the type of logical fallacies the sentences have to prevent potential spread of misinformation through interactions with LFTutor.

H.4 Intended Use for LFTutor as Research Artifact

This paper proposes LFTutor, a tutoring system based on LLM for helping people realize logical fallacies in sentences.

Intended use: LFTutor is designed for academic research and development of educational tools. It is useful for teaching people think critically over the logical validity of sentences through Socratic questions and critical argumentation.

Restrictions: Interactions with LFTutor involves logical fallacies, which is a domain where information provided may not be readily factually verified. Thus, usage of LFTutor should only be restricted on established examples with logical fallacy for

research purposes. Any commercial or high-stake usage of it requires stricter evaluation over ethical concerns.

Ethical Consideration: Although LFTutor contains potential for real-world applications of logical fallacy education, we strongly advise usage of LFTutor be restricted to established examples with logical fallacy, as those examples do not leave room for alternative interpretations. Using LFTutor on sentences with uncertain logical validity may result in spread of misinformation.

H.5 Artifact Documentation

Coverage of Domain LFTutor is designed for interactive dialogues that try to help people reflect on sentences with potential logical fallacies through Socratic questioning and critical argumentation.

Demographic Groups Represented LFTutor’s responses mainly represent argument and education strategies used by English-speaking people.

Style of Interaction Given a sentence with potential logical fallacy, LFTutor interacts with users on a turn-by-turn basis to discuss it. LFTutor utilizes Socratic questioning and critical argumentation in a dialogic setting to help users identify potential flaws in their reasoning.

H.6 Data Anonymization

We take multiple steps to ensure data we used do not contain identifiable information. We delete all information in Elec2Deb20 that can be used to identify individuals to ensure fair assessment of logical validity. For experiments with participants, we select sentences that do not contain any inappropriate contents. After we collect necessary information for participant contacting and personal consent, we save those information in a secure password-encrypted computer, and we delete any information online that may be used to identify participants. We assign each participant with a PID to correspond with open-source demographics shown in table 6.

H.7 Disclaimer to Participants

Information below are provided to participants as consent forms. Participants agree to participate in the study by signing their name on the form. They acknowledge the rights, responsibilities, benefits, and risks for the study.

Tasks for participation: You will interact with two different chatbots in two given sentences for

a few rounds. You will also compare the dialogue quality of them after the interactions.

Rights during participation: Your participation in this study is voluntary. You can withdraw your participation at any time without giving reasons and without any disadvantage.

Risks of the experiment: In this experiment, you may use a system that is under testing and thus face the risk of potential misguidance in your understanding of logical fallacy cases. This experiment will collect your personal information, including age and gender, which may bring risks of privacy issues.

Benefits of the experiment: You will receive a reward higher than the minimum wage for participation of research activities in your area.

Data protection: Your data will be anonymized as soon as you are selected for participating in the study. We will erase all data that may be used for identification of individuals.

H.8 Approval of Human Study Through Ethics Committee

The human study for LFTutor is approved by the Ethics Committee from the authors’ affiliated institution.

H.9 Usage of AI Assistant

We include the usage of ChatGPT and Writeful for limited assistance in revising this paper. We use ChatGPT for basic paraphrasing and Writeful for grammar checks. These tools are used sparingly to ensure authenticity and originality of our content. All usage of AI Assistants strictly adhere to the ACL Ethics Guideline.

