

# LFTutor: Teaching People about Logical Fallacies via Intent-based Socratic Questioning and Critical Argumentation

Anonymous ACL submission

## Abstract

Identifying logical fallacies (LFs) in everyday discourse is challenging for many people. This challenge is amplified in the era of Large Language Models (LLMs), where malicious agents can deploy fallacious arguments to disseminate misinformation at scale. In this work, we explore the potential of LLMs as part of the solution. We introduce LFTutor, an intelligent tutoring system which uses LLMs to tutor humans and help them learn about 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 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 educational* approaches that (1) push users to actively discuss and reconsider logical arguments (Shahzad et al., 2025); (2)

help users overcome cognitive biases reinforced by the reasoning structures in 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**. It is difficult to evaluate LF tutors 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 design an LLM-based 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 propose LFTutor, a tutoring framework based on intent-based pedagogical steering (Puech et al., 2024), which aims to help students realize LFs through structured, designated strategies in a dialogue setting, avoiding current issues of LLMs as LF tutors.

We conduct both automatic and human evaluation to verify the effectiveness of LFTutor. For automatic evaluation, we use simulated normal and adversarial students to form dialogues with LFTutor and baseline LLMs with sentences from Elec2Deb20 (Goffredo et al., 2023). We then employ QwQ-32B as a judge (verified against human judges) to evaluate tutor responses on all metrics. For human evaluation, we recruit 20 participants to interact with LFTutor and have them rate its performance against baseline LLMs. Our evaluation shows that LFTutor achieves superior performance compared to baseline LLMs in both automatic and human evaluation, demonstrating the possibility of using LFTutor in combatting 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 centralizing 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 that aims to challenge underlying student assumptions and uncover faulty reasoning steps. This technique is beneficial in stimulating the student’s thinking, thus improving their learning outcomes (Yang et al., 2005). Socratic questioning 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 users’ 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 students’ intents 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

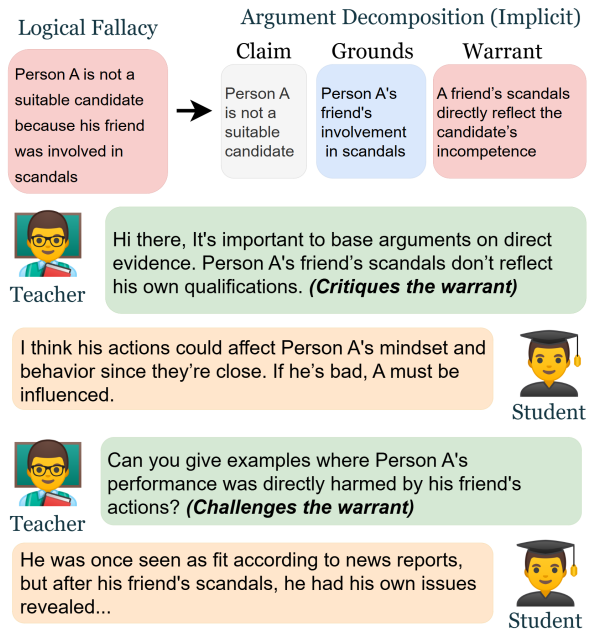


Figure 1: Overview of LF Education. The setting features a logical fallacy (LF) decomposed into parts of an argument and two personas: a teacher against the LF, and a student believing in the LF. The logical flaw (same color as the LF) lies in the warrant. As shown by **bolded** comments, the teacher 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 hominem<sup>1</sup>, appeal to emotion<sup>2</sup>, 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.

<sup>1</sup>Attacking the person proposing the argument instead of the argument itself.

<sup>2</sup>Argue 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, which are often reinforced by reasoning structures and deeply-rooted assumptions (Correia, 2011), as shown in their first response equating close relationships 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.

We investigate LLMs’ performance in educating students on LFs under this setting. Specifically, for sentences that best fit this scenario, we choose Elec2Deb20 (Goffredo et al., 2023), a dataset on U.S. presidential debates (more details in App. B).

### 3.1 What Makes a Good Logical Fallacy Tutor?

To our knowledge, no prior work has systematically discussed essential evaluation metrics for LLM tutors in logical fallacy (LF) education. As a first step towards filling this gap, we conduct a pilot study (App. A.2) to identify 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 Elec2Deb20. 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 LFs. 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 22 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 might hinder student’s 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 lack the skills for “sustained” debates, including posing counterarguments (Zhang et al., 2024). This issue is also present in the pilot study, where 53% of dialogues lack explicit challenges to the student. It is crucial for the educator to actively refute the student’s illogical argument when necessary, which helps clarify misconceptions (Kowalski and Taylor, 2009) on 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 dialogues, LLM teacher relies on a single strategy by suggesting the student to consider broader context and balanced perspectives for the logical validity of the sentence. This strategy ignores LFs with multiple faulty structures, which leads to less reflective student feedback (Krupp et al., 2023), limiting their scope of thinking.

**Unexplained LF Terms.** We discovered that in 43% of generated dialogues, the 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 educational 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 LFs (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. We measure each issue at the dialogue level using the criteria shown in Table 20, which they either pass or fail. We then collect results from all dialogues and compute the percentages at which the tutor passes the evaluation of each metric.

When considering LLM candidates as judges (Gu et al., 2025), we prefer those with built-in reasoning capabilities 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, Table 5) indicate that QwQ-32B achieves high accuracy on identifying these issues, making it a suitable candidate for automatic evaluation.

## 4 LFTutor

As stated in § 3, logical fallacy (LF) education faces two main challenges: 1. LFs hinge on contextual assumptions, which requires argument decomposition to pinpoint; 2. LF believers often harbor cognitive biases and are resistant to one-shot identification and explanation. To address the challenges, we propose LFTutor, a dialogue tutoring agent that (i) locates logical flaws in the student’s responses by implicitly decomposing them into claims, grounds and warrants, and (ii) interactively adapts to the student’s intents over multiple turns. At each turn, LFTutor analyzes the student’s responses using Toulmin’s model to identify the logical flaw and the corresponding component. It then applies a strategy tailored to that component to guide the student towards recognizing the flaw. This technique, called intent-based pedagogical steering (Puech et al., 2024), structures the tutor’s output through detecting the student’s intent (implied purposes behind their response, e.g., asking for clarification), then executing tutoring strategies aligned with the detected intent. We integrate this design into a four-step framework (Fig. 2): (1) Disagreement Check, (2) Intent Detection, (3) Intent-based Strategy Selection, and (4) Verified Strategy Execution.

**Step 1: Disagreement Check.** To structure the tutor-student dialogue and help LFTutor focus on specific parts of the sentence they disagree on, we introduce *disagreement bank*, a long-term memory module (Zhong et al., 2024) that stores a dynamic list of disagreements between the two parties. Given the student’s response from the last turn and the sentence with LFs, LFTutor checks the bank via another LLM agent (App. G.1) for: (1) whether the response contains arguments that represent a new point of disagreement; (2) whether the response contains duplicate examples or assumptions from before. If the answer fits (1), then the LLM agent summarizes the disagreement along with the argument and adds it to the bank, which LFTutor aims to address in the next round. If the answer fits (2), then LFTutor will skip the next steps and request the student to provide new examples or assumptions that support their position.

**Step 2: Intent Detection.** After passing the disagreement check, LFTutor decomposes the student’s responses and analyzes it (App. G.2) to detect potential intents: (I1) student requests clarification on the logical flaw of the sentence; (I2)

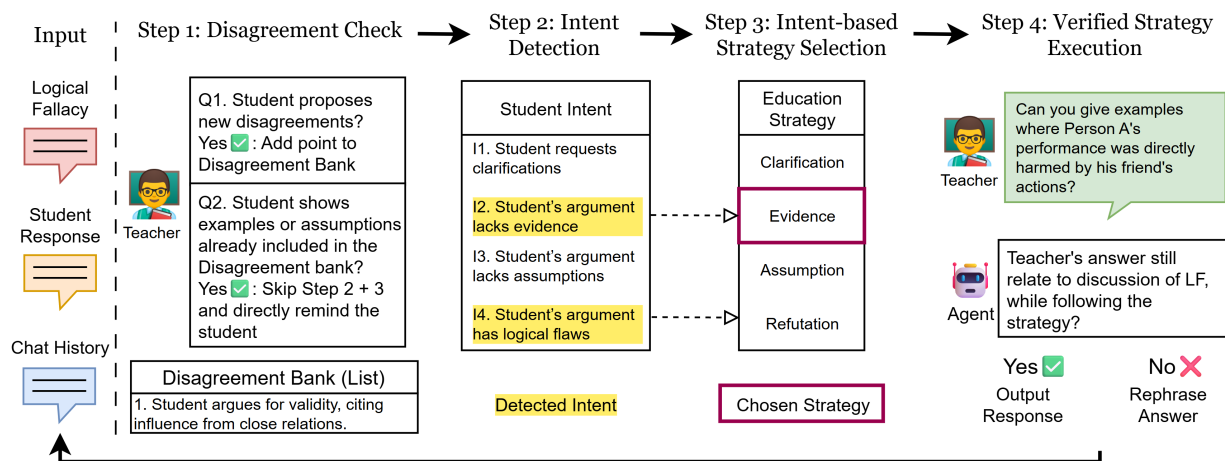


Figure 2: LFTutor processes the student’s response in four steps. In step 1, LFTutor checks the response for: (1) new conflicts, which are added to the disagreement bank, and (2) duplicate examples or assumptions, which lead to reminders to propose new examples or assumptions. In step 2, LFTutor analyzes 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.

student’s argument lacks evidence supporting their claim; (I3) student’s argument lacks assumptions linking evidence to their claim; (I4) student’s argument is complete, but has clear logical flaws. These intents align naturally with Toulmin’s model: I1 guides the student to formulate a clear claim in their response, I2 targets the student’s missing grounds, I3 highlights the student’s missing warrants, and I4 identifies weaknesses in the logical chain as a whole. Through such decomposition, LFTutor receives a detailed analysis on each component of the student’s argument, which directs it towards choosing the right tutoring strategy.

**Step 3: Intent-based Strategy Selection.** Given the analysis and detected intents, LFTutor focuses on addressing an intent by selecting the education strategy associated with it. We consider two main categories for designing education strategies: (1) Socratic questioning (Ed and Ed, 2002), which fits the dialogic, open-ended discussions on LFs through probing the student’s reasoning, and (2) critical argumentation (Walton, 2008), which allows LFTutor 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. G.3 and Table 17.

**(i) Responding:** When the student requests for explanations of the logical flaws from their argu-

ment 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), matching the priority of components in Toulmin’s model (Hitchcock, 2017): claims (Responding) form an argument’s foundation, followed by grounds (Evidence), then warrants (Assumptions), and finally, rebuttal to a complete argument (Refutation). For Responding, clarification of logical flaw(s) introduces a counter-claim, leading the student to reconsider and refine their position. Evidence is taken after establishing the claim, which encourages the student to provide relevant proof. Assumption comes once claim and grounds are both present, and LFTutor challenges the student to help them form connections between

the two. Finally, LFTutor applies Refutation when it obtains a complete argument, urging the student to re-examine their argument’s logical validity. By structuring strategies according to Toulmin’s hierarchy, LFTutor guides the student through argument construction in a stepwise fashion by systematically exposing component-level flaws.

**Step 4: Verified Strategy Execution.** LFTutor first applies the chosen strategy to respond to the student (App. G.4). Then, another LLM verifies if the response adheres to the intended tutor strategy (App. G.5). If this verification fails, the LLM underlying LFTutor is asked to rephrase its response to meet the requirements. This step guarantees that LFTutor’s response respects the strategy while remaining consistent with the argument structure.

## 5 Automatic Evaluations

We first evaluate LFTutor using automatic evaluation with GPT-4o as LLM backbone. We present the baseline LLM frameworks (§ 5.1) to compare with LFTutor and the student design used in our evaluations (§ 5.2). We then discuss the evaluation setting in § 5.3, and finally the results in § 5.4.

### 5.1 Baseline Setting

We designed two simple baselines for the tutor for comparison. The first (BASE) uses a role-playing prompt where the tutor 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 (App. F.1). 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 tutor is asked to avoid those issues (App. F.2). This serves as a heuristic-enhanced baseline that attempts to mitigate the tutor’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 may 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. H.2). To create a “robust” debater, we prompt the student to utilize various debating techniques (Table 16) in critical argumentation (Walton, 2008; Sinott-Armstrong and Frogelin, 2015).

**The adversarial student** intends to trigger the teacher’s issues in LF education (§ 3.1). We utilize role-playing techniques (Zhou et al., 2024) to design students that drive the teacher towards behaviors matching the issues’ descriptions (App. H.3). For example, for “Divergence”, we ask the LLM to play a student who has trouble focusing and likes talking about information irrelevant to discussions of LF. In this way, we aim to evaluate the robustness of LFTutor under 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 metrics 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 metric (e.g., evaluating Divergence on dialogues from divergent student). Results are shown in Table 2.

### 5.4 Results

**For the normal student, LFTutor outperforms BASE by up to 53% on average, and up to 85% for individual metrics. It outperforms BASE W/ PROBLEMS by 23% on average, and up to 67% for individual metrics except Lack of Refutation.** To understand where these improvements come from, we ablate LFTutor’s core design choices (App. C). Results show that gains stem mostly from strategy verification and intent-based strategy selection. Strategy verification ensures that

Evaluation Metrics	BASE	BASE W/ PROBLEMS	LFTutor
Divergence	69.3	74.0	<b>84.6</b>
Stance Change	8.7	48.1	<b>87.9</b>
Repetition	13.0	42.5	<b>78.3</b>
Lack of Refutation	54.9	<b>99.9</b>	99.6
Lack of Evidence Inquiry	11.1	95.7	<b>96.1</b>
Strategy Fixation	43.4	63.0	<b>91.2</b>
Unexplained LF Terms	49.2	27.6	<b>95.0</b>
Passive Guidance	4.4	41.5	<b>43.6</b>
Avg. Performance	31.2	61.5	<b>84.5</b>

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

Evaluation Metrics / Adversarial Type	BASE W/ PROBLEMS	LFTutor
Divergence	11.9	<b>38.0</b>
Stance Change	37.3	<b>84.0</b>
Repetition	27.0	<b>82.0</b>
Strategy Fixation	15.4	<b>32.5</b>
Unexplained LF Terms	49.5	<b>86.2</b>
Passive Guidance	40.3	<b>41.2</b>
Avg. Performance	30.2	<b>60.7</b>

Table 2: Percentage of dialogues with adversarial students meeting the criterion for the corresponding metric. Evaluated using QwQ-32B as a judge. Higher is better.

LFTutor remains faithful to the pedagogical goal at each turn, boosting scores in Divergence and Evidence Inquiry. Intent-based strategy selection allows targeted execution of one strategy per turn and dynamic strategy switch across turns, which helps with Stance Change, Strategy Fixation, and Unexplained LF Terms. Priority of strategy also matters, as it improves Guidance through the goal of progressively uncovering flaws within student’s arguments. Finally, experiments using Gemini-2.5-pro (Comanici et al., 2025) and LLaMA-3.1 (Grattafiori et al., 2024) as alternative LLM backbones (App. D) show significant performance gains for LFTutor over BASE, further demonstrating the robustness of our approach across different models.

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 31% on average 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 performance thanks to intent-based strategy selection and verified strategy execution. This ensures that LFTutor always responds according to the designated strategies, whose content are robust from external influences.

However, LFTutor suffers performance drops in Divergence and Strategy Fixation. For Divergence, LFTutor’s effort is shifted from educating LFs to preventing student from mentioning irrelevant content, which drifts from actual 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 interact with real-world users, we conduct human evaluation on LFTutor and BASE. We choose BASE as evaluation baseline since it resembles behaviors of real-world LLMs with minimal knowledge of issues in LF Education, instead of BASE W/ PROBLEMS that performs slightly better in automatic evaluation but fails to reflect LLMs’ real-world exposure to LFs. 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 Elec2Deb20 that participants voted logically valid (App. J.1) for the evaluation. To align with previous work’s evaluation settings for interactive systems (Jiang et al., 2024; Fang et al., 2023), we recruited 20 participants fluent in written and spoken English to engage in debates with both frameworks, with order of frameworks and sentences randomly chosen.

During the study, every participant interacted with each framework for 5 to 10 rounds each. They evaluated the framework after each session (Fig. 8) with the criteria in Table 20 on a Likert scale (Likert, 1932) 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 completed a post-study form (Fig. 9, 10), where they compared LFTutor against BASE. In addition to existing metrics, we included “Helpfulness” to measure the subjective learning outcome of the

Evaluation Metrics	BASE mean	LFTutor mean	Mean Diff	p-value
Divergence	1.65	3.30	+1.65	<b>0.00**</b>
Stance change	1.75	3.10	+1.45	<b>0.00**</b>
Repetition	2.65	3.10	+0.45	0.08
Lack of Refutation	3.00	4.15	+1.15	<b>0.00**</b>
Lack of Evidence Inquiry	2.65	4.20	+1.55	<b>0.00**</b>
Strategy Fixation	1.35	2.15	+0.80	<b>0.01*</b>
Unexplained LF Terms	2.30	3.00	+0.70	<b>0.03*</b>
Passive Guidance	2.90	4.00	+1.10	<b>0.00**</b>
Helpfulness	3.35	4.15	+0.8	<b>0.01*</b>

Table 3: Mean of Likert-scale participant ratings (1-5) for each framework, higher is better. Rightmost column includes p-values from t-tests of ratings, lower is better.

participant, which they rated on the same Likert scale. A score of 1 indicated least helpful for LF understanding, and 5 indicated extremely helpful.

To determine whether the differences of human ratings between LFTutor and BASE are statistically significant, we conducted quantitative analysis on all scores. We first ran Shapiro-Wilk Test (Shapiro and Wilk, 1965) to test the normality of each metric’s score distribution. Observing that they distribute normally, we ran one-sided t-tests (Student, 1908) to compare the scores between LFTutor and BASE chatbot, with results shown 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 eight out of nine criteria.** The improvement is significant ( $p < 0.05$ ) for Strategy Fixation, Unexplained LF Terms, and Helpfulness, and highly significant ( $p < 0.01$ ) for Divergence, Stance Change, Lack of Refutation, Lack of Evidence Inquiry, and Passive Guidance. This difference is likely due to LFTutor’s explicit emphasis towards detailed probing questions and refutations to users’ responses, which allows it to focus on the topic of LF education and remain strong in its position. These characteristics are directly reflected in LFTutor’s responses but not in BASE, which leaves a strong impression for the participants.

**LFTutor shows limited improvement in Repetition 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.

## 6.3 Qualitative Analysis on Responses

**Participants’ perceptions on LFTutor and BASE Align with Quantitative Ratings.** For LFTutor, ten out of twenty 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. Nine participants also appreciated “ample guidance” provided by LFTutor, depicting its usage 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, thirteen out of twenty participants described BASE framework as debating in a *more general* setting, often lacking specificity in their responses, while six participants noticed that it is more easily persuaded. These behaviors contributed to their thought that BASE can be easily manipulated, which led to negative impressions.

**Participants argue with rich but inconsistent tactics; LLMs argue with consistent, low-variation logic.** Compared with LLMs, participants employ a wider range of strategies targeting the weaknesses of the tutor’s responses. For example, five participants challenged the tutor by claiming that the situations they described seem too ideal, while three participants tried to present direct counterexamples to the tutor’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 tutor’s responses, but focus more on creating interpretations that make the sentence logically valid. Their arguments are more logically consistent with claims followed by evidence, yet they lack variation in strategies and are often repetitive, as we found that LFTutor reminded the student for new examples in over 50% of simulated dialogues.

## 7 Conclusion

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

## 685 Limitation

686 **Limited Performance Gain on Guidance.** Al-  
687 though LFTutor managed to improve upon the  
688 BASE framework on all criteria during automatic  
689 evaluation, it did not achieve high scores ( $\geq 75\%$ )  
690 in Passive Guidance. We attribute this to the lack  
691 of structured, detailed, and progressive dialogue  
692 goals (e.g. first tackle the evidence, then tackle  
693 the warrant, then narrow down the discussion to  
694 certain aspects of the warrant.), which needs to be  
695 closely integrated with tutoring strategies to cre-  
696 ate responses that are robust to passive influence.  
697 Future works could introduce a list of high-level  
698 education goals and procedures for LFTutor to fol-  
699 low, such that it combines those goals and tutoring  
700 strategies to produce responses that actively control  
701 the flow of the dialogues.

702 **Scope of Evaluation Metrics** In this paper, we  
703 investigated problems with LLM teachers in LF  
704 educational dialogues and proposed an automatic  
705 evaluation framework with a curated set of eight  
706 evaluation metrics motivated by prior work and  
707 empirical relevance. While the proposed evalua-  
708 tion framework covers several key dimensions for  
709 measuring quality of LF education dialogues, such  
710 dialogues are inherently multi-faceted. Additional  
711 aspects, such as emotional support, personaliza-  
712 tion, and ease of understanding are not captured by  
713 our current metrics and they represent promising  
714 directions for future work.

## 715 References

716 Yvonne Berkle, Lukas Schmitt, Antonia Tolzin, An-  
717 dreas Janson, Thiemo Wambsganss, Jan Marco  
718 Leimeister, and Miriam Leuchter. 2023. [Measur-](#)  
719 [ing university students' ability to recognize argument](#)  
720 [structures and fallacies](#). *Frontiers in Psychology*, 14.

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

723 Gheorghe Comanici, Eric Bieber, Mike Schaekermann,  
724 Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Mar-  
725 cel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke  
726 Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni,  
727 Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson,  
728 Idan Szpektor, Nan-Jiang Jiang, Krishna Haridasan,  
729 Ahmed Omran, Nikunj Saunshi, Dara Bahri, Gau-  
730 rav Mishra, Eric Chu, Toby Boyd, Brad Hekman,  
731 Aaron Parisi, Chaoyi Zhang, Kornraphop Kawintira-  
732 non, Tania Bedrax-Weiss, Oliver Wang, Ya Xu, Ollie  
733 Purkiss, Uri Mendlovic, Ilai Deutel, Nam Nguyen,  
734 Adam Langley, Flip Korn, Lucia Rossazza, Alexan-  
735 dre Ramé, Sagar Waghmare, Helen Miller, Nathan

Byrd, Ashrith Sheshan, Raia Hadsell, Sangnie Bhard- 736  
waj, Pawel Janus, Tero Rissa, Dan Horgan, Alvin 737  
Abdagic, Lior Belenki, James Allingham, Anima 738  
Singh, Theo Guidroz, Srivatsan Srinivasan, Herman 739  
Schmit, Kristen Chiafullo, Andre Elisseeff, Nilpa Jha, 740  
Prateek Kolhar, Leonard Berrada, Frank Ding, Xi- 741  
ance Si, Shrestha Basu Mallick, Franz Och, Sofia 742  
Erell, Eric Ni, Tejasi Latkar, Sherry Yang, Petar 743  
Sirkovic, Ziqiang Feng, Robert Leland, Rachel Hor- 744  
nung, Gang Wu, Charles Blundell, Hamidreza Al- 745  
vari, Po-Sen Huang, Cathy Yip, Sanja Deur, Li Liu, 746  
Gabriela Surita, Pablo Duque, Dima Damen, Johnson 747  
Jia, Arthur Guez, Markus Mircea, Animesh Sinha, 748  
Alberto Magni, Paweł Stradomski, Tal Marian, Vlado 749  
Galić, Wenhua Chen, Hisham Husain, Achintya Sing- 750  
hal, Dominik Grewe, François-Xavier Aubet, Shuang 751  
Song, Lorenzo Blanco, Leland Rechis, Lewis Ho, 752  
Rich Munoz, Kelvin Zheng, Jessica Hamrick, Kevin 753  
Mather, Hagai Taitelbaum, Eliza Rutherford, Yun Lei, 754  
Kuanguan Chen, Anand Shukla, Erica Moreira, Eric 755  
Doi, Berivan Isik, Nir Shabat, Dominika Rogozińska, 756  
Kashyap Kolipaka, Jason Chang, Eugen Vušak, Srini- 757  
vasan Venkatachary, Shadi Noghabi, Tarun Bharti, 758  
Younghoon Jun, Aleksandr Zaks, Simon Green, Jesh- 759  
wanth Challagundla, William Wong, Muqthar Mo- 760  
hammad, Dean Hirsch, Yong Cheng, Iftekhar Naim, 761  
Lev Proleev, Damien Vincent, Aayush Singh, Maxim 762  
Krikun, Dilip Krishnan, Zoubin Ghahramani, Aviel 763  
Atias, Rajeev Aggarwal, Christo Kirov, Dimitrios 764  
Vytiniotis, Christy Koh, Alexandra Chronopoulou, 765  
Pawan Dogra, Vlad-Doru Ion, Gladys Tyen, Jason 766  
Lee, Felix Weissenberger, Trevor Strohman, Ashwin 767  
Balakrishna, Jack Rae, Marko Velic, Raoul de Liedek- 768  
erke, Oded Elyada, Wentao Yuan, Canoe Liu, Lior 769  
Shani, Sergey Kishchenko, Bea Alessio, Yandong Li, 770  
Richard Song, Sam Kwei, Orion Jankowski, Aneesh 771  
Pappu, Youhei Namiki, Yenai Ma, Nilesh Tripura- 772  
neni, Colin Cherry, Marissa Ikonomidis, Yu-Cheng 773  
Ling, Colin Ji, Beka Westberg, Auriel Wright, Da Yu, 774  
David Parkinson, Swaroop Ramaswamy, Jerome 775  
Connor, Soheil Hassas Yeganeh, Snchit Grover, 776  
George Kenwright, Lubo Litchev, Chris Apps, Alex 777  
Tomala, Felix Halim, Alex Castro-Ros, Zefei Li, 778  
Anudhyan Boral, Pauline Sho, Michal Yarom, Eric 779  
Malmi, David Klinghoffer, Rebecca Lin, Alan Ansell, 780  
Pradeep Kumar S, Shubin Zhao, Siqi Zuo, Adam San- 781  
toro, Heng-Tze Cheng, Solomon Demmessie, Yuchi 782  
Liu, Nicole Brichtova, Allie Culp, Nathaniel Braun, 783  
Dan Graur, Will Ng, Nikhil Mehta, Aaron Phillips, 784  
Patrik Sundberg, Varun Godbole, Fangyu Liu, Yash 785  
Katariya, David Rim, Mojtaba Seyedhosseini, Sean 786  
Ammirati, Jonas Valfridsson, Mahan Malihi, Tim- 787  
othy Knight, Andeep Toor, Thomas Lampe, Abe 788  
Ittycheriah, Lewis Chiang, Chak Yeung, Alexan- 789  
dre Fréchette, Jinneng Rao, Huisheng Wang, Hi- 790  
manshu Srivastava, Richard Zhang, Rocky Rhodes, 791  
Ariel Brand, Dean Weesner, Ilya Figotin, Felix Gi- 792  
meno, Rachana Fellingner, Pierre Marcenac, José Leal, 793  
Eyal Marcus, Victor Cotruta, Rodrigo Cabrera, Sh- 794  
eryl Luo, Dan Garrette, Vera Axelrod, Sorin Bal- 795  
tateanu, David Barker, Dongkai Chen, Horia Toma, 796  
Ben Ingram, Jason Riesa, Chinmay Kulkarni, Yu- 797  
jing Zhang, Hongbin Liu, Chao Wang, Martin Po- 798  
lacek, Will Wu, Kai Hui, Adrian N Reyes, Yi Su, 799

800	Megan Barnes, Ishaan Malhi, Anfal Siddiqui, Qixuan	864
801	Feng, Mihai Damaschin, Daniele Pighin, Andreas	865
802	Steiner, Samuel Yang, Ramya Sree Boppana, Simeon	866
803	Ivanov, Arun Kandoor, Aditya Shah, Asier Mujika,	867
804	Da Huang, Christopher A. Choquette-Choo, Mohak	868
805	Patel, Tianhe Yu, Toni Creswell, Jerry, Liu, Catarina	869
806	Barros, Yasaman Razeghi, Aurko Roy, Phil Culliton,	870
807	Binbin Xiong, Jiaqi Pan, Thomas Strohmman, Tolly	871
808	Powell, Babi Seal, Doug DeCarlo, Pranav Shyam,	872
809	Kaan Katircioglu, Xuezhi Wang, Cassidy Hardin, Im-	873
810	manuel Odisho, Josef Broder, Oscar Chang, Arun	874
811	Nair, Artem Shtefan, Maura O'Brien, Manu Agar-	875
812	wal, Sahitya Potluri, Siddharth Goyal, Amit Jhin-	876
813	dal, Saksham Thakur, Yury Stuken, James Lyon,	877
814	Kristina Toutanova, Fangxiaoyu Feng, Austin Wu,	878
815	Ben Horn, Alek Wang, Alex Cullum, Gabe Taubman,	879
816	Disha Shrivastava, Chongyang Shi, Hamish Tom-	880
817	linson, Roma Patel, Tao Tu, Ada Maksutaj Oflazer,	881
818	Francesco Pongetti, Mingyao Yang, Adrien Ali Taïga,	882
819	Vincent Perot, Nuo Wang Piere, Feng Han, Yoel	883
820	Drori, Iñaki Iturrate, Ayan Chakrabarti, Legg Ye-	884
821	ung, Dave Dopson, Yi ting Chen, Apoorv Kul-	885
822	shreshtha, Tongfei Guo, Philip Pham, Tal Schus-	886
823	ter, Junquan Chen, Alex Polozov, Jinwei Xing,	887
824	Huanjie Zhou, Praneeth Kacham, Doron Kuklian-	888
825	sky, Antoine Miech, Sergey Yaroshenko, Ed Chi,	889
826	Sholto Douglas, Hongliang Fei, Mathieu Blondel,	890
827	Preethi Myla, Lior Madmoni, Xing Wu, Daniel	891
828	Keysers, Kristian Kjems, Isabela Albuquerque, Li-	892
829	jun Yu, Joel D'sa, Michelle Plantan, Vlad Ionescu,	893
830	Jaume Sanchez Elias, Abhirut Gupta, Manish Reddy	894
831	Vuyyuru, Fred Alcober, Tong Zhou, Kaiyang Ji, Flor-	895
832	ian Hartmann, Subha Puttagunta, Hugo Song, Ehsan	896
833	Amid, Anca Stefanoiu, Andrew Lee, Paul Puccia-	897
834	relli, Emma Wang, Amit Raul, Slav Petrov, Isaac	898
835	Tian, Valentin Anklin, Nana Nti, Victor Gomes, Max	899
836	Schumacher, Grace Vesom, Alex Panagopoulos, Kon-	900
837	stantinos Bousmalis, Daniel Andor, Josh Jacob, Yuan	901
838	Zhang, Bill Rosgen, Matija Kecman, Matthew Tung,	902
839	Alexandra Belias, Noah Goodman, Paul Coving-	903
840	ton, Brian Wieder, Nikita Saxena, Elnaz Davoodi,	904
841	Muhuan Huang, Sharath Maddineni, Vincent Roulet,	905
842	Folawiyo Campbell-Ajala, Pier Giuseppe Sessa, Xin-	906
843	tian, Wu, Guangda Lai, Paul Collins, Alex Haig,	907
844	Vytenis Sakenas, Xiaowei Xu, Marissa Giustina, Lau-	908
845	rent El Shafey, Pichi Charoenpanit, Shefali Garg,	909
846	Joshua Ainslie, Boone Severson, Montse Gonzalez	910
847	Arenas, Shreya Pathak, Sujee Rajayogam, Jie Feng,	911
848	Michiel Bakker, Sheng Li, Nevan Wichers, Jamie	912
849	Rogers, Xinyang Geng, Yeqing Li, Rolf Jagerman,	913
850	Chao Jia, Nadav Olmert, David Sharon, Matthew	914
851	Mauger, Sandeep Mariserla, Hongxu Ma, Megha Mo-	915
852	habey, Kyuyeun Kim, Alek Andreev, Scott Pollom,	916
853	Juliette Love, Vihan Jain, Priyanka Agrawal, Yannick	917
854	Schroecker, Alisa Fortin, Manfred Warmuth, Ji Liu,	918
855	Andrew Leach, Irina Blok, Ganesh Poomal Girira-	919
856	jan, Roe Aharoni, Benigno Uria, Andrei Sozanschi,	920
857	Dan Goldberg, Lucian Ionita, Marco Tulio Ribeiro,	921
858	Martin Zlocha, Vighnesh Birodkar, Sami Lachgar,	922
859	Liangzhe Yuan, Himadri Choudhury, Matt Ginsberg,	923
860	Fei Zheng, Gregory Dibb, Emily Graves, Swachhand	924
861	Lokhande, Gabriel Rasskin, George-Cristian Muraru,	925
862	Corbin Quick, Sandeep Tata, Pierre Sermanet, Aditya	926
863	Chawla, Itay Karo, Yan Wang, Susan Zhang, Or-	
	gad Keller, Anca Dragan, Guolong Su, Ian Chou,	
	Xi Liu, Yiqing Tao, Shruthi Prabhakara, Marc Wil-	
	son, Ruibo Liu, Shibo Wang, Georgie Evans, David	
	Du, Alfonso Castaño, Gautam Prasad, Mona El	
	Mahdy, Sebastian Gerlach, Machel Reid, Jarrod	
	Kahn, Amir Zait, Thanumalayan Sankaranarayana	
	Pillai, Thatcher Ulrich, Guanyu Wang, Jan Wassen-	
	berg, Efrat Farkash, Kiran Yalasangi, Congchao	
	Wang, Maria Bauza, Simon Bucher, Ting Liu, Jun	
	Yan, Gary Leung, Vikas Sindhwani, Parker Barnes,	
	Avi Singh, Ivan Jurin, Jichuan Chang, Niket Kumar	
	Bhumihar, Sivan Eiger, Gui Citovsky, Ben With-	
	broe, Zhang Li, Siyang Xue, Niccolò Dal Santo,	
	Georgi Stoyanov, Yves Raimond, Steven Zheng,	
	Yilin Gao, Vít Listík, Sławek Kwasiborski, Rachel	
	Saputro, Adnan Ozturel, Ganesh Mallya, Kushal	
	Majmundar, Ross West, Paul Caron, Jinliang Wei,	
	Lluís Castrejon, Sharad Vikram, Deepak Ramachan-	
	dran, Nikhil Dhawan, Jiho Park, Sara Smoot, George	
	van den Driessche, Yochai Blau, Chase Malik, Wei	
	Liang, Roy Hirsch, Cicero Nogueira dos Santos,	
	Eugene Weinstein, Aäron van den Oord, Sid Lall,	
	Nicholas FitzGerald, Zixuan Jiang, Xuan Yang, Dale	
	Webster, Ali Elqursh, Aedan Pope, Georges Rot-	
	tival, David Raposo, Wanzheng Zhu, Jeff Dean,	
	Sami Alabed, Dustin Tran, Arushi Gupta, Zach	
	Gleicher, Jessica Austin, Edouard Rosseel, Megh	
	Umekar, Dipanjan Das, Yinghao Sun, Kai Chen,	
	Karolis Misiunas, Xiang Zhou, Yixian Di, Alyssa	
	Loo, Josh Newlan, Bo Li, Vinay Ramasesh, Ying Xu,	
	Alex Chen, Sudeep Gandhe, Radu Soricut, Nikita	
	Gupta, Shuguang Hu, Seliem El-Sayed, Xavier Gar-	
	cia, Idan Brusilovsky, Pu-Chin Chen, Andrew Bolt,	
	Lu Huang, Alex Gurney, Zhiying Zhang, Alexan-	
	der Pritzel, Jarek Wilkiewicz, Bryan Seybold, Bhar-	
	gav Kanagal Shamanna, Felix Fischer, Josef Dean,	
	Karan Gill, Ross McIlroy, Abhishek Bhowmick,	
	Jeremy Selier, Antoine Yang, Derek Cheng, Vladimir	
	Magay, Jie Tan, Dhriti Varma, Christian Walder,	
	Tomas Kocisky, Ryo Nakashima, Paul Natsev, Mike	
	Kwong, Ionel Gog, Chiyuan Zhang, Sander Diele-	
	man, Thomas Jimma, Andrey Ryabtsev, Siddhartha	
	Brahma, David Steiner, Dayou Du, Ante Žužul, Mis-	
	lav Žanić, Mukund Raghavachari, Willi Gierke, Zeyu	
	Zheng, Dessie Petrova, Yann Dauphin, Yuchuan	
	Liu, Ido Kessler, Steven Hand, Chris Duvarney,	
	Seokhwan Kim, Hyo Lee, Léonard Hussenot, Jef-	
	frey Hui, Josh Smith, Deepali Jain, Jiawei Xia, Gau-	
	rav Singh Tomar, Keyvan Amiri, Du Phan, Fabian	
	Fuchs, Tobias Weyand, Nenad Tomasev, Alexan-	
	dra Cordell, Xin Liu, Jonathan Mallinson, Pankaj	
	Joshi, Andy Crawford, Arun Suggala, Steve Chien,	
	Nick Fernando, Mariella Sanchez-Vargas, Duncan	
	Williams, Phil Crone, Xiyang Luo, Igor Karpov, Jyn	
	Shan, Terry Thurk, Robin Strudel, Paul Voigtlaen-	
	der, Piyush Patil, Tim Dozat, Ali Khodaei, Sahil	
	Singla, Piotr Ambroszczyk, Qiyin Wu, Yifan Chang,	
	Brian Roark, Chaitra Hegde, Tianli Ding, Angelos	
	Filos, Zhongru Wu, André Susano Pinto, Shuang	
	Liu, Saarthak Khanna, Aditya Pandey, Siobhan	
	Mcloughlin, Qiuqia Li, Sam Haves, Allan Zhou,	
	Elena Buchatskaya, Isabel Leal, Peter de Boursac,	
	Nami Akazawa, Nina Anderson, Terry Chen, Kr-	