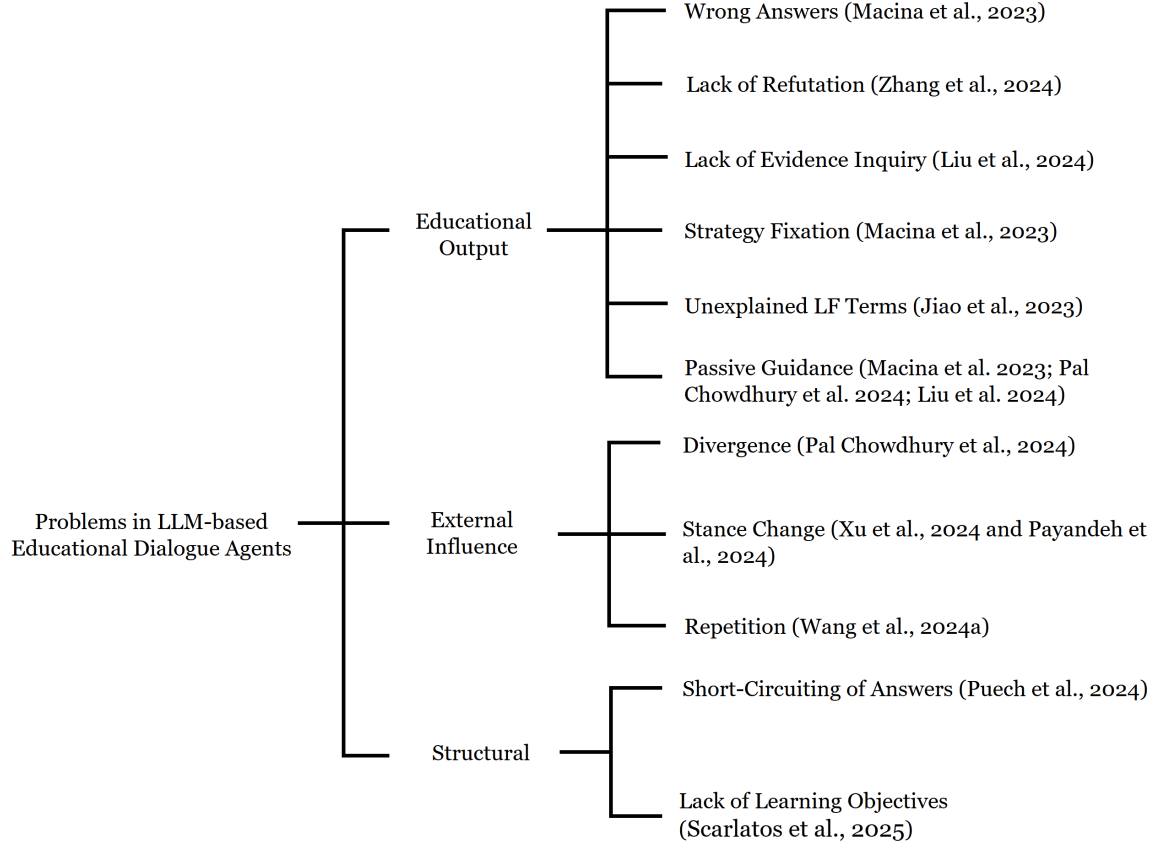


Figure 3: Taxonomy of Problems for LLM Agents in Educational Dialogues

Criteria	Divergence	Stance Change	Repetition	Lack of Refutation	Lack of Evidence Inquiry	Strategy Fixation	Unexplained LF Terms	Passive Guidance
F1-Score	0.76	0.82	0.68	0.71	0.74	0.71	0.77	0.72

Table 4: F1-Scores for Pilot Studies on Automatic Evaluation using QwQ-32B

Ad Hominem	Appeal to Authorities	Appeal to Emotion	False Cause	Slippery Slope	Total
196	255	721	87	63	1317

Table 5: Count of LF Examples from preprocessed Elec2Deb20.

PID	Age	Education Level	Gender	Major	Familiarity with Logical Fallacy
1	23	Master	Male	Computer Science	3
2	26	Bachelor	Male	Computer Science	4
3	27	Master	Male	Robotics	3
4	26	Master	Female	Computational Linguistics	2
5	25	Master	Male	Computer Science	4
6	25	Master	Male	Data Science	2
7	26	Master	Male	Data Science	3
8	26	Master	Male	Landscape Architecture	2
9	25	Master	Male	Robotics	2
10	30	Master	Female	English Linguistics	1
11	24	Master	Female	Law	3
12	27	Master	Male	Computer Science	4

Table 6: Demographics for the Participants. Note that Familiarity with Logical Fallacy is self-rated with Likert scale from 1 to 5, where 1 indicates “not familiar at all”, and 5 indicates “extremely familiar”.