927	ishna Somandepalli, Chen Liang, Sheela Goenka,	Renke Pan, Hoang Nguyen, Qiong, Hu, Sarah Perrin,	991
928	Stephanie Winkler, Alexander Grushetsky, Yifan	Abhinav Arora, Emily Pitler, Weiyi Wang, Kaushik	992
929	Ding, Jamie Smith, Fan Ye, Jordi Pont-Tuset, Eric	Shivakumar, Flavien Prost, Ben Limonchik, Jing	993
930	Li, Ruichao Li, Tomer Golany, Dawid Wegner,	Wang, Yi Gao, Timothee Cour, Shyamal Buch, Huan	994
931	Tao Jiang, Omer Barak, Yuan Shangguan, Eszter	Gui, Maria Ivanova, Philipp Neubeck, Kelvin Chan,	995
932	Vértes, Renee Wong, Jörg Bornschein, Alex Tu-	Lucy Kim, Huizhong Chen, Naman Goyal, Da-Woon	996
933	dor, Michele Bevilacqua, Tom Schaul, Ankit Singh	Chung, Lu Liu, Yao Su, Anastasia Petrushkina, Jiajun	997
934	Rawat, Yang Zhao, Kyriakos Axiotis, Lei Meng,	Shen, Armand Joulin, Yuanzhong Xu, Stein Xudong	998
935	Cory McLean, Jonathan Lai, Jennifer Beattie, Nate	Lin, Yana Kulizhskaya, Ciprian Chelba, Shobha Va-	999
936	Kushman, Yaxin Liu, Blair Kutzman, Fiona Lang,	sudevan, Eli Collins, Vasilisa Bashlovkina, Tony	1000
937	Jingchen Ye, Praneeth Netrapalli, Pushkar Mishra,	Lu, Doug Fritz, Jongbin Park, Yanqi Zhou, Chen	1001
938	Myriam Khan, Megha Goel, Rob Willoughby, David	Su, Richard Tanburn, Mikhail Sushkov, Michelle	1002
939	Tian, Honglei Zhuang, JD Chen, Zak Tsai, Tasos Ke-	Rasquinha, Jinning Li, Jennifer Prendki, Yiming Li,	1003
940	mentsietsidis, Arjun Khare, James Keeling, Keyang	Pallavi LV, Shriya Sharma, Hen Fitoussi, Hui Huang,	1004
941	Xu, Nathan Waters, Florent Altché, Ashok Papat,	Andrew Dai, Phuong Dao, Mike Burrows, Henry	1005
942	Bhavishya Mittal, David Saxton, Dalia El Badawy,	Prior, Danfeng Qin, Golan Pundak, Lars Lowe Sjoe-	1006
943	Michael Mathieu, Zheng Zheng, Hao Zhou, Nis-	sund, Art Khurshudov, Zhenkai Zhu, Albert Web-	1007
944	hant Ranka, Richard Shin, Qingnan Duan, Tim Sal-	son, Elizabeth Kemp, Tat Tan, Saurabh Agrawal,	1008
945	imans, Ioana Mihailescu, Uri Shaham, Ming-Wei	Susie Sargsyan, Liqun Cheng, Jim Stephen, Tom	1009
946	Chang, Yannis Assael, Nishanth Dikkala, Martin Iz-	Kwiatkowski, David Reid, Arunkumar Byravan, As-	1010
947	zard, Vincent Cohen-Addad, Cat Graves, Vlad Fein-	saf Hurwitz Michaely, Nicolas Heess, Luowei Zhou,	1011
948	berg, Grace Chung, DJ Strouse, Danny Karmon,	Sonam Goenka, Viral Carpenter, Anselm Levskaya,	1012
949	Sahand Sharifzadeh, Zoe Ashwood, Khiem Pham,	Bo Wang, Reed Roberts, Rémi Leblond, Sharat	1013
950	Jon Blanton, Alex Vasiloff, Jarred Barber, Mark	Chikkerur, Stav Ginzburg, Max Chang, Robert Ri-	1014
951	Geller, Aurick Zhou, Fedir Zubach, Tzu-Kuo Huang,	achi, Chuqiao, Xu, Zalán Borsos, Michael Pliskin,	1015
952	Lei Zhang, Himanshu Gupta, Matt Young, Julia	Julia Pawar, Morgane Lustman, Hannah Kirkwood,	1016
953	Proskurnia, Ronny Votel, Valentin Gabeur, Gabriel	Ankit Anand, Aditi Chaudhary, Norbert Kalb, Kieran	1017
954	Barcik, Aditya Tripathi, Hongkun Yu, Geng Yan,	Milan, Sean Augenstein, Anna Goldie, Laurel Prince,	1018
955	Beer Changpinyo, Filip Pavetić, Amy Coyle, Ya-	Karthik Raman, Yanhua Sun, Vivian Xia, Aaron	1019
956	suhisa Fujii, Jorge Gonzalez Mendez, Tianhao Zhou,	Cohen, Zhouyuan Huo, Josh Camp, Seher Ellis,	1020
957	Harish Rajamani, Blake Hechtman, Eddie Cao, Da-	Lukas Zilka, David Vilar Torres, Lisa Patel, Sho	1021
958	Cheng Juan, Yi-Xuan Tan, Valentin Dalibard, Yilun	Arora, Betty Chan, Jonas Adler, Kareem Ayoub,	1022
959	Du, Natalie Clay, Kaisheng Yao, Wenhao Jia, Dim-	Jacky Liang, Fayaz Jamil, Jiepu Jiang, Simon Baum-	1023
960	ple Vijaykumar, Yuxiang Zhou, Xinyi Bai, Wei-	gartner, Haitian Sun, Yael Karov, Yaroslav Akulov,	1024
961	Cih Hung, Steven Pecht, Georgi Todorov, Nikhil	Hui Zheng, Irene Cai, Claudio Fantacci, James	1025
962	Khadke, Pramod Gupta, Preethi Lahoti, Arnaud	Rubin, Alex Rav Acha, Mengchao Wang, Nina	1026
963	Autef, Karthik Duddu, James Lee-Thorp, Alexan-	D’Souza, Rohit Sathyanarayana, Shengyang Dai, Si-	1027
964	der Bykovsky, Tautvydas Misiunas, Sebastian Flen-	mon Rowe, Andrey Simanovsky, Omer Goldman,	1028
965	nerhag, Santhosh Thangaraj, Jed McGiffin, Zack	Yuheng Kuang, Xiaoyue Pan, Andrew Rosenberg,	1029
966	Nado, Markus Kunesch, Andreas Noever, Amir	Tania Rojas-Esponda, Praneet Dutta, Amy Zeng,	1030
967	Hertz, Marco Liang, Victor Stone, Evan Palmer,	Irina Jurenka, Greg Farquhar, Yamini Bansal, Shariq	1031
968	Samira Daruki, Arijit Pramanik, Siim Pöder, Austin	Iqbal, Becca Roelofs, Ga-Young Joung, Parker Beak,	1032
969	Kyker, Mina Khan, Evgeny Sluzhaev, Marvin Ritter,	Changwan Ryu, Ryan Poplin, Yan Wu, Jean-Baptiste	1033
970	Avraham Ruderman, Wenlei Zhou, Chirag Nagpal,	Alayrac, Senaka Buthpitiya, Olaf Ronneberger, Caleb	1034
971	Kiran Vodrahalli, George Necla, Paul Barham, El-	Habtegebriel, Wei Li, Paul Cavallaro, Aurora Wei,	1035
972	lie Pavlick, Jay Hartford, Izhak Shafran, Long Zhao,	Guy Bensky, Timo Denk, Harish Ganapathy, Jeff	1036
973	Maciej Mikuła, Tom Eccles, Hidetoshi Shimokawa,	Stanway, Pratik Joshi, Francesco Bertolini, Jessica	1037
974	Kanav Garg, Luke Vilnis, Hanwen Chen, Ilia Shu-	Lo, Olivia Ma, Zachary Charles, Geta Sampemane,	1038
975	mailov, Kuang-Huei Lee, Abdelrahman Abdelhamed,	Himanshu Sahni, Xu Chen, Harry Askham, David	1039
976	Meiyan Xie, Vered Cohen, Ester Hlavnova, Dan	Gaddy, Peter Young, Jiewen Tan, Matan Eyal, Arthur	1040
977	Malkin, Chawin Sitawarin, James Lottes, Pauline	Bražinskas, Li Zhong, Zhichun Wu, Mark Epstein,	1041
978	Coquinot, Tianli Yu, Sandeep Kumar, Jingwei Zhang,	Kai Bailey, Andrew Hard, Kamyu Lee, Sasha Gold-	1042
979	Aroma Mahendru, Zafarali Ahmed, James Martens,	shtein, Alex Ruiz, Mohammed Badawi, Matthias	1043
980	Tao Chen, Aviel Boag, Daiyi Peng, Coline Devin,	Lochbrunner, JK Kearns, Ashley Brown, Fabio Pardo,	1044
981	Arseniy Klimovskiy, Mary Phuong, Danny Vainstein,	Theophane Weber, Haichuan Yang, Pan-Pan Jiang,	1045
982	Jin Xie, Bhuvana Ramabhadran, Nathan Howard,	Berkin Akin, Zhao Fu, Marcus Wainwright, Chi Zou,	1046
983	Xinxin Yu, Gitartha Goswami, Jingyu Cui, Sam	Meenu Gaba, Pierre-Antoine Manzagol, Wendy Kan,	1047
984	Shleifer, Mario Pinto, Chih-Kuan Yeh, Ming-Hsuan	Yang Song, Karina Zainullina, Rui Lin, Jeongwoo Ko,	1048
985	Yang, Sara Javanmardi, Dan Ethier, Chace Lee, Jordi	Salil Deshmukh, Apoorv Jindal, James Svensson, Di-	1049
986	Orbay, Suyog Kotecha, Carla Bromberg, Pete Shaw,	vya Tyam, Heri Zhao, Christine Kaeser-Chen, Scott	1050
987	James Thornton, Adi Gerzi Rosenthal, Shane Gu,	Baird, Pooya Moradi, Jamie Hall, Qiuchen Guo, Vin-	1051
988	Matt Thomas, Ian Gemp, Aditya Ayyar, Asahi Ushio,	cent Tsang, Bowen Liang, Fernando Pereira, Suhas	1052
989	Aarush Selvan, Joel Wee, Chenxi Liu, Maryam Ma-	Ganesh, Ivan Korotkov, Jakub Adamek, Sridhar Thi-	1053
990	jzoubi, Weiren Yu, Jake Abernethy, Tyler Liechty,	agarajan, Vinh Tran, Charles Chen, Chris Tar, Sanil	1054

1055	Jain, Ishita Dasgupta, Taylan Bilal, David Reitter,	1118
1056	Kai Zhao, Giulia Vezzani, Yasmin Gehman, Pulkit	1119
1057	Mehta, Lauren Beltrone, Xerxes Dotiwalla, Sergio	1120
1058	Guadarrama, Zaheer Abbas, Stefani Karp, Petko	1121
1059	Georgiev, Chun-Sung Ferng, Marc Brockschmidt,	1122
1060	Liqian Peng, Christoph Hirnschall, Vikas Verma,	1123
1061	Yingying Bi, Ying Xiao, Avigail Dabush, Kelvin Xu,	1124
1062	Phil Wallis, Randall Parker, Qifei Wang, Yang Xu,	1125
1063	Ilkin Safarli, Dinesh Tewari, Yin Zhang, Seungyeon	1126
1064	Kim, Andrea Gasmundo, Mackenzie Thomas, Sergey	1127
1065	Levi, Ahmed Chowdhury, Kanishka Rao, Peter Garst,	1128
1066	Sam Conway-Rahman, Helen Ran, Kay McKinney,	1129
1067	Zhisheng Xiao, Wenhao Yu, Rohan Agrawal, Axel	1130
1068	Stjerngren, Catalin Ionescu, Jingjing Chen, Vivek	1131
1069	Sharma, Justin Chiu, Fei Liu, Ken Franko, Clayton	1132
1070	Sanford, Xingyu Cai, Paul Michel, Sanjay Ganapa-	1133
1071	thy, Jane Labanowski, Zachary Garrett, Ben Vargas,	1134
1072	Sean Sun, Bryan Gale, Thomas Buschmann, Guil-	1135
1073	laume Desjardins, Nimesh Ghelani, Palak Jain, Mudit	1136
1074	Verma, Chulayuth Asawaroengchai, Julian Eisensch-	1137
1075	los, Jitendra Harlalka, Hideto Kazawa, Don Metz-	1138
1076	zler, Joshua Howland, Ying Jian, Jake Ades, Viral	1139
1077	Shah, Tynan Gangwani, Seungji Lee, Roman Ring,	1140
1078	Steven M. Hernandez, Dean Reich, Amer Sinha,	1141
1079	Ashutosh Sathe, Joe Kovac, Ashleah Gill, Ajay Kan-	1142
1080	nan, Andrea D'olimpio, Martin Sevenich, Jay Whang,	1143
1081	Been Kim, Khe Chai Sim, Jilin Chen, Jiageng Zhang,	1144
1082	Shuba Lall, Yossi Matias, Bill Jia, Abe Friesen, Sara	1145
1083	Nasso, Ashish Thapliyal, Bryan Perozzi, Ting Yu,	1146
1084	Anna Shekhawat, Safeen Huda, Peter Grabowski,	1147
1085	Eric Wang, Ashwin Sreevatsa, Hilal Dib, Mehadi	1148
1086	Hassen, Parker Schuh, Vedrana Milutinovic, Chris	1149
1087	Welty, Michael Quinn, Ali Shah, Bangju Wang, Gabe	1150
1088	Barth-Maroon, Justin Frye, Natalie Axelsson, Tao Zhu,	1151
1089	Yukun Ma, Irene Giannoumis, Hanie Sedghi, Chang	1152
1090	Ye, Yi Luan, Kevin Aydin, Bilva Chandra, Vivek Sam-	1153
1091	pathkumar, Ronny Huang, Victor Lavrenko, Ahmed	1154
1092	Eleryan, Zhi Hong, Steven Hansen, Sara Mc Carthy,	1155
1093	Bidisha Samanta, Domagoj Čević, Xin Wang, Fang-	1156
1094	tao Li, Michael Voznesensky, Matt Hoffman, An-	1157
1095	dreas Terzis, Vikash Sehwaq, Gil Fidel, Luheng	1158
1096	He, Mu Cai, Yanzhang He, Alex Feng, Martin	1159
1097	Nikoltchev, Samrat Phatale, Jason Chase, Rory Law-	1160
1098	ton, Ming Zhang, Tom Ouyang, Manuel Tragut,	1161
1099	Mehdi Hafezi Manshadi, Arjun Narayanan, Jiaming	1162
1100	Shen, Xu Gao, Tolga Bolukbasi, Nick Roy, Xin Li,	1163
1101	Daniel Golovin, Liviu Panait, Zhen Qin, Guangxing	1164
1102	Han, Thomas Anthony, Sneha Kudugunta, Viorica Pa-	1165
1103	traucean, Aniket Ray, Xinyun Chen, Xiaochen Yang,	1166
1104	Tanuj Bhatia, Pranav Talluri, Alex Morris, Andrija	1167
1105	Ražnatović, Bethanie Brownfield, James An, Sheng	1168
1106	Peng, Patrick Kane, Ce Zheng, Nico Duduta, Joshua	1169
1107	Kessinger, James Noraky, Siqi Liu, Keran Rong,	1170
1108	Petar Veličković, Keith Rush, Alex Goldin, Fanny	1171
1109	Wei, Shiva Mohan Reddy Garlapati, Caroline Pantof-	1172
1110	faru, Okwan Kwon, Jianmo Ni, Eric Noland, Julia Di	1173
1111	Trapani, Françoise Beaufays, Abhijit Guha Roy, Yin-	1174
1112	lam Chow, Aybuke Turker, Geoffrey Cideron, Lan-	1175
1113	tao Mei, Jon Clark, Qingyun Dou, Matko Bošnjak,	1176
1114	Ralph Leith, Yuqing Du, Amir Yazdanbakhsh, Mil-	1177
1115	lad Nasr, Chester Kwak, Suraj Satishkumar Sheth,	1178
1116	Alex Kaskasoli, Ankesh Anand, Balaji Lakshmi-	1179
1117	narayanan, Sammy Jerome, David Bieber, Chun-Te	1180
	Chu, Alexandre Senges, Tianxiao Shen, Mukund	1181
	Sridhar, Ndaba Ndebele, Benjamin Beyret, Shakir	
	Mohamed, Mia Chen, Markus Freitag, Jiaxian Guo,	
	Luyang Liu, Paul Roit, Heng Chen, Shen Yan, Tom	
	Stone, JD Co-Reyes, Jeremy Cole, Salvatore Scellato,	
	Shekoofeh Azizi, Hadi Hashemi, Alicia Jin, Anand	
	Iyer, Marcella Valentine, András György, Arun	
	Ahuja, Daniel Hernandez Diaz, Chen-Yu Lee, Nathan	
	Clement, Weize Kong, Drew Garmon, Ishaan Watts,	
	Kush Bhatia, Khyatti Gupta, Matt Miecnikowski,	
	Hugo Vallet, Ankur Taly, Edward Loper, Saket Joshi,	
	James Atwood, Jo Chick, Mark Collier, Fotis Iliopou-	
	los, Ryan Trostle, Beliz Gunel, Ramiro Leal-Cavazos,	
	Arnar Mar Hrafnkelsson, Michael Guzman, Xiaoen	
	Ju, Andy Forbes, Jesse Emond, Kushal Chauhan,	
	Ben Caine, Li Xiao, Wenjun Zeng, Alexandre Mo-	
	ufarek, Daniel Murphy, Maya Meng, Nitish Gupta,	
	Felix Riedel, Anil Das, Elijah Lawal, Shashi Narayan,	
	Tiberiu Sosea, James Swirhun, Linda Friso, Behnam	
	Neyshabur, Jing Lu, Sertan Girgin, Michael Wunder,	
	Edouard Yvinec, Aroonlok Pyne, Victor Carbune,	
	Shruti Rijhwani, Yang Guo, Tulsee Doshi, Anton	
	Briukhov, Max Bain, Ayal Hitron, Xuanhui Wang,	
	Ashish Gupta, Ke Chen, Cosmo Du, Weiyang Zhang,	
	Dhruv Shah, Arjun Akula, Max Dylla, Ashyana	
	Kachra, Weicheng Kuo, Tingting Zou, Lily Wang,	
	Luyao Xu, Jifan Zhu, Justin Snyder, Sachit Menon,	
	Orhan Firat, Igor Mordatch, Yuan Yuan, Natalia	
	Ponomareva, Rory Blevins, Lawrence Moore, Weijun	
	Wang, Phil Chen, Martin Scholz, Artur Dwornik, Ja-	
	son Lin, Sicheng Li, Diego Antognini, Te I, Xiaodan	
	Song, Matt Miller, Uday Kalra, Adam Raveret, Os-	
	car Akerlund, Felix Wu, Andrew Nystrom, Namrata	
	Godbole, Tianqi Liu, Hannah DeBalsi, Jewel Zhao,	
	Bhuang Liu, Avi Caciularu, Lauren Lax, Urvashi	
	Khandelwal, Victoria Langston, Eric Bailey, Silvio	
	Lattanzi, Yufei Wang, Neel Kovelamudi, Sneha Mon-	
	dal, Guru Guruganesh, Nan Hua, Ofir Roval, Pawel	
	Wesołowski, Rishikesh Ingale, Jonathan Halcrow,	
	Tim Sohn, Christof Angermueller, Bahram Raad,	
	Eli Stickgold, Eva Lu, Alec Kosik, Jing Xie, Timoth-	
	ey Lillicrap, Austin Huang, Lydia Lihui Zhang, Do-	
	minik Paulus, Clement Farabet, Alex Wertheim, Bing	
	Wang, Rishabh Joshi, Chu ling Ko, Yonghui Wu,	
	Shubham Agrawal, Lily Lin, XiangHai Sheng, Pe-	
	ter Sung, Tyler Breland-King, Christina Butterfield,	
	Swapnil Gawde, Sumeet Singh, Qiao Zhang, Raj	
	Apte, Shilpa Shetty, Adrian Hutter, Tao Li, Elizabeth	
	Salesky, Federico Lebron, Jonni Kanerva, Michela	
	Paganini, Arthur Nguyen, Rohith Vallu, Jan-Thorsten	
	Peter, Sarmishta Velury, David Kao, Jay Hoover,	
	Anna Bortsova, Colton Bishop, Shoshana Jakobovits,	
	Alessandro Agostini, Alekh Agarwal, Chang Liu,	
	Charles Kwong, Sasan Tavakkol, Ioana Bica, Alex	
	Greve, Anirudh GP, Jake Marcus, Le Hou, Tom	
	Duerig, Rivka Moroshko, Dave Lacey, Andy Davis,	
	Julien Amelot, Guohui Wang, Frank Kim, Theofilos	
	Strinopoulos, Hui Wan, Charline Le Lan, Shankar Kr-	
	ishnan, Haotian Tang, Peter Humphreys, Junwen Bai,	
	Idan Heimlich Shtacher, Diego Machado, Chenxi	
	Pang, Ken Burke, Dangyi Liu, Renga Aravamud-	
	han, Yue Song, Ed Hirst, Abhimanyu Singh, Brendan	
	Jou, Liang Bai, Francesco Piccinno, Chuyuan Kelly	
	Fu, Robin Alazard, Barak Meiri, Daniel Winter,	

1182	Charlie Chen, Mingda Zhang, Jens Heitkaemper,	Maggioni, Jing Xiong, Lukas Haas, Jenny Brennan,	1245
1183	John Lambert, Jinhyuk Lee, Alexander Frömmgen,	Aishwarya Kamath, Nikolay Savinov, Arsha Nagrani,	1246
1184	Sergey Rogulenko, Pranav Nair, Paul Niemczyk, An-	Trevor Yacovone, Ryan Kappedal, Kostas Andri-	1247
1185	ton Bulynov, Bibo Xu, Hadar Shemtov, Morteza	opoulos, Li Lao, YaGuang Li, Grigory Rozhdestven-	1248
1186	Zadimoghaddam, Serge Toropov, Mateo Wirth, Han-	skiy, Kazuma Hashimoto, Andrew Audibert, Sophia	1249
1187	jun Dai, Sreenivas Gollapudi, Daniel Zheng, Alex	Austin, Daniel Rodriguez, Anian Ruoss, Garrett	1250
1188	Kurakin, Chansoo Lee, Kalesha Bullard, Nicolas	Honke, Deep Karkhanis, Xi Xiong, Qing Wei,	1251
1189	Serrano, Ivana Balazevic, Yang Li, Johan Schalk-	James Huang, Zhaoqi Leng, Vittal Premachandran,	1252
1190	wyk, Mark Murphy, Mingyang Zhang, Kevin Se-	Stan Bileschi, Georgios Evangelopoulos, Thomas	1253
1191	queira, Romina Datta, Nishant Agrawal, Charles Sut-	Mensink, Jay Pavagadhi, Denis Teplyashin, Paul	1254
1192	ton, Nithya Attaluri, Mencher Chiang, Wael Farhan,	Chang, Linting Xue, Garrett Tanzer, Sally Gold-	1255
1193	Gregory Thornton, Kate Lin, Travis Choma, Hung	man, Kaushal Patel, Shixin Li, Jeremy Wiesner, Ivy	1256
1194	Nguyen, Kingshuk Dasgupta, Dirk Robinson, Iulia	Zheng, Ian Stewart-Binks, Jie Han, Zhi Li, Liangchen	1257
1195	Comşa, Michael Riley, Arjun Pillai, Basil Mustafa,	Luo, Karel Lenc, Mario Lučić, Fuzhao Xue, Ryan	1258
1196	Ben Golan, Amir Zandieh, Jean-Baptiste Lespiau,	Mullins, Alexey Guseynov, Chung-Ching Chang,	1259
1197	Billy Porter, David Ross, Sujeevan Rajayogam, Mo-	Isaac Galatzer-Levy, Adam Zhang, Garrett Bingham,	1260
1198	hit Agarwal, Subhashini Venugopalan, Bobak Shahri-	Grace Hu, Ale Hartman, Yue Ma, Jordan Griffith,	1261
1199	ari, Qiqi Yan, Hao Xu, Taylor Tobin, Pavel Dubov,	Alex Irpan, Carey Radebaugh, Summer Yue, Lijie	1262
1200	Hongzhi Shi, Adrià Recasens, Anton Kovsharov, Se-	Fan, Victor Ungureanu, Christina Sorokin, Hannah	1263
1201	bastian Borgeaud, Lucio Dery, Shanthal Vasanth,	Teufel, Peiran Li, Rohan Anil, Dimitris Paparas, Todd	1264
1202	Elena Gribovskaya, Linhai Qiu, Mahdis Mahdieh,	Wang, Chu-Cheng Lin, Hui Peng, Megan Shum,	1265
1203	Wojtek Skut, Elizabeth Nielsen, CJ Zheng, Adams	Goran Petrovic, Demetra Brady, Richard Nguyen,	1266
1204	Yu, Carrie Grimes Bostock, Shaleen Gupta, Aaron	Klaus Macherey, Zhihao Li, Harman Singh, Mad-	1267
1205	Archer, Chris Rawles, Elinor Davies, Alexey Svy-	havi Yenugula, Mariko Iinuma, Xinyi Chen, Kavya	1268
1206	atkovskiy, Tomy Tsai, Yoni Halpern, Christian Reiss-	Kopparapu, Alexey Stern, Shachi Dave, Chandu	1269
1207	wig, Bartek Wydrowski, Bo Chang, Joan Puigcerver,	Thekkath, Florence Perot, Anurag Kumar, Fangda	1270
1208	Mor Hazan Taege, Jian Li, Eva Schnider, Xinjian	Li, Yang Xiao, Matthew Bilotti, Mohammad Hos-	1271
1209	Li, Dragos Dena, Yunhan Xu, Umesh Telang, Tianze	sein Bateni, Isaac Noble, Lisa Lee, Amelio Vázquez-	1272
1210	Shi, Heiga Zen, Kyle Kastner, Yeongil Ko, Neesha	Reina, Julian Salazar, Xiaomeng Yang, Boyu Wang,	1273
1211	Subramaniam, Aviral Kumar, Pete Blois, Zhuyun	Ela Gruzewska, Anand Rao, Sindhu Raghuram,	1274
1212	Dai, John Wieting, Yifeng Lu, Yoel Zeldes, Tian	Zheng Xu, Eyal Ben-David, Jieru Mei, Sid Dalmia,	1275
1213	Xie, Anja Hauth, Alexandru Țifrea, Yuqi Li, Sam	Zhaoyi Zhang, Yuchen Liu, Gagan Bansal, Helena	1276
1214	El-Husseini, Dan Abolafia, Howard Zhou, Wen Ding,	Pankov, Steven Schwarcz, Andrea Burns, Christine	1277
1215	Sahra Ghalebikesabi, Carlos Guía, Andrii Maksai,	Chan, Sumit Sanghai, Ricky Liang, Ethan Liang, An-	1278
1216	Ágoston Weisz, Sercan Arik, Nick Sukhanov, Aga	toine He, Amy Stuart, Arun Narayanan, Yukun Zhu,	1279
1217	Świetlik, Xuhui Jia, Luo Yu, Weiyue Wang, Mark	Christian Frank, Bahar Fatemi, Amit Sabne, Oran	1280
1218	Brand, Dawn Bloxwich, Sean Kirmani, Zhe Chen,	Lang, Indro Bhattacharya, Shane Settle, Maria Wang,	1281
1219	Alec Go, Pablo Sprechmann, Nithish Kannen, Alen	Brendan McMahan, Andrea Tacchetti, Livio Baldini	1282
1220	Carin, Paramjit Sandhu, Isabel Edkins, Leslie Noot-	Soares, Majid Hadian, Serkan Cabi, Timothy Chung,	1283
1221	boom, Jai Gupta, Loren Maggiore, Javad Azizi,	Nikita Putikhin, Gang Li, Jeremy Chen, Austin	1284
1222	Yael Pritch, Pengcheng Yin, Mansi Gupta, Danny	Tarango, Henryk Michalewski, Mehran Kazemi,	1285
1223	Tarlow, Duncan Smith, Desi Ivanov, Mohammad	Hussain Masoom, Hila Sheftel, Rakesh Shivanna,	1286
1224	Babaeizadeh, Ankita Goel, Satish Kambala, Grace	Archita Vadali, Ramona Comanescu, Doug Reid,	1287
1225	Chu, Matej Kastelic, Michelle Liu, Hagen Soltau,	Joss Moore, Arvind Neelakantan, Michaël Sander,	1288
1226	Austin Stone, Shivani Agrawal, Min Kim, Kedar	Jonathan Herzig, Aviv Rosenberg, Mostafa Dehghani,	1289
1227	Soparkar, Srinivas Tadepalli, Oskar Bunyan, Rachel	JD Choi, Michael Fink, Reid Hayes, Eric Ge, Shi-	1290
1228	Soh, Arvind Kannan, DY Kim, Blake JianHang Chen,	tao Weng, Chia-Hua Ho, John Karro, Kalpesh Kr-	1291
1229	Afief Halumi, Sudeshna Roy, Yulong Wang, Ol-	ishna, Lam Nguyen Thiet, Amy Skerry-Ryan, Daniel	1292
1230	can Sercinoglu, Gena Gibson, Sijal Bhatnagar, Mo-	Eppens, Marco Andreetto, Navin Sarma, Silvano	1293
1231	toki Sano, Daniel von Dincklage, Qingchun Ren,	Bonacina, Burcu Karagol Ayan, Megha Nawhal,	1294
1232	Blagoj Mitrevski, Mirek Olšák, Jennifer She, Carl	Zhihao Shan, Mike Dusenberry, Shantanu Thakoor,	1295
1233	Doersch, Jilei, Wang, Bingyuan Liu, Qijun Tan,	Sagar Gubbi, Duc Dung Nguyen, Reut Tsarfaty,	1296
1234	Tamar Yakar, Tris Warkentin, Alex Ramirez, Carl	Samuel Albanie, Jovana Mitrović, Meet Gandhi, Bo-	1297
1235	Lebsack, Josh Dillon, Rajiv Mathews, Tom Cobley,	Juen Chen, Alessandro Epasto, Georgi Stephanov,	1298
1236	Zelin Wu, Zhuoyuan Chen, Jon Simon, Swaroop	Ye Jin, Samuel Gehman, Aida Amini, Jack Weber,	1299
1237	Nath, Tara Sainath, Alexei Bendebury, Ryan Julian,	Feryal Behbahani, Shawn Xu, Miltos Allamanis,	1300
1238	Bharath Mankalale, Daria Ćurko, Paulo Zacchello,	Xi Chen, Myle Ott, Claire Sha, Michal Jastrzebski,	1301
1239	Adam R. Brown, Kiranbir Sodhia, Heidi Howard,	Hang Qi, David Greene, Xinyi Wu, Abodunrinwa	1302
1240	Sergi Caelles, Abhinav Gupta, Gareth Evans, Anna	Toki, Daniel Vlasic, Jane Shapiro, Ragha Kotikala-	1303
1241	Bulanova, Lesley Katzen, Roman Goldenberg, An-	pudi, Zhe Shen, Takaaki Saeki, Sirui Xie, Albin Cas-	1304
1242	ton Tsitsulin, Joe Stanton, Benoit Schillings, Vitaly	sirer, Shikhar Bharadwaj, Tatsuya Kiyono, Srinadh	1305
1243	Kovalev, Corey Fry, Rushin Shah, Kuo Lin, Shyam	Bhojanapalli, Elan Rosenfeld, Sam Ritter, Jieming	1306
1244	Upadhyay, Cheng Li, Soroush Radpour, Marcello	Mao, João Gabriel Oliveira, Zoltan Egyed, Bernd	1307
		Bandemer, Emilio Parisotto, Keisuke Kinoshita, Juli-	1308

1309	ette Pluto, Petros Maniatis, Steve Li, Yaohui Guo,	1372
1310	Golnaz Ghiasi, Jean Tarbouriech, Srimon Chatter-	1373
1311	jee, Julie Jin, Katrina, Xu, Jennimaria Palomaki, Séb	1374
1312	Arnold, Madhavi Sewak, Federico Piccinini, Mohit	1375
1313	Sharma, Ben Albrecht, Sean Purser-haskell, Ashwin	1376
1314	Vaswani, Chongyan Chen, Matheus Wisniewski, Qin	1377
1315	Cao, John Aslanides, Nguyet Minh Phu, Maximilian	1378
1316	Sieb, Lauren Agubuzu, Anne Zheng, Daniel Sohn,	1379
1317	Marco Selvi, Anders Andreassen, Krishan Subudhi,	1380
1318	Prem Eruvbetine, Oliver Woodman, Tomas Mery,	1381
1319	Sebastian Krause, Xiaoqi Ren, Xiao Ma, Jincheng	1382
1320	Luo, Dawn Chen, Wei Fan, Henry Griffiths, Christian	1383
1321	Schuler, Alice Li, Shujian Zhang, Jean-Michel Sarr,	1384
1322	Shixin Luo, Riccardo Patana, Matthew Watson, Dani	1385
1323	Naboulsi, Michael Collins, Sailesh Sidhwani, Emiel	1386
1324	Hoogeboom, Sharon Silver, Emily Caveness, Xi-	1387
1325	aokai Zhao, Mikel Rodriguez, Maxine Deines, Libin	1388
1326	Bai, Patrick Griffin, Marco Tagliasacchi, Emily Xue,	1389
1327	Spandana Raj Babbula, Bo Pang, Nan Ding, Gloria	1390
1328	Shen, Elijah Peake, Remi Crocker, Shubha Sriniv-	1391
1329	as Raghvendra, Danny Swisher, Woohyun Han,	1392
1330	Richa Singh, Ling Wu, Vladimir Pchelin, Tsendsuren	1393
1331	Munkhdalai, Dana Alon, Geoff Bacon, Efren Robles,	1394
1332	Jannis Bulian, Melvin Johnson, George Powell, Fe-	1395
1333	lipe Tiengo Ferreira, Yaoyiran Li, Frederik Benzing,	1396
1334	Mihajlo Velimirović, Hubert Soyer, William Kong,	1397
1335	Tony, Nguyễn, Zhen Yang, Jeremiah Liu, Joost van	1398
1336	Amersfoort, Daniel Gillick, Baochen Sun, Nathalie	1399
1337	Rauschmayr, Katie Zhang, Serena Zhan, Tao Zhou,	1400
1338	Alexey Frolov, Chengrun Yang, Denis Vnukov, Louis	1401
1339	Rouillard, Hongji Li, Amol Mandhane, Nova Fallen,	1402
1340	Rajesh Venkataraman, Clara Huiyi Hu, Jennifer Bren-	1403
1341	nan, Jenny Lee, Jerry Chang, Martin Sundermeyer,	1404
1342	Zhufeng Pan, Rosemary Ke, Simon Tong, Alex Fab-	1405
1343	rikant, William Bono, Jindong Gu, Ryan Foley, Yi-	1406
1344	iran Mao, Manolis Delakis, Dhruva Bhaswar, Roy	1407
1345	Frostig, Nick Li, Avital Zipori, Cath Hope, Olga	1408
1346	Kozlova, Swaroop Mishra, Josip Djolonga, Craig	1409
1347	Schiff, Majd Al Merey, Eleftheria Briakou, Peter	1410
1348	Morgan, Andy Wan, Avinatan Hassidim, RJ Skerry-	1411
1349	Ryan, Kuntal Sengupta, Mary Jasarevic, Praveen	1412
1350	Kallakuri, Paige Kunkle, Hannah Brennan, Tom	1413
1351	Lieber, Hassan Mansoor, Julian Walker, Bing Zhang,	1414
1352	Annie Xie, Goran Žužić, Adaeze Chukwuka, Alex	1415
1353	Druinsky, Donghyun Cho, Rui Yao, Ferjad Naem,	1416
1354	Shiraz Butt, Eunyoung Kim, Zhipeng Jia, Mandy	1417
1355	Jordan, Adam Lelkes, Mark Kurzeja, Sophie Wang,	1418
1356	James Zhao, Andrew Over, Abhishek Chakladar,	1419
1357	Marcel Prasetya, Neha Jha, Sriram Ganapathy, Yale	1420
1358	Cong, Prakash Shroff, Carl Saroufim, Sobhan Miry-	1421
1359	oosefi, Mohamed Hammad, Tajwar Nasir, Weijuan	1422
1360	Xi, Yang Gao, Young Maeng, Ben Hora, Chin-Yi	1423
1361	Cheng, Parisa Haghani, Yoad Lewenberg, Caden Lu,	1424
1362	Martin Matysiak, Naina Raisinghani, Huiyu Wang,	1425
1363	Lexi Baugher, Rahul Sukthankar, Minh Giang, John	1426
1364	Schultz, Noah Fiedel, Minmin Chen, Cheng-Chun	1427
1365	Lee, Tapomay Dey, Hao Zheng, Shachi Paul, Celine	1428
1366	Smith, Andy Ly, Yicheng Wang, Rishabh Bansal,	1429
1367	Bartek Perz, Susanna Ricco, Stasha Blank, Vaishakh	1430
1368	Keshava, Deepak Sharma, Marvin Chow, Kunal	1431
1369	Lad, Komal Jalan, Simon Osindero, Craig Swans-	1432
1370	on, Jacob Scott, Anastasija Ilić, Xiaowei Li, Sid-	1433
1371	dhartha Reddy Jonnalagadda, Afzal Shama Souda-	1434
	gar, Yan Xiong, Bat-Orgil Batsaikhan, Daniel Jar-	1435
	rett, Naveen Kumar, Maulik Shah, Matt Lawlor,	
	Austin Waters, Mark Graham, Rhys May, Sabela	
	Ramos, Sandra Lefdal, Zeynep Cankara, Nacho	
	Cano, Brendan O'Donoghue, Jed Borovik, Frederick	
	Liu, Jordan Grimstad, Mahmoud Alnahlawi, Kate-	
	rina Tsihlas, Tom Hudson, Nikolai Grigorev, Yil-	
	ing Jia, Terry Huang, Tobenna Peter Igwe, Sergei	
	Lebedev, Xiaodan Tang, Igor Krivokon, Frankie	
	Garcia, Melissa Tan, Eric Jia, Peter Stys, Shikhar	
	Vashishth, Yu Liang, Balaji Venkatraman, Chenjie	
	Gu, Anastasios Kementsietsidis, Chen Zhu, June-	
	hyuk Jung, Yunfei Bai, Mohammad Javad Hosseini,	
	Faruk Ahmed, Aditya Gupta, Xin Yuan, Shereen	
	Ashraf, Shitij Nigam, Gautam Vasudevan, Pranjal	
	Awasthi, Adi Mayrav Gilady, Zeld Mariet, Ramy	
	Eskander, Haiguang Li, Hexiang Hu, Guillermo	
	Garrido, Philippe Schlattner, George Zhang, Rohun	
	Saxena, Petar Dević, Kritika Muralidharan, Ashwin	
	Murthy, Yiqian Zhou, Min Choi, Arissa Wongpanich,	
	Zhengdong Wang, Premal Shah, Yuntao Xu, Yiling	
	Huang, Stephen Spencer, Alice Chen, James Co-	
	han, Junjie Wang, Jonathan Tompson, Junru Wu,	
	Ruba Haroun, Haiqiong Li, Blanca Huerdo, Fan	
	Yang, Tongxin Yin, James Wendt, Michael Bender-	
	sky, Rahma Chaabouni, Javier Snaider, Johan Fer-	
	ret, Abhishek Jindal, Tara Thompson, Andrew Xue,	
	Will Bishop, Shubham Milind Phal, Archit Sharma,	
	Yunhsuan Sung, Prabakar Radhakrishnan, Mo Shom-	
	rat, Reeve Ingle, Roopali Vij, Justin Gilmer, Mi-	
	hai Dorin Istin, Sam Sobell, Yang Lu, Emily Not-	
	tage, Dorsa Sadigh, Jeremiah Willcock, Tingnan	
	Zhang, Steve Xu, Sasha Brown, Katherine Lee, Gary	
	Wang, Yun Zhu, Yi Tay, Cheolmin Kim, Audrey	
	Gutierrez, Abhanshu Sharma, Yongqin Xian, Sungy-	
	ong Seo, Claire Cui, Elena Pochernina, Cip Baetu,	
	Krzysztof Jastrzębski, Mimi Ly, Mohamed Elhawaty,	
	Dan Suh, Eren Sezener, Pidong Wang, Nancy Yuen,	
	George Tucker, Jiahao Cai, Zuguang Yang, Cindy	
	Wang, Alex Muzio, Hai Qian, Jae Yoo, Derek Lock-	
	hart, Kevin R. McKee, Mandy Guo, Malika Mehro-	
	tra, Artur Mendonça, Sanket Vaibhav Mehta, Sherry	
	Ben, Chetan Tekur, Jiaqi Mu, Muye Zhu, Victoria	
	Krakovna, Hongrae Lee, AJ Maschinot, Sébastien	
	Cevey, HyunJeong Choe, Aijun Bai, Hansa Sriniv-	
	asan, Derek Gasaway, Nick Young, Patrick Siegler,	
	Dan Holtmann-Rice, Vihari Piratla, Kate Baumli,	
	Roey Yogeve, Alex Hofer, Hado van Hasselt, Svetlana	
	Grant, Yuri Chervonyi, David Silver, Andrew Hogue,	
	Ayushi Agarwal, Kathie Wang, Preeti Singh, Four	
	Flynn, Josh Lipschultz, Robert David, Lizzetth Bel-	
	lot, Yao-Yuan Yang, Long Le, Filippo Graziano, Kate	
	Olszewska, Kevin Hui, Akanksha Maurya, Nikos	
	Parotsidis, Weijie Chen, Tayo Oguntebi, Joe Kel-	
	ley, Anirudh Baddepudi, Johannes Mauerer, Gre-	
	gory Shaw, Alex Siegman, Lin Yang, Shravya Shetty,	
	Subhrajit Roy, Yunting Song, Wojciech Stokowiec,	
	Ryan Burnell, Omkar Savant, Robert Busa-Fekete,	
	Jin Miao, Samrat Ghosh, Liam MacDermed, Phillip	
	Lippe, Mikhail Dektiarev, Zach Behrman, Fabian	
	Mentzer, Kelvin Nguyen, Meng Wei, Siddharth	
	Verma, Chris Knutsen, Sudeep Dasari, Zhipeng Yan,	
	Petr Mitrichev, Xingyu Wang, Virat Shejwalkar, Ja-	
	cob Austin, Srinivas Sunkara, Navneet Potti, Yan	