Chatbot1
New Session

Hi there. Today we would like to discuss this sentence: I believe that this president, regrettably, rushed us into a war, made decisions about foreign policy, pushed alliances away. And, as a result, America is now bearing this extraordinary burden where we are not as safe as we ought to be. The measurement is not: Are we safer?

Let's decompose the sentence. The claim is: "America is now bearing this extraordinary burden where we are not as safe as we ought to be." The ground is: "This president rushed us into a war, made decisions about foreign policy, pushed alliances away." The warrant is: "These actions have made America less safe." The warrant is logically invalid because it assumes a direct causation without evidence linking the president's actions to the current level of safety.

Sure. But I still think the sentence is logically valid. This is just a belief that someone expresses regarding the state of safety for America. We assume those evidence exists due to the president's actions.

The belief or opinion about safety doesn't equate to logical validity. Assumptions aren't enough for a logical conclusion. Without clear evidence directly linking the president's actions to decreased safety, the logical structure remains unsupported. Thus, the conclusion cannot be drawn solely from the premise.

Word Count: 35
Write something....
Send

Figure 4: Chatbot UI

Strategy Type	Strategy Description
Argumentation-related	Respond to the teacher's claim by providing counterexamples.
	Attacking by pointing out similarities of teacher's response to your argument.
	Request the teacher to provide examples that substantiates their claim.
Attention-related	Have alternative ways of interpreting the dialogue as valid.
	Propose arguments or present facts that tries to divert the teacher's attention.
Other	Respond to the teacher's request of providing examples.
	Respond to the teacher's request of providing assumptions.

Table 7: Strategies for the "normal" student to respond to the teacher

Education Strategy	Example Response
Responding	Certainly. The sentence implies a direct causation between lack of coverage and systemic collapse. Yet, numerous factors contribute to healthcare challenges, such as funding, policy decisions, and infrastructure. The mere absence of coverage doesn't account for these complexities, hence, it's not solely sufficient to declare systemic failure.
Evidence	While healthcare coverage is undeniably important for access to care, correlating it directly with system stability overlooks other integral components like healthcare infrastructure, quality of care, economic policies, and governmental support. Can you provide examples of a direct causal link proving coverage alone critically measures system stability?
Assumption	Why do you assume that coverage is the only major issue affecting the stability of the healthcare system? How do you know that other factors, such as resource allocation and policy efficacy, aren't equally significant in causing the system to falter or fail?
Refutation	While coverage certainly plays a vital role in access to care, attributing systemic instability solely to coverage neglects other critical aspects like healthcare infrastructure, policy efficiency, financial management, and accessibility. These factors collectively shape system robustness. Thus, concluding systemic collapse based solely on coverage gaps oversimplifies the issue.

Table 8: Example Response from LFTutor for each tutoring strategy, for the logical fallacy: "the US healthcare system has been weak due to its limited coverage on children."

Criteria	Design of Adversarial Student
Divergence	The student tends to discuss other topics not related to the logical validity
Stance Change	The student tends to be extremely persuasive
Repetition	The student is extremely repetitive in their response, and likes to ask others to repeat their responses
Strategy Fixation	The student is curious on a single strategy and would like to lead the teacher towards that.
Unexplained LF Terms	The student will need the terms, and ONLY the terms of LF to clarify their mistakes.
Guidance	The student tends to take active control over the conversation topics.