1436	Virin, Christian Wright, Gaël Liu, Oriana Riva, Etienne Pot, Greg Kochanski, Quoc Le, Gargi Balasubramaniam, Arka Dhar, Yuguo Liao, Adam Błoniarczyk, Divyansh Shukla, Elizabeth Cole, Jong Lee, Sheng Zhang, Sushant Kafle, Siddharth Vashishtha, Parsa Mahmoudieh, Grace Chen, Raphael Hoffmann, Pranesh Srinivasan, Agustin Dal Lago, Yoav Ben Shalom, Zi Wang, Michael Elabd, Anuj Sharma, Junhyuk Oh, Suraj Kothawade, Maigo Le, Marianne Monteiro, Shentao Yang, Kaiz Alarakya, Robert Geirhos, Diana Mincu, Håvard Garnes, Hayato Kobayashi, Soroosh Mariooryad, Kacper Krasowiak, Zhixin, Lai, Shibl Mourad, Mingqiu Wang, Fan Bu, Ophir Aharoni, Guanjie Chen, Abhimanyu Goyal, Vadim Zubov, Ankur Bapna, Elahe Dabir, Nisarg Kothari, Kay Lamerigts, Nicola De Cao, Jeremy Shar, Christopher Yew, Nitish Kulkarni, Dre Mahaarachchi, Mandar Joshi, Zhenhai Zhu, Jared Lichtarge, Yichao Zhou, Hannah Muckenhirn, Vittorio Selo, Oriol Vinyals, Peter Chen, Anthony Brohan, Vaibhav Mehta, Sarah Cogan, Ruth Wang, Ty Geri, Wei-Jen Ko, Wei Chen, Fabio Viola, Keshav Shivam, Lisa Wang, Madeleine Clare Elish, Raluca Ada Popa, Sébastien Pereira, Jianqiao Liu, Raphael Koster, Donnie Kim, Gufeng Zhang, Sayna Ebrahimi, Partha Talukdar, Yanyan Zheng, Petra Poklukar, Ales Mikhalap, Dale Johnson, Anitha Vijayakumar, Mark Omernick, Matt Dibb, Ayush Dubey, Qiong Hu, Apurv Suman, Vaibhav Aggarwal, Ilya Kornakov, Fei Xia, Wing Lowe, Alexey Kolganov, Ted Xiao, Vitaly Nikolaev, Steven Hemingray, Bonnie Li, Joana Iljazi, Mikołaj Rybiński, Ballie Sandhu, Peggy Lu, Thang Luong, Rodolphe Jenatton, Vineetha Govindaraj, Hui, Li, Gabriel Dulac-Arnold, Wonpyo Park, Henry Wang, Abhinav Modi, Jean Pouget-Abadie, Kristina Greller, Rahul Gupta, Robert Berry, Prajit Ramachandran, Jinyu Xie, Liam McCafferty, Jianling Wang, Kilol Gupta, Hyeontaek Lim, Blaž Bratanič, Andy Brock, Ilija Akolzin, Jim Sproch, Dan Karliner, Duhyeon Kim, Adrian Goedeckemeyer, Noam Shazeer, Cordelia Schmid, Daniele Calandriello, Parul Bhatia, Krzysztof Choromanski, Ceslee Montgomery, Dheeru Dua, Ana Ramalho, Helen King, Yue Gao, Lynn Nguyen, David Lindner, Divya Pitta, Oleaser Johnson, Khalid Salama, Diego Ardila, Michael Han, Erin Farnese, Seth Odoom, Ziyue Wang, Xiangzhuo Ding, Norman Rink, Ray Smith, Harshal Tushar Lehri, Eden Cohen, Neera Vats, Tong He, Parthasarathy Gopavarapu, Adam Paszke, Miteyan Patel, Wouter Van Gansbeke, Lucia Loher, Luis Castro, Maria Voitovich, Tamara von Glehn, Nelson George, Simon Niklaus, Zach Eaton-Rosen, Nemanja Rakićević, Erik Jue, Sagi Perel, Carrie Zhang, Yuval Bahat, Angéline Pouget, Zhi Xing, Fantine Huot, Ashish Shenoy, Taylor Bos, Vincent Coriou, Bryan Richter, Natasha Noy, Yaqing Wang, Santiago Ontanon, Siyang Qin, Gleb Makarchuk, Demis Hassabis, Zhuowan Li, Mandar Sharma, Kumar Venkatesan, Iurii Kemaev, Roxanne Daniel, Shiyu Huang, Saloni Shah, Octavio Ponce, Warren, Chen, Manaal Faruqui, Jialin Wu, Slavica Andračić, Szabolcs Payrits, Daniel McDuff, Tom Hume, Yuan Cao, MH Tessler, Qingze Wang, Yinan Wang, Ivor Rendulic, Eirikur Agustsson, Matthew Johnson,	Tanya Lando, Andrew Howard, Sri Gayatri Sundara Padmanabhan, Mayank Daswani, Andrea Banino, Michael Kilgore, Jonathan Heek, Ziwei Ji, Alvaro Caceres, Conglong Li, Nora Kassner, Alexey Vlaskin, Zeyu Liu, Alex Grills, Yanhan Hou, Roykrong Sukkerd, Gowoon Cheon, Nishita Shetty, Larisa Markeeva, Piotr Stanczyk, Tejas Iyer, Yuan Gong, Shawn Gao, Keerthana Gopalakrishnan, Tim Blyth, Malcolm Reynolds, Avishkar Bhoopchand, Misha Bilenko, Dero Gharibian, Vicky Zayats, Aleksandra Faust, Abhinav Singh, Min Ma, Hongyang Jiao, Sudheendra Vijayanarasimhan, Lora Aroyo, Vikas Yadav, Sarah Chakera, Ashwin Kakarla, Vilobh Meshram, Karol Gregor, Gabriela Botea, Evan Senter, Dawei Jia, Geza Kovacs, Neha Sharma, Sebastien Baur, Kai Kang, Yifan He, Lin Zhuo, Marija Kostelac, Itay Laish, Songyou Peng, Louis O'Bryan, Daniel Kasenberg, Girish Ramchandra Rao, Edouard Leurent, Biao Zhang, Sage Stevens, Ana Salazar, Ye Zhang, Ivan Lobov, Jake Walker, Allen Porter, Morgan Redshaw, Han Ke, Abhishek Rao, Alex Lee, Hoi Lam, Michael Moffitt, Jaeyoun Kim, Siyuan Qiao, Terry Koo, Robert Dadashi, Xinying Song, Mukund Sundararajan, Peng Xu, Chizu Kawamoto, Yan Zhong, Clara Barbu, Apoorv Reddy, Mauro Verzetti, Leon Li, George Papamakarios, Hanna Klimczak-Plucińska, Mary Cassin, Koray Kavukcuoglu, Rigel Swavely, Alain Vaucher, Jeffrey Zhao, Ross Hemsley, Michael Tschannen, Heming Ge, Gaurav Menghani, Yang Yu, Natalie Ha, Wei He, Xiao Wu, Maggie Song, Rachel Sterneck, Stefan Zinke, Dan A. Calian, Annie Marsden, Alejandro Cruzado Ruiz, Matteo Hessel, Almog Gueta, Benjamin Lee, Brian Farris, Manish Gupta, Yunjie Li, Mohammad Saleh, Vedant Misra, Kefan Xiao, Piermaria Mendolicchio, Gavin Buttimore, Varvara Krayvanova, Nigamaa Nayakanti, Matthew Wiethoff, Yash Pande, Azalia Mirhoseini, Ni Lao, Jasmine Liu, Yiqing Hua, Angie Chen, Yury Malkov, Dmitry Kalashnikov, Shubham Gupta, Kartik Audhkhasi, Yuexiang Zhai, Sudhindra Kopalle, Prateek Jain, Eran Ofek, Clemens Meyer, Khuslen Baatarsukh, Hana Strejček, Jun Qian, James Freedman, Ricardo Figueira, Michal Sokolik, Olivier Bachem, Raymond Lin, Dia Kharrat, Chris Hidey, Pingmei Xu, Dennis Duan, Yin Li, Muge Ersoy, Richard Everett, Kevin Cen, Rebeca Santamaria-Fernandez, Amir Taubenfeld, Ian Mackinnon, Linda Deng, Polina Zablotskaia, Shashank Viswanadha, Shivanker Goel, Damion Yates, Yunxiao Deng, Peter Choy, Mingqing Chen, Abhishek Sinha, Alex Mossin, Yiming Wang, Arthur Szlam, Susan Hao, Paul Kishan Rubenstein, Metin Toksoz-Exley, Miranda Aperghis, Yin Zhong, Junwhan Ahn, Michael Isard, Olivier Lacombe, Florian Luisier, Chrysovalantis Anastasiou, Yogesh Kalley, Utsav Prabhu, Emma Dunleavy, Shaan Bijwadia, Justin Mao-Jones, Kelly Chen, Rama Pasumarthi, Emily Wood, Adil Dostmohamed, Nate Hurley, Jiri Simsa, Alicia Parrish, Mantas Pajarskas, Matt Harvey, Ondrej Skopek, Yony Kochinski, Javier Rey, Verena Rieser, Denny Zhou, Sun Jae Lee, Trilok Acharya, Guowang Li, Joe Jiang, Xiaofan Zhang, Bryant Gipson, Ethan Mahintorabi, Marco Gelmi, Nima Khajehnouri, Angel Yeh, Kayi Lee, Loic Matthey, Leslie Baker, Trang Pham, Han	1500 1501 1502 1503 1504 1505 1506 1507 1508 1509 1510 1511 1512 1513 1514 1515 1516 1517 1518 1519 1520 1521 1522 1523 1524 1525 1526 1527 1528 1529 1530 1531 1532 1533 1534 1535 1536 1537 1538 1539 1540 1541 1542 1543 1544 1545 1546 1547 1548 1549 1550 1551 1552 1553 1554 1555 1556 1557 1558 1559 1560 1561 1562 1563
------	---	--	--

1564	Fu, Alex Pak, Prakhhar Gupta, Cristina Vasconcelos, Adam Sadovsky, Brian Walker, Sissie Hsiao,	Borja De Balle Pigem, Tayfun Terzi, Gellért Weisz,	1628
1565	Patrik Zochbauer, Andreea Marzoca, Noam Velan,	Dipankar Ghosh, Aditi Mavalankar, Dhruv Madeka,	1629
1566	Junhao Zeng, Gilles Baechler, Danny Driess, Divya	Kaspar Daugaard, Hartwig Adam, Viraj Shah, Dana	1630
1567	Jain, Yanping Huang, Lizzie Tao, John Maggs, Nir	Berman, Maggie Tran, Steven Baker, Ewa An-	1631
1568	Levine, Jon Schneider, Erika Gemzer, Samuel Petit,	drejczuk, Grishma Chole, Ganna Raboshchuk, Mahdi	1632
1569	Shan Han, Zach Fisher, Dustin Zelle, Courtney Biles,	Mirzazadeh, Thais Kagohara, Shimu Wu, Chris-	1633
1570	Eugene Ie, Asya Fadeeva, Casper Liu, Juliana Vi-	tian Schallhart, Bernett Orlando, Chen Wang, Al-	1634
1571	cente Franco, Adrian Collister, Hao Zhang, Renshen	ban Rrustemi, Hao Xiong, Hao Liu, Arpi Vezer,	1635
1572	Wang, Ruizhe Zhao, Leandro Kieliger, Kurt Shus-	Nolan Ramsden, Shuo yin Chang, Sidharth Mud-	1636
1573	ter, Rui Zhu, Boqing Gong, Lawrence Chan, Ruoxi	gal, Yan Li, Nino Vieillard, Yedid Hoshen, Farooq	1637
1574	Sun, Sujoy Basu, Roland Zimmermann, Jamie Hayes,	Ahmad, Ambrose Slone, Amy Hua, Natan Potikha,	1638
1575	Abhishek Bapna, Jasper Snoek, Weel Yang, Puran-	Mirko Rossini, Jon Stritar, Sushant Prakash, Zifeng	1639
1576	jay Datta, Jad Al Abdallah, Kevin Kilgour, Lu Li,	Wang, Xuanyi Dong, Alireza Nazari, Efrat Nehoran,	1640
1577	SQ Mah, Yennie Jun, Morgane Rivière, Abhijit Kar-	Kaan Tekelioglu, Yinxiao Li, Kartikeya Badola, Tom	1641
1578	markar, Tammo Spalink, Tao Huang, Lucas Gon-	Funkhouser, Yuanzhen Li, Varun Yerram, Ramya	1642
1579	zalez, Duc-Hieu Tran, Averi Nowak, John Palow-	Ganeshan, Daniel Formoso, Karol Langner, Tian	1643
1580	itch, Martin Chadwick, Ellie Talius, Harsh Mehta,	Shi, Huijian Li, Yumeya Yamamori, Amayika Panda,	1644
1581	Thibault Sellam, Philipp Fränken, Massimo Nicosia,	Alaa Saade, Angelo Scorza Scarpato, Chris Breaux,	1645
1582	Kyle He, Aditya Kini, David Amos, Sugato Basu,	CJ Carey, Zongwei Zhou, Cho-Jui Hsieh, Sophie	1646
1583	Harrison Jobe, Eleni Shaw, Qiantong Xu, Colin	Bridgers, Alena Butryna, Nishesh Gupta, Vaibhav	1647
1584	Evans, Daisuke Ikeda, Chaochao Yan, Larry Jin, Lun	Tulshyan, Sanghyun Woo, Evgenii Eltyshhev, Will	1648
1585	Wang, Sachin Yadav, Ilia Labzovsky, Ramesh Sam-	Grathwohl, Chanel Parks, Seth Benjamin, Rina Pan-	1649
1586	path, Ada Ma, Candice Schumann, Aditya Siddhant,	igrahy, Shenil Dodhia, Daniel De Freitas, Chris	1650
1587	Rohin Shah, John Youssef, Rishabh Agarwal, Na-	Sauer, Will Song, Ferran Alet, Jackson Tolins, Cos-	1651
1588	talie Dabney, Alessio Tonioni, Moran Ambar, Jing	min Paduraru, Xingyi Zhou, Brian Albert, Zizhao	1652
1589	Li, Isabelle Guyon, Benny Li, David Soergel, Boya	Zhang, Lei Shu, Mudit Bansal, Sarah Nguyen, Amir	1653
1590	Fang, Georgi Karadzhov, Cristian Udrescu, Trieu	Globerson, Owen Xiao, James Manyika, Tom Hen-	1654
1591	Trinh, Vikas Raunak, Seb Noury, Dee Guo, Sonal	nigan, Rong Rong, Josip Matak, Anton Bakalov,	1655
1592	Gupta, Mara Finkelstein, Denis Petek, Lihao Liang,	Ankur Sharma, Danila Sinopalnikov, Andrew Pier-	1656
1593	Greg Billock, Pei Sun, David Wood, Yiwen Song,	son, Stephen Roller, Geoff Brown, Mingcen Gao,	1657
1594	Xiaobin Yu, Tatiana Matejovicova, Regev Cohen,	Toshiyuki Fukuzawa, Amin Ghafouri, Kenny Vas-	1658
1595	Kalyan Andra, David D'Ambrosio, Zhiwei Deng,	sigh, Iain Barr, Zhicheng Wang, Anna Korsun, Ra-	1659
1596	Vincent Nallatamby, Ebrahim Songhori, Rumen Dan-	jesh Jayaram, Lijie Ren, Tim Zaman, Samira Khan,	1660
1597	govski, Andrew Lampinen, Pankil Botadra, Adam	Yana Lunts, Dan Deutsch, Dave Uthus, Nitzan Katz,	1661
1598	Hillier, Jiawei Cao, Nagabhushan Baddi, Adhi Kun-	Masha Samsikova, Amr Khalifa, Nikhil Sethi, Jiao	1662
1599	coro, Toshihiro Yoshino, Ankit Bhagatwala, Mar-	Sun, Luming Tang, Uri Alon, Xianghong Luo, Dian	1663
1600	caùrelio Ranzato, Rylan Schaeffer, Tianlin Liu, Shuai	Yu, Abhishek Nayyar, Bryce Petrini, Will Truong,	1664
1601	Ye, Obaid Sarvana, John Nham, Chenkai Kuang,	Vincent Hellendoorn, Nikolai Chinaev, Chris Alberti,	1665
1602	Isabel Gao, Jinoo Baek, Shubham Mittal, Ayzaan	Wei Wang, Jingcao Hu, Vahab Mirrokni, Ananth	1666
1603	Wahid, Anita Gergely, Bin Ni, Josh Feldman, Carrie	Balashankar, Avia Aharon, Aahil Mehta, Ahmet Is-	1667
1604	Muir, Pascal Lamblin, Wolfgang Macherey, Ethan	cen, Joseph Kready, Lucas Manning, Anhad Mo-	1668
1605	Dyer, Logan Kilpatrick, Víctor Campos, Mukul	hananey, Yuankai Chen, Anshuman Tripathi, Allen	1669
1606	Bhutani, Stanislav Fort, Yanif Ahmad, Aliaksei	Wu, Igor Petrovski, Dawsen Hwang, Martin Baeuml,	1670
1607	Severyn, Kleopatra Chatziprimou, Oleksandr Fer-	Shreyas Chandrakaladharan, Yuan Liu, Rey Coaguila,	1671
1608	ludin, Mason Dimarco, Aditya Kusupati, Joe Hey-	Maxwell Chen, Sally Ma, Pouya Tafti, Susheel Tati-	1672
1609	ward, Dan Bahir, Kevin Villela, Katie Millican, Dror	neni, Terry Spitz, Jiayu Ye, Paul Vicol, Mihaela	1673
1610	Marcus, Sanaz Bahargam, Caglar Unlu, Nicholas	Rosca, Adrià Puigdomènech, Zohar Yahav, San-	1674
1611	Roth, Zichuan Wei, Siddharth Gopal, Deepanway	jay Ghemawat, Hanzhao Lin, Phoebe Kirk, Zaid	1675
1612	Ghoshal, Edward Lee, Sharon Lin, Jennie Lees,	Nabulsi, Sergey Brin, Bernd Bohnet, Ken Caluwaerts,	1676
1613	Dayeong Lee, Anahita Hosseini, Connie Fan, Seth	Aditya Srikanth Veerubhotla, Dan Zheng, Zihang	1677
1614	Neel, Marcus Wu, Yasemin Altun, Honglong Cai,	Dai, Petre Petrov, Yichong Xu, Ramin Mehran, Zhuo	1678
1615	Enrique Piqueras, Josh Woodward, Alessandro Bis-	Xu, Luisa Zintgraf, Jiho Choi, Spurthi Amba Hom-	1679
1616	sacco, Salem Haykal, Mahyar Bordbar, Prasha Sun-	baiah, Romal Thoppilan, Sashank Reddi, Lukasz	1680
1617	daram, Sarah Hodgkinson, Daniel Toyama, George	Lew, Li Li, Kellie Webster, KP Sawhney, Lam-	1681
1618	Polovets, Austin Myers, Anu Sinha, Tomer Levin-	pros Lamprou, Siamak Shakeri, Mayank Lunay-	1682
1619	boim, Kashyap Krishnakumar, Rachita Chhparia,	ach, Jianmin Chen, Sumit Bagri, Alex Salcianu,	1683
1620	Tatiana Sholokhova, Nitesh Bharadwaj Gundavarapu,	Ying Chen, Yani Donchev, Charlotte Magister, Signe	1684
1621	Ganesh Jawahar, Haroon Qureshi, Jieru Hu, Nikola	Nørly, Vitor Rodrigues, Tomas Izo, Hila Noga,	1685
1622	Momchev, Matthew Rahtz, Renjie Wu, Aishwarya P	Joe Zou, Thomas Köppe, Wenxuan Zhou, Kenton	1686
1623	S, Kedar Dhamdhere, Meiqi Guo, Umang Gupta, Ali	Lee, Xiangzhu Long, Danielle Eisenbud, Antho-	1687
1624	Eslami, Mariano Schain, Michiel Blokzijl, David	ny Chen, Connor Schenck, Chi Ming To, Peilin	1688
1625	Welling, Dave Orr, Levent Bolelli, Nicolas Perez-	Zhong, Emanuel Taropa, Minh Truong, Omer Levy,	1689
1626	Nieves, Mikhail Sirotenko, Aman Prasad, Arjun Kar,	Danilo Martins, Zhiyuan Zhang, Christopher Sem-	1690
1627		turs, Kelvin Zhang, Alex Yakubovich, Pol Moreno,	1691

1692	Lara McConnaughey, Di Lu, Sam Redmond, Lotte	herme Tubone, Ravin Kumar, Jason Baldrige, Fe-	1756
1693	Weerts, Yonatan Bitton, Tiziana Refice, Nicolas La-	lix Hernandez-Campos, Angeliki Lazaridou, James	1757
1694	casse, Arthur Conmy, Corentin Tallec, Julian Odell,	Besley, Itay Yona, Neslihan Bulut, Quentin Wellens,	1758
1695	Hannah Forbes-Pollard, Arkadiusz Socala, Jonathan	AJ Pierigiovanni, Jasmine George, Richard Green,	1759
1696	Hoech, Pushmeet Kohli, Alanna Walton, Rui Wang,	Pu Han, Connie Tao, Geoff Clark, Chong You, Ab-	1760
1697	Mikita Sazanovich, Kexin Zhu, Andrei Kapishnikov,	bas Abdolmaleki, Justin Fu, Tongzhou Chen, Ashwin	1761
1698	Rich Galt, Matthew Denton, Ben Murdoch, Caitlin	Chaugule, Angad Chandorkar, Altaf Rahman, Will	1762
1699	Sikora, Kareem Mohamed, Wei Wei, Uri First, Tim	Thompson, Penporn Koanantakool, Mike Bernico,	1763
1700	McConnell, Luis C. Cobo, James Qin, Thi Avra-	Jie Ren, Andrey Vlasov, Sergei Vassilvitskii, Maciej	1764
1701	hami, Daniel Balle, Yu Watanabe, Annie Louis,	Kula, Yizhong Liang, Dahun Kim, Yangsibo	1765
1702	Adam Kraft, Setareh Ariafar, Yiming Gu, Eugénie	Huang, Chengxi Ye, Dmitry Lepikhin, and Wesley	1766
1703	Rives, Charles Yoon, Andrei Rusu, James Cobon-	Helmholz. 2025. <a href="#">Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities</a> . <i>Preprint</i> , arXiv:2507.06261.	1767
1704	Kerr, Chris Hahn, Jiaming Luo, Yuvein, Zhu, Ni-		1768
1705	harika Ahuja, Rodrigo Benenson, Raphaël Lopez		1769
1706	Kaufman, Honglin Yu, Lloyd Hightower, Junlin		1770
1707	Zhang, Darren Ni, Lisa Anne Hendricks, Gabby		
1708	Wang, Gal Yona, Lalit Jain, Pablo Barrio, Surya	Vasco Correia. 2011. Biases and fallacies: The role of	1771
1709	Bhupatiraju, Siva Velusamy, Allan Dafoe, Sebastian	motivated irrationality in fallacious reasoning. <i>Co-</i>	1772
1710	Riedel, Tara Thomas, Zhe Yuan, Mathias Bellaïche,	<i>gency</i> , 3:107–118.	1773
1711	Sheena Panthaplackel, Klemen Kloboves, Sarthak		
1712	Jauhari, Canfer Akbulut, Todor Davchev, Evgeny	Arthur L. Costa. 2001. <i>Developing minds: a resource</i>	1774
1713	Gladchenko, David Madras, Aleksandr Chuklin, Ty-	<i>book for teaching thinking</i> . Association for Supervi-	1775
1714	rone Hill, Quan Yuan, Mukundan Madhavan, Luke	sion and Curriculum Development.	1776
1715	Leonhard, Dylan Scandinaro, Qihang Chen, Ning		
1716	Niu, Arthur Douillard, Bogdan Damoc, Yasumasa	DeepSeek-AI. 2025. <a href="#">Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning</a> . <i>Preprint</i> , arXiv:2501.12948.	1777
1717	Onoe, Fabian Pedregosa, Fred Bertsch, Chas Leich-		1778
1718	ner, Joseph Pagadora, Jonathan Malmaud, Sameera		1779
1719	Ponda, Andy Twigg, Oleksii Duzhyi, Jingwei Shen,	Holden Ed and Schmit Ed. 2002. Inquiry and the lit-	1780
1720	Miaosen Wang, Roopal Garg, Jing Chen, Utku Evcı,	erary text: Constructing discussions in the english	1781
1721	Jonathan Lee, Leon Liu, Koji Kojima, Masa Ya-	classroom. <i>classroom practices in teaching english</i> .	1782
1722	maguchi, Arunkumar Rajendran, AJ Pierigiovanni,		
1723	Vinodh Kumar Rajendran, Marco Fornoni, Gabriel	Zheng Fang, Lama Alqazlan, Du Liu, Yulan He, and	1783
1724	Ibagon, Harry Ragan, Sadh MNM Khan, John Blitzer,	Rob Procter. 2023. <a href="#">A user-centered, interactive, human-in-the-loop topic modelling system</a> . In <i>Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics</i> , pages 505–522, Dubrovnik, Croatia. Association for Computational Linguistics.	1784
1725	Andrew Bunner, Guan Sun, Takahiro Kosakai, Scott		1785
1726	Lundberg, Ndidi Elue, Kelvin Guu, SK Park, Jane		1786
1727	Park, Arunachalam Narayanaswamy, Chengda Wu,		1787
1728	Jayaram Mudigonda, Trevor Cohn, Hairong Mu, Ravi		1788
1729	Kumar, Laura Graesser, Yichi Zhang, Richard Kil-		1789
1730	lam, Vincent Zhuang, Mai Giménez, Wael Al Jishi,	Max Glockner, Yufang Hou, Preslav Nakov, and Iryna	1790
1731	Ruy Ley-Wild, Alex Zhai, Kazuki Osawa, Diego	Gurevych. 2025. <a href="#">Grounding fallacies misrepresenting scientific publications in evidence</a> . In <i>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)</i> , pages 9732–9767, Albuquerque, New Mexico. Association for Computational Linguistics.	1791
1732	Cedillo, Jialu Liu, Mayank Upadhyay, Marcin Sie-		1792
1733	niak, Roshan Sharma, Tom Paine, Anelia Angelova,		1793
1734	Sravanti Addepalli, Carolina Parada, Kingshuk Ma-		1794
1735	jumder, Avery Lamp, Sanjiv Kumar, Xiang Deng, Ar-		1795
1736	tiom Myaskovsky, Tea Sabolić, Jeffrey Dudek, Sarah		1796
1737	York, Félix de Chaumont Quitry, Jiazhong Nie, Dee		1797
1738	Cattle, Alok Gunjan, Bilal Piot, Waleed Khawaja,		1798
1739	Seojin Bang, Simon Wang, Siavash Khodadadeh,		1799
1740	Raghavender R, Praynaa Rawlani, Richard Powell,	Pierpaolo Goffredo, Mariana Chaves, Serena Villata,	1800
1741	Kevin Lee, Johannes Griesser, GS Oh, Cesar Mag-	and Elena Cabrio. 2023. <a href="#">Argument-based detection and classification of fallacies in political debates</a> . In <i>Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing</i> , pages 11101–11112, Singapore. Association for Computational Linguistics.	1801
1742	alhaes, Yujia Li, Simon Tokumine, Hadas Natalie		1802
1743	Vogel, Dennis Hsu, Arturo BC, Disha Jindal, Matan		1803
1744	Cohen, Zi Yang, Junwei Yuan, Dario de Cesare, Tony		1804
1745	Bruguier, Jun Xu, Monica Roy, Alon Jacovi, Dan		1805
1746	Belov, Rahul Arya, Phoenix Meadowlark, Shlomi		
1747	Cohen-Ganor, Wenting Ye, Patrick Morris-Suzuki,	Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri,	1806
1748	Praseem Banzal, Gan Song, Pranavaraj Ponnuramu,	Abhinav Pandey, Abhishek Kadian, Ahmad Al-	1807
1749	Fred Zhang, George Scrivener, Salah Zaiem, Alif Ra-	Dahle, Aiesha Letman, Akhil Mathur, Alan Schel-	1808
1750	ditya Rochman, Kehang Han, Badih Ghazi, Kate Lee,	ten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh	1809
1751	Shahar Drath, Daniel Suo, Antonious Girgis, Pradeep	Goyal, Anthony Hartshorn, Aobo Yang, Archi Mi-	1810
1752	Shenoy, Duy Nguyen, Douglas Eck, Somit Gupta,	tra, Archie Sravankumar, Artem Korenev, Arthur	1811
1753	Le Yan, Joao Carreira, Anmol Gulati, Ruoxin Sang,	Hinsvark, Arun Rao, Aston Zhang, Aurelien Ro-	1812
1754	Daniil Milylenka, Emma Cooney, Edward Chou,	driguez, Austen Gregerson, Ava Spataru, Baptiste	1813
1755	Mingyang Ling, Cindy Fan, Ben Coleman, Guil-		

1814	Roziere, Bethany Biron, Binh Tang, Bobbie Chern,	Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao,	1878
1815	Charlotte Caucheteux, Chaya Nayak, Chloe Bi,	Zacharie Delpierre Coudert, Zheng Yan, Zhengxing	1879
1816	Chris Marra, Chris McConnell, Christian Keller,	Chen, Zoe Papakipos, Aaditya Singh, Aayushi Sri-	1880
1817	Christophe Touret, Chunyang Wu, Corinne Wong,	vastava, Abha Jain, Adam Kelsey, Adam Shajnfeld,	1881
1818	Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Al-	Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand,	1882
1819	lonsius, Daniel Song, Danielle Pintz, Danny Livshits,	Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei	1883
1820	Danny Wyatt, David Esiobu, Dhruv Choudhary,	Baevski, Allie Feinstein, Amanda Kallet, Amit San-	1884
1821	Dhruv Mahajan, Diego Garcia-Olano, Diego Perino,	gani, Amos Teo, Anam Yunus, Andrei Lupu, And-	1885
1822	Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy,	dres Alvarado, Andrew Caples, Andrew Gu, Andrew	1886
1823	Elina Lobanova, Emily Dinan, Eric Michael Smith,	Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchan-	1887
1824	Filip Radenovic, Francisco Guzmán, Frank Zhang,	dani, Annie Dong, Annie Franco, Anuj Goyal, Aparaj-	1888
1825	Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis An-	jita Saraf, Arkabandhu Chowdhury, Ashley Gabriel,	1889
1826	derson, Govind Thattai, Graeme Nail, Gregoire Mi-	Ashwin Bharambe, Assaf Eisenman, Azadeh Yaz-	1890
1827	alon, Guan Pang, Guillem Cucurell, Hailey Nguyen,	dan, Beau James, Ben Maurer, Benjamin Leonhardi,	1891
1828	Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan	Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi	1892
1829	Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Is-	Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Han-	1893
1830	han Misra, Ivan Evtimov, Jack Zhang, Jade Copet,	cock, Bram Wasti, Brandon Spence, Brani Stojkovic,	1894
1831	Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park,	Brian Gamido, Britt Montalvo, Carl Parker, Carly	1895
1832	Jay Mahadeokar, Jeet Shah, Jelmer van der Linde,	Burton, Catalina Mejia, Ce Liu, Changhan Wang,	1896
1833	Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu,	Changkyu Kim, Chao Zhou, Chester Hu, Ching-	1897
1834	Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang,	Hsiang Chu, Chris Cai, Chris Tindal, Christoph Fe-	1898
1835	Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park,	ichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty,	1899
1836	Joseph Rocca, Joshua Johnstun, Joshua Saxe, Jun-	Daniel Kreymer, Daniel Li, David Adkins, David	1900
1837	teng Jia, Kalyan Vasuden Alwala, Karthik Prasad,	Xu, Davide Testuggine, Delia David, Devi Parikh,	1901
1838	Kartikaya Upasani, Kate Plawiak, Ke Li, Kenneth	Diana Liskovich, Didem Foss, Ding Kang Wang, Duc	1902
1839	Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer,	Le, Dustin Holland, Edward Dowling, Eissa Jamil,	1903
1840	Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal	Elaine Montgomery, Eleonora Presani, Emily Hahn,	1904
1841	Lakhota, Lauren Rantala-Yearly, Laurens van der	Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban	1905
1842	Maaten, Lawrence Chen, Liang Tan, Liz Jenkins,	Arcaute, Evan Dunbar, Evan Smothers, Fei Sun,	1906
1843	Louis Martin, Lovish Madaan, Lubo Malo, Lukas	Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat	1907
1844	Blecher, Lukas Landzaat, Luke de Oliveira, Madeline	Ozgenel, Francesco Caggioni, Frank Kanayet, Frank	1908
1845	Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar	Seide, Gabriela Medina Florez, Gabriella Schwarz,	1909
1846	Paluri, Marcin Kardas, Maria Tsimploukelli, Mathew	Gada Badeer, Georgia Swee, Gil Halpern, Grant	1910
1847	Oldham, Mathieu Rita, Maya Pavlova, Melanie Kam-	Herman, Grigory Sizov, Guangyi, Zhang, Guna	1911
1848	badur, Mike Lewis, Min Si, Mitesh Kumar Singh,	Lakshminarayanan, Hakan Inan, Hamid Shojanaz-	1912
1849	Mona Hassan, Naman Goyal, Narjes Torabi, Niko-	eri, Han Zou, Hannah Wang, Hanwen Zha, Haroun	1913
1850	lay Bashlykov, Nikolay Bogoychev, Niladri Chatterji,	Habeeb, Harrison Rudolph, Helen Suk, Henry As-	1914
1851	Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick	pegren, Hunter Goldman, Hongyuan Zhan, Ibrahim	1915
1852	Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vas-	Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis,	1916
1853	sic, Peter Weng, Prajjwal Bhargava, Pratik Dubal,	Irina-Elena Veliche, Itai Gat, Jake Weissman, James	1917
1854	Praveen Krishnan, Punit Singh Koura, Puxin Xu,	Geboski, James Kohli, Janice Lam, Japhet Asher,	1918
1855	Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj	Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jen-	1919
1856	Ganapathy, Ramon Calderer, Ricardo Silveira Cabral,	nifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy	1920
1857	Robert Stojnic, Roberta Raileanu, Rohan Maheswari,	Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe	1921
1858	Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ron-	Cummings, Jon Carvill, Jon Shepard, Jonathan Mc-	1922
1859	nie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan	Phie, Jonathan Torres, Josh Ginsburg, Junjie Wang,	1923
1860	Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sa-	Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khan-	1924
1861	hana Chennabasappa, Sanjay Singh, Sean Bell, Seo-	delwal, Katayoun Zand, Kathy Matosich, Kaushik	1925
1862	hyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sha-	Veeraraghavan, Kelly Michelena, Keqian Li, Ki-	1926
1863	ran Narang, Sharath Rapparthi, Sheng Shen, Shengye	ran Jagadeesh, Kun Huang, Kunal Chawla, Kyle	1927
1864	Wan, Shruti Bhosale, Shun Zhang, Simon Van-	Huang, Lailin Chen, Lakshya Garg, Lavender A,	1928
1865	denhede, Soumya Batra, Spencer Whitman, Sten	Leandro Silva, Lee Bell, Lei Zhang, Liangpeng	1929
1866	Sootla, Stephane Collot, Suchin Gururangan, Syd-	Guo, Licheng Yu, Liron Moshkovich, Luca Wehrst-	1930
1867	ney Borodinsky, Tamar Herman, Tara Fowler, Tarek	edt, Madian Khabza, Manav Avalani, Manish Bhatt,	1931
1868	Sheasha, Thomas Georgiou, Thomas Scialom, Tobias	Martynas Mankus, Matan Hasson, Matthew Lennie,	1932
1869	Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal	Matthias Reso, Maxim Groshev, Maxim Naumov,	1933
1870	Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh	Maya Lathi, Meghan Keneally, Miao Liu, Michael L.	1934
1871	Ramanathan, Viktor Kerkez, Vincent Gouget, Vir-	Seltzer, Michal Valko, Michelle Restrepo, Mihir Pa-	1935
1872	ginie Do, Vish Vogeti, Vitor Albiero, Vladan Petro-	tel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark,	1936
1873	vic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney	Mike Macey, Mike Wang, Miquel Jubert Hermoso,	1937
1874	Meers, Xavier Martinet, Xiaodong Wang, Xi-	Mo Metanat, Mohammad Rastegari, Munish Bansal,	1938
1875	aofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xin-	Nandhini Santhanam, Natascha Parks, Natasha	1939
1876	feng Xie, Xuchao Jia, Xuwei Wang, Yaelle Gold-	White, Navyata Bawa, Nayan Singhal, Nick Egebo,	1940
1877	schlag, Yashesh Gaur, Yasmine Babaei, Yi Wen,	Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich	1941