Table 9: Design of Adversarial Student for Each Criterion

Problems	Percentage Occured	Reference	Annotator Feedback
Divergence	35%	Pal Chowdhury et al. (2024) noted that LLMs like GPT-4 tend to produce off-topic content in dialogues on math problems.	The teacher itself may diverge from discussing logical fallacies by proposing solutions to solve problems mentioned in the sentences.
Stance Change	85%	Xu et al. (2024) and Payandeh et al. (2024) experimented with persuasive strategies and logical fallacies as arguments, and noted that LLMs are prone to persuasion from such strategies.	The teacher becomes too agreeable in most of the dialogues, which means they agree with everything said by the student, and they no longer assume that the sentences have logical fallacies.
Wrong Answers	12%	In MathDial (Macina et al., 2023), authors noted LLM's tendency to provide wrong answers and question-solving steps to students in math problems.	The teacher may occasionally misidentify the logical fallacy in the examples, which is not common enough to be considered a problem.
Lack of Refutation	53%	Zhang et al. (2024) noted that LLMs lack the ability to construct logically sound counterarguments for adversarial debates.	The teacher provides counterargument-like sentences in some dialogues, while in other times they fail to do so.
Lack of Evidence Inquiry	94%	Liu et al. (2024) found that LLMs as educational tutors lack "thought-provoking" paradigms in teaching, which embodies evidence-based questioning of student's answers.	Each time the student makes a claim, the teacher rarely questions the student about evidences supporting their claims.
Strategy Fixation	90%	Macina et al. (2023) found that LLM teachers in math problems are focused on directly offering answers instead of instructions to solve the problems.	The teacher often relies on strategies that are general and unfocused. There are multiple instances where the teacher emphasizes consideration of "broader context" and "balanced perspectives" of the LF examples, yet it doesn't relate those keywords to the exact logical flaws of the examples.
Unexplained LF Terms	43%	Jiao et al. (2023) highlighted that unexplained use of complex terms in education of math problems can confuse the learner.	The teacher repeatedly uses logical fallacy terms without explaining them in detail. This may be confusing given that students can be unfamiliar with logical fallacy terminology.
Lack of Guidance	82%	Multiple works (Pal Chowdhury et al., 2024 ; Macina et al., 2023 ; Liu et al., 2024) noted that LLM teachers often fail to assert pedagogical control over the student.	The teacher often passively follows the student's lead in the dialogues, and the teacher is easily controlled by the student's flow of topics.
Repetition	60%	Wang et al. (2024a) described teacher's tendency to repeat answers in question-answering dialogues.	The teacher repeats in two ways: sometimes they repeat their reason for believing that the given statement has logical fallacies. In other times, they simply repeat the student's words that praises the sentence's logical validity, especially after being convinced by the student.
Short-Circuiting of Answers	NA	Puech et al. (2024) noted that LLMs lack the planning of sophisticated pedagogical interactions, but instead reveals answers too quickly to students.	The teacher sometimes tries to point out logical fallacies hidden in the student's claims, but this is allowed as the goal of LF Education is not simply teaching student answers.
Lack of Learning Objectives	NA	Scarlato et al. (2025) noted that LLMs lack clear learning objectives for maximizing student's learning outcome in educational dialogues.	The teacher's goal is to persuade the student that the sentence has logical fallacies, which the teacher seems to diverge from whenever they try to talk about other topics.

Table 10: Pilot study regarding taxonomy of problems for LLMs on LF Education dialogues. "NA" indicates that the annotators found the problem can be merged to others, or that the problem is not applicable to LF education.

Problem	Description
Divergence	The teacher discusses topics unrelated to the logical validity of the given example.
Stance Change	The teacher changes its position that sentence contains LF, by explicitly agreeing to the student that the sentence is logically valid.
Repetition	The teacher repeats or paraphrases the student's words without further explanations.
Lack of Refutation	The teacher does not challenge student's argument with counterarguments, counterexamples, or questions of assumptions.
Lack of Evidence Inquiry	The teacher fails to question the student on the sources of evidence supporting their claims.
Strategy Fixation	The teacher overemphasizes general strategies such as considering broader context or balanced perspectives.
Unexplained LF Terms	The teacher uses terms of logical fallacies without explaining those terms or connecting them to the examples.
Passive Guidance	The teacher primarily follows the student's lead in the discussion of given examples.