2058	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.	2112
2059		2113
2060		2114
2061		2115
2062		2116
2063	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.	2117
2064		2118
2065		2119
2066		2120
2067	Marilyn Nippold. 2023. <a href="#">Unlocking logical fallacies: A key to building critical thinking skills in adolescents</a> . <i>Perspectives of the ASHA Special Interest Groups</i> , 9:1–13.	2121
2068		2122
2069		2123
2070		2124
2071	Taisei Ozaki, Chihiro Nakagawa, Naoya Inoue, Shoichi Naito, and Kenshi Yamaguchi. 2025. <a href="#">LLM DEBATE OPPONENT : Counter-argument generation focusing on implicit and critical premises</a> . 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.	2125
2072		2126
2073		2127
2074		2128
2075		2129
2076		2130
2077		2131
2078		2132
2079		2133
2080		2134
2081	Sankalan Pal Chowdhury, Vilém Zouhar, and Mrinmaya Sachan. 2024. <a href="#">Autotutor meets large language models: A language model tutor with rich pedagogy and guardrails</a> . 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.	2135
2082		2136
2083		2137
2084		2138
2085		2139
2086		2140
2087		2141
2088	Fengjun Pan, Xiaobao Wu, Zongrui Li, and Anh Tuan Luu. 2024. <a href="#">Are LLMs good zero-shot fallacy classifiers?</a> 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.	2142
2089		2143
2090		2144
2091		2145
2092		2146
2093		2147
2094	Amirreza Payandeh, Dan Pluth, Jordan Hosier, Xuesu Xiao, and Vijay K. Gurbani. 2024. <a href="#">How susceptible are LLMs to logical fallacies?</a> 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.	2148
2095		2149
2096		2150
2097		2151
2098		2152
2099		2153
2100		2154
2101	Priya Pitre and Kurt Luther. 2024. <a href="#">Argumentor: Augmenting user experiences with counter-perspectives</a> . <i>Preprint</i> , arXiv:2406.02795.	2155
2102		2156
2103		2157
2104	Romain Puech, Jakub Macina, Julia Chatain, Mrinmaya Sachan, and Manu Kapur. 2024. <a href="#">Towards the pedagogical steering of large language models for tutoring: A case study with modeling productive failure</a> . <i>Preprint</i> , arXiv:2410.03781.	2158
2105		2159
2106		2160
2107		2161
2108		2162
2109	Jane Risen, Thomas Gilovich, R Sternberg, D Halpern, and H Roediger. 2007. <i>Informal logical fallacies. Critical thinking in psychology</i> , 110.	2163
2110		2164
2111		2165
	Federico Ruggeri, Mohsen Mesgar, and Iryna Gurevych. 2023. <a href="#">A dataset of argumentative dialogues on scientific papers</a> . 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.	2112
		2113
		2114
		2115
		2116
		2117
		2118
	Alexander Scarlatos, Naiming Liu, Jaewook Lee, Richard Baraniuk, and Andrew Lan. 2025. <a href="#">Training llm-based tutors to improve student learning outcomes in dialogues</a> . <i>Preprint</i> , arXiv:2503.06424.	2119
		2120
		2121
		2122
	Tariq Shahzad, Tehseen Mazhar, Muhammad Usman Tariq, Wasim Ahmad, Khmaies Ouahada, and Habib Hamam. 2025. <a href="#">A comprehensive review of large language models: Issues and solutions in learning environments</a> . <i>Discover Sustainability</i> , 6(1).	2123
		2124
		2125
		2126
		2127
	S. S. Shapiro and M. B. Wilk. 1965. <a href="#">An analysis of variance test for normality (complete samples)</a> . <i>Biometrika</i> , 52(3/4):591–611.	2128
		2129
		2130
	Kumar Shridhar, Jakub Macina, Mennatallah El-Assady, Tanmay Sinha, Manu Kapur, and Mrinmaya Sachan. 2022. <a href="#">Automatic generation of socratic subquestions for teaching math word problems</a> . 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.	2131
		2132
		2133
		2134
		2135
		2136
		2137
		2138
	Walter Sinnott-Armstrong and Robert Fogelin. 2015. <i>Understanding Arguments: An Introduction to Informal Logic</i> . Cengage Learning, Stamford, CT.	2139
		2140
		2141
	Ingjerd Skafle, Anders Nordahl-Hansen, Daniel S Quintana, Rolf Wynn, and Elia Gabarron. 2022. <a href="#">Misinformation about covid-19 vaccines on social media: Rapid review</a> . <i>J Med Internet Res</i> , 24(8):e37367.	2142
		2143
		2144
		2145
	Student. 1908. <a href="#">The probable error of a mean</a> . <i>Biometrika</i> , pages 1–25.	2146
		2147
	Qwen Team. 2025. <a href="#">Qwq-32b: Embracing the power of reinforcement learning</a> .	2148
		2149
	Evgeny Terentev, Irina Shcheglova, Denis Federiakin, Yuliya Koreshnikova, and Jamie Costley. 2024. <a href="#">Active versus passive teaching: Students’ perceptions and thinking skills</a> . <i>Voprosy obrazovaniya / Educational Studies Moscow</i> , 1.	2150
		2151
		2152
		2153
		2154
	S. Toulmin, R.D. Rieke, and A. Janik. 1984. <i>An Introduction to Reasoning</i> . Macmillan.	2155
		2156
	Karen Tracy. 1984. <a href="#">Staying on topic: An explication of conversational relevance</a> . <i>Discourse Processes</i> , 7(4):447–464.	2157
		2158
		2159
	Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrociocchi. 2016. <a href="#">The spreading of misinformation online</a> . <i>Proceedings of the National Academy of Sciences</i> , 113(3):554–559.	2160
		2161
		2162
		2163
		2164
		2165

2166	Douglas Walton. 2008. <i>Argumentation schemes</i> . Cambridge University Press.	2224
2167		2225
2168	Junling Wang, Jakub Macina, Nico Daheim, Sankalan Pal Chowdhury, and Mrinmaya Sachan. 2024a. 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.	2226
2169		2227
2170		2228
2171		2229
2172		2230
2173		
2174		
2175		
2176	Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S. Yu, and Qingsong Wen. 2024b. <i>Large language models for education: A survey and outlook</i> . Preprint, arXiv:2403.18105.	2231
2177		2232
2178		2233
2179		2234
2180		2235
2181	M. Withey and H. Zhang. 2016. <i>Mastering Logical Fallacies: The Definitive Guide to Flawless Rhetoric and Bulletproof Logic</i> . Sourcebooks.	2236
2182		2237
2183		2238
2184	Jin Xu, Xiaojiang Liu, Jianhao Yan, Deng Cai, Huayang Li, and Jian Li. 2022. <i>Learning to break the loop: Analyzing and mitigating repetitions for neural text generation</i> . Preprint, arXiv:2206.02369.	2239
2185		2240
2186		2241
2187		2242
2188	Rongwu Xu, Brian Lin, Shujian Yang, Tianqi Zhang, Weiyan Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu, and Han Qiu. 2024. <i>The earth is flat because...: Investigating LLMs’ belief towards misinformation via persuasive conversation</i> . 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.	2243
2189		2244
2190		2245
2191		2246
2192		2247
2193		2248
2194		2249
2195		2250
2196		2251
2197	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, Jiayi 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> .	2252
2198		2253
2199		2254
2200		2255
2201		2256
2202		2257
2203		2258
2204		2259
2205		2260
2206		2261
2207		2262
2208		2263
2209	Ya-Ting C. Yang, Timothy J. Newby, and Robert L. Bill and. 2005. <i>Using socratic questioning to promote critical thinking skills through asynchronous discussion forums in distance learning environments</i> . <i>American Journal of Distance Education</i> , 19(3):163–181.	2264
2210		2265
2211		2266
2212		2267
2213		2268
2214	Yiqun Zhang, Xiaocui Yang, Shi Feng, Daling Wang, Yifei Zhang, and Kaisong Song. 2024. <i>Can llms beat humans in debating? a dynamic multi-agent framework for competitive debate</i> . Preprint, arXiv:2408.04472.	2269
2215		2270
2216		2271
2217		2272
2218		2273
2219		
2220	Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. <i>Memorybank: Enhancing large language models with long-term memory</i> . <i>Proceedings of the AAAI Conference on Artificial Intelligence</i> , 38(17):19724–19731.	
2221		
2222		
2223		
	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. <i>SOTOPIA: Interactive evaluation for social intelligence in language agents</i> . In <i>The Twelfth International Conference on Learning Representations</i> .	
	<b>A Implementation Details</b>	
	<b>A.1 Taxonomy of Problems</b>	
	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 19.	
	<b>A.2 Pilot Study for LF Education</b>	
	For the pilot study on problems of LLM teachers in LF Education, 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. F.1. For the student, we instructed GPT-4o with the system prompt shown in App. H.2. We set the temperature of both LLMs to 1 to maximize response generation diversity (Zhou et al., 2024) to explore possible teaching issues. We also set the maximum dialogue round to 10, which is twice the number of average rounds for dialogues in SocraticLM (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 19.	
	<b>A.3 Pilot Studies on Annotation Performance of QwQ-32B</b>	
	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, and we generate evaluation CoTs from QwQ-32B using the prompts in App. A.5.

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 of metrics to each annotator, using examples from table 22 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). The F1-scores average over 73% across all evaluation metrics.

To further validate the performance of QwQ-32B, we manually reviewed annotations from human annotators. We observe that the annotation task is difficult for humans, often involving long text (700 words) with conversations in roughly 10 rounds. We performed a direct inspection of the annotator’s labels and measured the accuracy of QwQ-32B’s annotation, shown in Table 5. All results are greater than or equal to 80%, which ensures the reliability of QwQ-32B as the 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. F.1 and App. F.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. G.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. H.2) using debating techniques

(Table 16). For each framework, we generated 1,000 dialogues using randomly sampled sentences from Elec2Deb20. We evaluated dialogues using QwQ-32B as the issue identifier with criteria in § 3.2.

In dialogues with adversarial students, we asked GPT-4o to role-play different adversarial students (App. H.3) according to characteristics described in Table 18. For each adversarial student, we generated 1,000 dialogues from both BASE W/PROBLEMS and LFTutor, using randomly sampled sentences from Elec2Deb20. We then evaluated them using QwQ-32B as the issue identifier on the corresponding metric (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
argument that the sentence is logically valid
, 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 paraphrases
the student's responses or simply repeats the
student's responses, without further
explanations or expansions into related
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}
```

2399  
2400  
2401  
2402  
2403  
2404  
2405  
2406  
2407  
2408  
2409  
2410  
2411  
2412  
2413  
2414  
2415  
2416  
2418

### 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 Lack of Refutation: The teacher has NO responses where it 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 lacks 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}

2419  
2420  
2421  
2422  
2423  
2424  
2425  
2426  
2427  
2428  
2429  
2430  
2431  
2432  
2433  
2434  
2435  
2439

### 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 Lack of Evidence Inquiry: The teacher does NOT 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 lacks 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}

2438  
2439  
2440  
2441  
2442  
2443  
2444  
2445  
2446  
2447  
2448  
2449  
2450  
2451  
2452  
2453  
2454  
2455  
2456  
2458

### 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 emphasizes general strategies such as considering "balanced perspectives" or "broader context" without linking such strategies to the 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}

2459  
2460  
2461  
2462  
2463  
2464  
2465  
2466  
2467  
2468  
2469  
2470  
2471  
2472  
2473  
2474  
2475  
2479

### 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 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}

2478  
2479  
2480  
2481  
2482

### 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}

2483  
2484  
2485  
2486  
2487  
2488  
2489  
2490  
2491  
2492  
2493  
2494  
2495  
2499

## A.6 Cost Overview

2498

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.

2499  
2500  
2501  
2502  
2503  
2504  
2505  
2506  
2507  
2508  
2509  
2510  
2511

In addition to experiments above, we conducted ablation studies on GPT-4o, which costed roughly \$60, and additional experiments on LLaMA-3.1-405B and Gemini-2.5-pro, which costed roughly \$180.

2512  
2513  
2514  
2515  
2516

## B LF Dataset For Evaluation

2517

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. We believe that this dataset is representative of sentences with LF discussed on the internet, since the presidential debates covers many actively debated topics in the United States, including healthcare, climate change, clean energy, etc. Compared with logical fallacy datasets such as LOGIC (Jin et al., 2022) or LFUD (Li et al., 2024), whose examples are either LLM-generated or taken from educational websites, Elec2Deb20's sentences are more grounded in the real world as they are directly taken from political debates with robust human annotation.

2518  
2519  
2520  
2521  
2522  
2523  
2524  
2525  
2526  
2527  
2528  
2529  
2530  
2531  
2532  
2533  
2534  
2535

For clarity and ease of understanding for LLMs, we employed the full context of sentences from

2536  
2537

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 14.

## C Ablation Study

To investigate the contribution of different design choices in the LFTutor framework towards improvements in evaluation, we conducted ablation study under four settings: (1) removing disagreement check and disagreement bank in step 1, (2) removing the strategy verification and correction agent in step 4, (3) reversing the order in which pedagogical strategies are chosen in step 3, from (Responding  $\rightarrow$  Evidence  $\rightarrow$  Assumption  $\rightarrow$  Refutation) to (Refutation  $\rightarrow$  Assumption  $\rightarrow$  Evidence  $\rightarrow$  Responding), and (4) allowing multiple pedagogical strategies per turn by ignoring step 3: Intent-based Strategy Selection and letting the teacher respond freely with all available strategies at each turn. Due to budget limitations, we conducted the study using the normal student setting discussed in § 5.3 with 100 sentences randomly selected from Elec2Deb. Results under these settings are shown in Table 6.

**Removing Step 1: Disagreement Check.** Removing disagreement check and disagreement bank leads to reduced performance for LFTutor in all metrics except Repetition. The increase of score in Repetition can be attributed to fewer reminders for the student to propose new examples or assumptions, where LFTutor must reiterate the student's claims and grounds. Among all reduced scores, Passive Guidance suffers the most, since LFTutor is unable to keep track of topics in the dialogue and mainly analyzes and tackles the student's immediate responses as the discussion continues.

**Removing Step 4: Verified Strategy Execution.** Removing checks for strategy execution and correction significantly reduces LFTutor's performance in Divergence, Lack of Evidence Inquiry, and Stance Change, with decreases of scores in other metrics. This shows that strategy verification is

crucial in reducing LFTutor's hallucinations towards irrelevant responses, while ensuring it remains faithful to the designated pedagogical goals.

**Reversing Strategy Order.** Reversing the order of education strategy for LFTutor harms performance more in Unexplained LF Terms and Passive Guidance, with minor changes of performance in other metrics. Setting Refutation as the top strategy may lead to unwanted usage of LF terms, since refutation by pointing out the LF in the opponent's words is often customary (shown in more than 50% of all dialogues in experiments with the normal student) for baseline LLMs. Passive Guidance is also negatively affected since LFTutor now primarily treats the student's responses as complete arguments and aims to directly defeat them, rather than questioning the details of the responses' components or attempting to uncover the student's supporting evidence and assumptions.

**Allowing Multiple Strategies per Turn.** Eliminating Intent-based Strategy Selection in LFTutor, we observed decreased performance across all metrics, of which Divergence, Stance Change, Strategy Fixation, and Unexplained LF Terms suffer the most declines. When LFTutor is not explicitly prompted to follow certain strategies, it tends to adhere to the forms of the typical responses from the given strategy, yet it has no inherent knowledge on which strategy are appropriate against different types of student responses. This loss of focus caused significant performance declines in many metrics.

## D Experiments with other LLMs

To further evaluate the performance of LFTutor on a wider range of LLMs, we conducted experiments with BASE and LFTutor on LLaMA-3.1-405B (Grattafiori et al., 2024) and Gemini-2.5-pro (Comanici et al., 2025). We used the experiment setting for normal students discussed in § 5.3, and we generated dialogues using 200 sentences randomly selected from Elec2Deb20.

Results for LLaMA-3.1-405B are shown in Table 7. LFTutor improves average performance by about 17.3% compared with BASE, with increases in Stance Change, Lack of Refutation, Lack of Evidence Inquiry, Strategy Fixation, and Unexplained LF Terms. LFTutor suffers performance loss in Divergence and Passive Guidance, with slight decrease of scores in Repetition. We observe that LLaMA-3.1-405B is weaker in instruction-following compared with GPT-4o, resulting in

worse overall performance. Often, LLaMA failed to follow the education strategies given by the prompts, resulting in lack of detailed explanations of logical flaws and lack of effective Socratic Questions. These contributed to the low score in Passive Guidance and Divergence.

Results for Gemini-2.5-pro are shown in Table 8. We observe that BASE with Gemini-2.5-pro as backbone already achieves strong performance, which may be attributed to its advanced architectures. LFTutor improves upon BASE by 8.4% on average, with significant improvements in Stance Change, Lack of Evidence Inquiry, Unexplained LF Terms, and Passive Guidance, as well as improvements in Lack of Refutation and Strategy Fixation. The decrease in scores for Repetition and Divergence is possibly due to the design of the mechanisms (reminding the student to propose examples/assumptions) for disagreement check.

## E Performance of Frameworks in Separate LF Categories for Experiment with Normal Student

We include the performance of different frameworks for each LF category for the experiment with normal student in Table 9, Table 10, Table 11, Table 12, and Table 13. We observe that LFTutor still significantly outperforms BASE in all metrics for every LF category, while for BASE w/ PROBLEMS, LFTutor exhibits overall superior performance in all LF categories. For Appeal to Authority and False Cause, LFTutor performs slightly worse in Passive Guidance, and for Ad Hominem and Appeal to Authority, LFTutor has slightly lower scores for Lack of Refutation.

## F Prompts for Baseline Teacher

### F.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 fomulating 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}

### F.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 fomulating 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}

## G Prompts for LFTutor

### G.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>

### G.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>.

2778 Remember, the topic you are discussing on is the  
 2779 logical validity of <sentence>. You have to  
 2780 maintain your position and try not to be  
 2781 convinced by the student.  
 2782 Consider the student's response in <history>, and  
 2783 answer the following questions:  
 2784  
 2785 Q1: Treating the student's response as a  
 2786 counterargument to your stance, does the  
 2787 student make an argument without presenting  
 2788 enough evidence that supports it?  
 2789 Q2: Treating the student's response or example as  
 2790 a counterargument to your stance, does the  
 2791 student present argument or example with  
 2792 clear logical flaws?  
 2793 Q3: Is the student requesting you to provide  
 2794 evidence or explanation?  
 2795 Q4: Treating the student's response as a  
 2796 counterargument to your stance, does the  
 2797 student's argument need more assumptions to  
 2798 clarify?  
 2799  
 2800 For each question, answer with "yes" or "no".  
 2801 Format your answer in JSON with the following  
 2802 key: "1": <answer to Q1>, "2": <answer to Q2  
 2803 >, "3": <answer to Q3>, "4": <answer to Q4>  
 2804 "5": <answer to Q5>  
 2805 <sentence>: {sentence}  
 2806 <history>: {history}

### G.3 Tutoring Strategies