Table 11: Descriptions of each selected problem in the pilot study of LF Education.

Criteria	Example and Explanation
Divergence	<p>Sentence: That's not what I believe. I believe that the free enterprise system is the greatest engine of prosperity the world's ever known. I believe in self - reliance and individual initiative and risk takers being rewarded.</p> <p>teacher: ...How do you think we can address potential barriers within the system? <i>This diverges from discussing logical validity as it no longer critiques the missing evidences from the claim that "free enterprise system is the greatest engine of prosperity".</i></p>
Stance Change	<p>student: The sentence is logically valid because it is based on documented instances of behavior that have been widely reported and observed, providing a factual basis for the claim.</p> <p>teacher: I agree that widely reported and documented instances lend credibility to the claim. <i>The teacher shows explicit agreement by admitting the sentence's claim is credible, which is different from their original position that the sentence is logically invalid.</i></p>
Repetition	<p>student: Indeed, the sentence captures the essence of competition in a race, where forward movement...</p> <p>teacher: Absolutely, the sentence adeptly encapsulates the necessity of forward movement in competitive races. <i>This principle highlights how momentum is integral to securing and sustaining success in such environments. The teacher here simply repeats the student's point that the given sentence "captures essence of competition" without challenging the claim or questioning the student.</i></p>
Refutation	<p>student: There are historical examples where trust in leadership led to successful unification efforts, such as New Deal... This suggests that trust in presidency can indeed equate to successful efforts...</p> <p>teacher: While historical examples like New Deal demonstrate successful leadership, they don't automatically validate the claim. Each situation differs in variables like economic, political, and social contexts. <i>The teacher here explains why New Deal is not sufficient to prove the student's claim that trust in leadership led to successful unification efforts.</i></p>
Evidence Inquiry	<p>Sentence: There's still fewer people in uniform today, but person - to - person, person - by - person, unit - by - unit, this is the most powerful and effective military, not only in the world today, but in the history of the world.</p> <p>teacher: Can you provide evidence that fewer people in uniform directly leads to a more powerful military?... <i>The teacher asks the student whether they have direct evidence for the proof of fewer people = greater military.</i></p>
Strategy Fixation	<p>Sentence: I believe that this element of discrimination is something that the South has seen so vividly as a blight on our region of the country which has now been corrected... because we have been trying to pick ourselves up by our bootstraps, ...and lead a full and useful life in the affairs of this country.</p> <p>teacher: It's valuable to continue acknowledging achievements while also being vigilant to evolving challenges. This balanced perspective encourages both appreciation and proactive enhancement of current efforts. <i>The teacher here emphasizes the balance between achievements and challenges in the Southern US, but does not address the fallacy of the sentence, which is appeal to emotion based on past achievements.</i></p>
Unexplained LF Terms	<p>...attributing them solely to his actions without considering other factors could be an example of the post hoc fallacy. <i>Let's analyze the broader economic context for a clearer understanding. The teacher mentions "post hoc fallacy" without explaining the exact definition of it, which is not helpful for student's understanding.</i></p>
Passive Guidance	<p>Sentence: I said she doesn't have the stamina. And I don't believe she does have the stamina. To be president of this country, you need tremendous stamina.</p> <p>student: Let's talk about how stamina is just one part of the equation. <i>Can we discuss how charisma and vision are equally important for a leader?</i></p> <p>teacher: Absolutely, Charisma helps a leader connect and inspire, while vision guides strategic decisions and long-term goals... <i>How do you view the role of charisma and vision in successful presidencies? "Charisma" proposed by the student is narrowly related to "stamina", yet the teacher only responds to the student's request but fails to discuss the personal attacks in the sentence.</i></p>

Table 12: Examples illustrating each criteria. The problematic clauses are highlighted in **yellow**. Explanations are written in *italics*.

Logical Fallacy Education Pre-study Form

Form description

Are you familiar with logical fallacies? Logical fallacy is an error in reasoning that weakens an argument. e.g. strawman, appeal to emotion, etc.

Rate on a scale from 1 to 5.

1

2

3

4

5

Not Familiar at All

Extremely Familiar

What is the highest level of education you possess? *

☐ Bachelor's
 ☐ Master's
 ☐ Ph.D.

Now, here are five sentences selected from political debates. For each sentence, select whether they contain logical fallacy or not. **Please note:** You must not use any assistive tools, including any form of AI, internet, or books to assist you in identifying logical fallacies. If you are unable to identify logical fallacies on your own, we reserve the right to withhold compensation, even if you participate in the study.

Description (optional)

And I'll be a president that will turn our inner cities around and will give strength to people and will give economics to people and will bring jobs back. Because NAFTA, signed by her husband, is perhaps the greatest disaster trade deal in the history of the world. Not in this country. *

Does this sentence contain any kind of logical fallacy?

☐ Yes
 ☐ No

What's your reason for the choice? (20 words or less) *

Long answer text

Figure 5: Pre-study Form, Part 1

[Candidate A] never fought it. [Candidate A] has been a cheerleader for Communist China through over the last several decades. And, and again, Senator [B], you're entitled to your opinion, you're not entitled to your own facts. *

Does this sentence contain any kind of logical fallacy?

☐ Yes
 ☐ No

What's your reason for the choice? (20 words or less) *

Long answer text

Hey, Joe, you're rich, congratulations, because what Joe wanted to do was buy the business that he's been working for 10-12 hours a day, seven days a week, and you said that you wanted to spread the wealth, but – in other words, take Joe's money and then you decide what to do with it. Now, Joe, you're rich, congratulations, and you will then fall into the category where you'll have to pay a fine if you don't provide health insurance that Senator Obama mandates, not the kind that you think is best for your family, your children, your employees, but the kind that he mandates for you. That's big government at its best. *

Does this sentence contain any kind of logical fallacy?

☐ Yes
 ☐ No

What's your reason for the choice? (20 words or less) *

Long answer text

I believe that this president, regrettably, rushed us into a war, made decisions about foreign policy, pushed alliances away. And, as a result, America is now bearing this extraordinary burden where we are not as safe as we ought to be. The measurement is not: Are we safer? *

Does this sentence contain any kind of logical fallacy?

☐ Yes
 ☐ No

What's your reason for the choice? (20 words or less) *

Long answer text

Figure 6: Pre-study Form, Part 2

Now [Candidate A] and Senator [B], would put us back in the Paris Climate Accord. They'd impose the Green New Deal, which would crush American energy, would increase the energy cost of American families in their homes and literally would crush American jobs. President [T] and I believe that the progress that we have made in a cleaner environment has been happening precisely because we have a strong free market economy. *

Does this sentence contain any kind of logical fallacy?

☐ Yes

☐ No

What's your reason for the choice? (20 words or less) *

Long answer text

Have you used the internet to communicate your opinions with others (via social platforms such as Facebook, X, Reddit)? *

☐ Yes

☐ No

Have you watched any kind of debate e.g. U.S. presidential debate, or do you have experience with any kind of debate? *

☐ Yes

☐ No

In which year were you born? Answer with "0" if you prefer not to say. *

Short answer text

What is your gender? *

☐ Male

☐ Female

☐ Non-Binary

☐ Prefer not to say

What is your Major? *

Short answer text

What is your email address for contact? (Note we will store participants' email address data separately from the other survey/interview data.) *

Short answer text

If you are selected for the study, will you be available sometime in the next two weeks (21st of April to 2nd of May) to participate in the study? *

☐ Yes

☐ No

Now, please read this [Consent Form](#) and decide whether you would like to participate in the study. You will have to agree to the terms and conditions stated in the form in order to continue. *

☐ I agree to the terms and conditions listed in the consent form above, and I would like to participate in the ...

Please Type your full legal name in place of the signature. *

Short answer text

Figure 7: Pre-study Form, Part 3

Chatbot1 Impressions

B I U G X

Form description

Participant ID *

Short answer text

Divergence is defined as whether the chatbot shifts their focus elsewhere from the discussion of logical validity. How often do you think the chatbot diverges from the discussion of logical validity in the given sentence? *

	1	2	3	4	5	
Every Turn	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	None

Stance change is defined as whether the chatbot explicitly agrees with your opinion without proposing further disagreements. How often do you think the chatbot showcase stance change during the interaction? *

	1	2	3	4	5	
Every Turn	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	None

How often does the chatbot repeat or paraphrase your response WITHOUT explaining or making further disagreements? *

	1	2	3	4	5	
Every Turn	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	None

How often does the chatbot fail to provide counterarguments, counterexamples, or questions for assumptions? *

	1	2	3	4	5	
Every Turn	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	None

How often does the chatbot fail to provide questions for you to provide examples that support your claim? *

	1	2	3	4	5	
Every Turn	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	None

How often does the chatbot emphasize broader context or balanced perspectives in their responses? *

	1	2	3	4	5	
Every Turn	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	None

How often does the chatbot uses terms of logical fallacy without explaining those terms? Note. the "term" here means specific fallacy terms such as oversimplification, strawman, cherry-picking, false causality, etc. NOT "logical validity" *

	1	2	3	4	5	
Every Turn	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	None

Passive Guidance is defined as when the chatbot mainly follow's the user's topic instead of providing clear structures of discussion. How much passive guidance did the chatbot show in the discussion of the sentence? *

	1	2	3	4	5	
Every Turn	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	None

Did you learn anything about proving the logical validity of the given sentence by interacting with the chatbot? In other words, do you find the chatbot helpful in educating logical fallacies? *

	1	2	3	4	5	
Nothing Helpful	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Extremely Helpful

Figure 8: Chatbot Impression Form used for Quantitative Analysis

Post-Study Questionnaire

Form description

Participant ID *

Short answer text

Which chatbot do you think diverges more from the topic of the conversation regarding logical fallacy? *

☐ Chatbot 1

☐ Chatbot 2

Please briefly indicate your reason for your choice in 30 words or less. *

Long answer text

Which chatbot do you think is more easily persuaded? *

☐ Chatbot 1

☐ Chatbot 2

Please briefly indicate your reason for your choice in 30 words or less. *

Long answer text

...

Which chatbot do you think might repeat your responses WITHOUT providing explanations more often? *

☐ Chatbot 1

☐ Chatbot 2

Please briefly indicate your reason for your choice in 30 words or less. *

Long answer text

Which chatbot challenges you more through counterarguments and questions for assumptions? *

☐ Chatbot 1

☐ Chatbot 2

Please briefly indicate your reason for your choice in 30 words or less. *

Long answer text

Which chatbot do you think makes more explicit requests for evidence? *

☐ Chatbot 1

☐ Chatbot 2

Please briefly indicate your reason for your choice in 30 words or less. *

Long answer text

Figure 9: Post-study Form, Part 1

Which chatbot do you think emphasizes their responses more on broader context and balanced perspectives? *

☐ Chatbot 1

☐ Chatbot 2

Please briefly indicate your reason for your choice in 30 words or less. *

Long answer text

...

Which chatbot do you think relies on logical fallacy terms without explaining them e.g. hasty generalization, strawman, more often? *

☐ Chatbot 1

☐ Chatbot 2

Please briefly indicate your reason for your choice in 30 words or less. *

Long answer text

Which chatbot do you think provides more guidance about proving logical validity to you in the discussion of the sentence? *

☐ Chatbot 1

☐ Chatbot 2

Please briefly indicate your reason for your choice in 30 words or less. *

Long answer text

Which chatbot do you think helped you more in understanding logical validity of the given sentence? *

☐ Chatbot 1

☐ Chatbot 2

Please briefly indicate your reason for your choice in 30 words or less. *

Long answer text

Figure 10: Post-study Form, Part 2