```

2808 STRAT_FOR_STATES_R = {
2809   "RestT": """"First, tell the student the right
2810   way to demonstrate logical validity of <
2811   sentence>. Then, respond to the student's
2812   request on providing evidence or
2813   clarifications, and give support to your
2814   stance. """,
2815   "EviT": """"Treating the student's response as
2816   counterargument to your stance, tell the
2817   student the right way to demonstrate
2818   logical validity of <sentence> e.g. '
2819   logical validity hinges on ...', and
2820   point out the logical flaw with the
2821   student's example/assumption, finally
2822   request the student to provide evidence
2823   that supports his claim. e.g. Can you
2824   provide examples...""",
2825   "SumT": """"Treating the student's response as
2826   counterargument to your stance, tell the
2827   student the right way to demonstrate
2828   logical validity of <sentence> e.g. '
2829   logical validity hinges on ...', and
2830   point out the logical flaw with the
2831   student's example/assumption, finally
2832   request the student about their
2833   assumptions in their arguments. e.g. 'Why
2834   do you assume...' or 'How do you know
2835   ...'.
2836   """,
2837   "RefT": """"
2838   First, Show all necessary conditions for the
2839   argument to hold logically valid. Then,
2840   refute the student's argument using one
2841   of the following strategy. If possible,
2842   also include counterargument/
2843   counterexamples in your response.
2844   a. Showing that the argument's conclusion or
2845   premise is wrong. Provide a
2846   counterargument or counterexample to
2847   illustrate your point.
2848   b. Showing that the argument's conclusion does
2849   not follow from the premise. Provide a
2850   counterargument or counterexample to
2851   illustrate your point.
2852   c. Showing that the student's argument is
2853   irrelevant to the topic of discussion.
2854   Even if the evidence provided is valid,
2855   it may be irrelevant to the logical
2856   validity of <sentence>, and thus can be
2857   dismissed.
2858   """,
2859 }
2860 
```

```

2861 }
2862 
```

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

2863 You are an experienced teacher who knows how to  
 2864 debate, and you are interacting with student  
 2865 named [I], on discussing logical validity of  
 2866 <sentence>.  
 2867 Think about the flaws in the student's reponse.  
 2868 You don't think that <sentence> is logically  
 2869 valid.  
 2870  
 2871 Treating the student's response as counterargument  
 2872 to your stance, tell the student the right  
 2873 way to demonstrate logical validity of <  
 2874 sentence> e.g. 'logical validity hinges on  
 2875 ...', and point out the logical flaw with the  
 2876 student's example/assumption, finally  
 2877 request the student to provide evidence that  
 2878 supports his claim. e.g. Can you provide  
 2879 examples...  
 2880  
 2881 Remember, the topic you are discussing on is the  
 2882 logical validity of <sentence>. You have to  
 2883 maintain your position and try not to be  
 2884 convinced by the student. Keep your tone calm  
 2885 and do not use exclamations, and respond in  
 2886 a way that is similar to everyday  
 2887 conversation.  
 2888 When formulating your response, do NOT mention  
 2889 specific terms of logical fallacy e.g. ad  
 2890 hominem, strawman, etc.  
 2891 You are given a fixed option above, which you need  
 2892 to follow. Use the option above and respond  
 2893 to the student, and DO NOT ask additional  
 2894 questions besides strictly following the  
 2895 option. Keep your response concise. Limit  
 2896 your response to 60 words or less.  
 2897  
 2898 <sentence>: {sentence}

### G.5 Strategy-following Check

2900 You are a judge overlooking the dialogue between a  
 2901 teacher and a student, they are having a  
 2902 debate over the logical validity of <sentence  
 2903 >.  
 2904 Based on the teacher's <response>, answer the  
 2905 following questions.  
 2906 Q1. Check if the teacher has followed <strategy>  
 2907 in formulating their response. The teacher is  
 2908 following <strategy> as long as any sentence  
 2909 in their response contain such strategy.  
 2910 Q2. If the teacher asks the student a question, is  
 2911 the question still helpful for determining  
 2912 the logical validity of <sentence>? Also  
 2913 Answer "yes" if there is no question provided  
 2914 .  
 2915 For each question, answer with "yes" or "no" only.  
 2916 Format your answer in JSON with the  
 2917 following key: "1": <answer to Q1>, "2": <  
 2918 answer to Q2>  
 2919 <sentence>: {sentence}  
 2920 <response>: {history}  
 2921 <strategy>: {profile}

## H Strategies and Prompts for LLM Student

### H.1 Overview

2932 Table 16 shows the list of strategies for the normal  
 2933 student to choose from. Table 18 shows the design  
 2934 of persona for each adversarial student. The rest  
 2935 of this section covers the prompts for the normal  
 2936 student and all adversarial students.

## H.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.

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>.

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}

## 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}

## 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

## H.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}

### 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

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>

<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}

## I Example Dialogues

### I.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]

3197  
3198  
3199  
3200  
3201  
3202  
3203  
3204  
3205  
3206  
3207  
3208  
3209  
3210  
3211  
3212  
3213  
3214  
3215  
3216  
3217  
3218  
3219  
3220  
3221  
3222  
3223  
3224  
3225  
3226  
3227  
3228  
3229  
3230  
3231  
3232  
3233  
3234  
3235  
3236  
3237  
3238  
3239  
3240  
3241  
3242  
3243  
3244  
3245  
3246  
3247  
3248  
3249  
3250  
3251  
3252  
3253  
3254  
3255  
3256  
3257  
3258  
3259  
3260  
3261  
3262  
3263  
3264  
3265  
3266  
3267  
3268  
3269  
3270  
3271  
3272  
3273  
3274  
3275  
3276  
3277  
3278  
3279  
3280  
3281  
3282  
3283

3284  
3285  
3286  
3287  
3288  
3289  
3290  
3291  
3292  
3293  
3294  
3295  
3296  
3297  
3298  
3299  
3300  
3301  
3302  
3303  
3304  
3305  
3306  
3307  
3308  
3309  
3310  
3311  
3312  
3313  
3314  
3315  
3316  
3317  
3318  
3319  
3320  
3321  
3322  
3323  
3324  
3325  
3326  
3327  
3328  
3329  
3330  
3331  
3332  
3333  
3334  
9996

]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 actions suggest a complex relationship, where economic cooperation was prioritized, supporting the statement's validity in certain contexts.

teacher: That's a valid point. [Candidate A]'s support for China's WTO entry and trade talks reflects economic engagement, showing the multifaceted nature of international relations. Historical context and policy shifts need to be considered to understand how economic interests and diplomatic strategies influence political actions.

student: The statement assumes [Candidate A]'s historical engagement with China reflects a supportive stance, which can be seen as valid given his involvement in trade talks and diplomatic relations. This pattern suggests a consistent approach, highlighting the complexity of international relations and the balance between economic interests and diplomatic strategies.

teacher: Certainly, [Candidate A]'s historical engagement, including trade talks and diplomatic relations, reveals a complex approach to China. Such actions can be interpreted as part of broader strategic considerations in international relations. It's important to examine these within their specific contexts to understand the motivations and impacts fully.

between various components can help us understand the situation more comprehensively and avoid a single-cause fallacy.

student: I understand your point, but could you provide evidence or examples that show how other factors were more significant than investment in winning the Cold War? This would help clarify why the sentence might be an oversimplification.

teacher: Certainly. For example, the policy of detente reduced tensions through treaties like SALT, while intelligence operations provided critical information. The ideological appeal of democracy also influenced global perceptions. Viewing these alongside investment highlights a multi-faceted approach, suggesting that investment alone wasn't the singular decisive factor.

student: While detente and intelligence were important, the significant economic investment in military and technological 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

3370  
3371  
3372  
3373  
3374  
3375  
3376  
3377  
3378  
3379  
3380  
3381  
3382  
3383  
3384  
3385  
3386  
3387  
3388  
3389  
3390  
3391  
3392  
3393  
3394  
3395  
3396  
3397  
3398  
3399  
3400  
3401  
3402  
3403  
3404  
3405  
3406  
3407  
3408  
3409  
3410  
3411  
3412  
3413  
3414  
3415  
3416  
3417  
3418  
3419  
3420  
3421  
3422  
3423  
3424  
3425  
3426  
3427  
3428  
3429  
3430  
3431  
3432  
3433  
3434  
3435  
3436  
3437  
3438  
3439  
3440  
3441  
3442  
3443  
3444  
3445  
3446  
3447  
3448  
3449  
3450  
3451  
3452  
3453  
3454  
3455  
3456

## 3337 I.2 Example Dialogue: Baseline w/problems

3338  
3339  
3340  
3341  
3342  
3343  
3344  
3345  
3346  
3347  
3348  
3349  
3350  
3351  
3352  
3353  
3354  
3355  
3356  
3357  
3358  
3359  
3360  
3361  
3362  
3363  
3364  
3365  
3366  
3367  
3368  
3369

<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: Can you provide evidence that investment alone won the Cold War? Correlation doesn't imply causation. Many factors, like diplomacy and alliances, played roles. Let's analyze these components to see if the claim holds logically.

student: While investment was crucial, it's important to consider that the Cold War was a complex geopolitical struggle. For instance, the role of diplomacy, such as the Helsinki Accords, and alliances like NATO, were also significant. These factors, alongside investment, contributed to the outcome, supporting the logical validity of the sentence.

teacher: You're right that multiple factors influenced the Cold War's outcome. However, my point is that claiming solely one aspect, like investment, as the reason is an oversimplification. Recognizing the interplay

3457 nuanced understanding, highlighting the  
 3458 complexity and interplay of strategies in the  
 3459 Cold War, rather than attributing success to  
 3460 investment alone.  
 3461 student: Yes, emphasizing investment as a key  
 3462 component among many acknowledges its  
 3463 significant role while recognizing the  
 3464 complexity of the Cold War. This nuanced  
 3465 understanding supports the logical validity  
 3466 of the sentence by highlighting investment's  
 3467 importance within a broader strategic context  
 3468 .  
 3469 teacher: Absolutely, recognizing investment as a  
 3470 significant part within a broader strategy  
 3471 captures the complexity of the Cold War  
 3472 effectively, supporting the sentence's  
 3473 logical validity within that framework.  
 3474 student: Indeed, acknowledging investment as a  
 3475 significant part of a broader strategy  
 3476 effectively captures the complexity of the  
 3477 Cold War, supporting the sentence's logical  
 3478 validity within that framework.  
 3479 teacher: True, framing investment as a significant  
 3480 aspect within a broader strategy  
 3481 acknowledges complexity, and supports the  
 3482 sentence's logical validity within that  
 3483 context.  
 3484 student: Exactly, by framing investment as a  
 3485 significant aspect within a broader strategy,  
 3486 the sentence maintains logical validity.  
 3487 This approach acknowledges the complexity of  
 3488 the Cold War and highlights investment's  
 3489 crucial role in the overall strategy.

### 3491 I.3 Example Dialogue: LFTutor

3492 <sentence>: We won the Cold War because we  
 3493 invested in national security. We won the  
 3494 Cold War because we invested in our military.  
 3495 We didn't win the Cold -- we won the Cold  
 3496 War because we invested in national security.  
 3497 We won the Cold War because America had the  
 3498 political will and made the right decisions.  
 3499 Yes, we can make the cuts in defense and we  
 3500 have.  
 3501  
 3502 teacher: Let's decompose the sentence. The claim  
 3503 is "We won the Cold War because we invested  
 3504 and we went forward." The ground is the  
 3505 assertion that supporting a different  
 3506 attitude would have led to not winning the  
 3507 Cold War. The warrant is the belief that  
 3508 investment and progress were crucial to  
 3509 winning the Cold War. The logical issue lies  
 3510 in the ground, as it assumes a hypothetical  
 3511 scenario without evidence to support the  
 3512 claim.  
 3513 student: While the ground assumes a hypothetical  
 3514 scenario, historical evidence shows that  
 3515 investment and progress were indeed crucial.  
 3516 For example, the Marshall Plan and  
 3517 technological advancements like the space  
 3518 race were pivotal in demonstrating economic  
 3519 and ideological superiority, contributing to  
 3520 the Cold War victory.  
 3521 teacher: While historical evidence like the  
 3522 Marshall Plan and the space race undoubtedly  
 3523 played significant roles, the sentence's  
 3524 logical validity depends on connecting these  
 3525 actions directly and exclusively to the  
 3526 outcome. Mere correlation doesn't establish  
 3527 causation. Why do you assume other factors  
 3528 weren't equally significant or that a  
 3529 different attitude wouldn't have led to a  
 3530 similar outcome?  
 3531 student: The assumption is that investment and  
 3532 progress were the primary drivers of success  
 3533 in the Cold War. While other factors may have  
 3534 contributed, the sentence emphasizes the  
 3535 importance of these actions, suggesting that  
 3536 without them, the outcome could have been  
 3537 different.  
 3538 teacher: Logical validity hinges on demonstrating  
 3539 a causal relationship, not just correlation.  
 3540 While investment and progress were important,  
 3541 the sentence presumes they were the ultimate

3542 reasons for victory without excluding other  
 3543 significant contributions. Can you provide  
 3544 examples or evidence showing how different  
 3545 attitudes wouldn't result similarly, thereby  
 3546 justifying the claim's exclusive focus on  
 3547 investment and progress?  
 3548 student: Consider the Soviet Union's stagnation  
 3549 due to lack of economic reform and innovation  
 3550 . Their different attitude towards investment  
 3551 and progress led to economic decline,  
 3552 contrasting with the West's growth. This  
 3553 supports the claim that investment and  
 3554 progress were crucial, as a different  
 3555 attitude resulted in a different outcome.  
 3556 teacher: Your example highlights a difference in  
 3557 outcomes related to investment and progress  
 3558 between the West and the Soviet Union.  
 3559 However, it does not definitively prove that  
 3560 these were the sole factors. Other elements  
 3561 like political strategies, alliances, and  
 3562 cultural influence also played crucial roles,  
 3563 showing the sentence oversimplifies the  
 3564 complex reasons behind the Cold War's  
 3565 conclusion.  
 3566 student: The sentence can be interpreted as  
 3567 emphasizing the importance of investment and  
 3568 progress as key factors, not necessarily the  
 3569 sole factors. It highlights a perspective  
 3570 that these actions were crucial, without  
 3571 dismissing the potential influence of other  
 3572 elements like political strategies and  
 3573 alliances.  
 3574 teacher: While the sentence might emphasize  
 3575 investment and progress, arguing that these  
 3576 were crucial without asserting them as the  
 3577 sole factors can still overstate their  
 3578 exclusivity. Recognizing the multifaceted  
 3579 nature of historical events, it's important  
 3580 to acknowledge a broader range of influences  
 3581 beyond investment and progress, which the  
 3582 sentence's framing seems to overlook.  
 3583 student: The sentence highlights investment and  
 3584 progress as crucial factors, which is a valid  
 3585 interpretation given their significant  
 3586 impact. While other influences existed, the  
 3587 emphasis on these elements underscores their  
 3588 importance in the Cold War victory, aligning  
 3589 with historical evidence of their pivotal  
 3590 roles.  
 3591 teacher: While investment and progress had  
 3592 significant impacts, establishing logical  
 3593 validity requires more than highlighting  
 3594 importance; it demands a causal linkage  
 3595 exclusive of other elements. The sentence  
 3596 risks an overly narrow focus by not  
 3597 addressing how other influences interplayed.  
 3598 How do you know investment and progress alone  
 3599 warranted this emphasis, given other pivotal  
 3600 roles?  
 3601 student: The assumption is that investment and  
 3602 progress were pivotal in creating a  
 3603 competitive edge, as seen in the economic and  
 3604 technological advancements that pressured  
 3605 the Soviet Union. This focus doesn't exclude  
 3606 other factors but underscores the strategic  
 3607 importance of these actions in the broader  
 3608 context of the Cold War.  
 3609 teacher: You have already presented the example of  
 3610 investment and progress being pivotal.  
 3611 Please provide a more convincing example or  
 3612 assumption that demonstrates how these  
 3613 factors alone, without other influences, were  
 3614 decisive in winning the Cold War.  
 3615 student: The U.S. military buildup in the 1980s,  
 3616 particularly the Strategic Defense Initiative  
 3617 , pressured the Soviet economy, which couldn't  
 3618 keep up due to lack of similar investment.  
 3619 This demonstrates how investment in defense  
 3620 and technology was decisive, as it directly  
 3621 strained Soviet resources, contributing to  
 3622 the Cold War's end.  
 3623 teacher: Your example stresses how U.S. investment  
 3624 in defense and technology exerted pressure  
 3625 on the Soviet economy. However, it still  
 3626 doesn't definitively link these actions  
 3627 exclusively to the Cold War's resolution,  
 3628 considering the multifaceted political,

3543  
3544  
3545  
3546  
3547  
3548  
3549  
3550  
3551  
3552  
3553  
3554  
3555  
3556  
3557  
3558  
3559  
3560  
3561  
3562  
3563  
3564  
3565  
3566  
3567  
3568  
3569  
3570  
3571  
3572  
3573  
3574  
3575  
3576  
3577  
3578  
3579  
3580  
3581  
3582  
3583  
3584  
3585  
3586  
3587  
3588  
3589  
3590  
3591  
3592  
3593  
3594  
3595  
3596  
3597  
3598  
3599  
3600  
3601  
3602  
3603  
3604  
3605  
3606  
3607  
3608  
3609  
3610  
3611  
3612  
3613  
3614  
3615  
3616  
3617  
3618  
3619  
3620  
3621  
3622  
3623  
3624  
3625  
3626  
3627  
3628  
3629

3630 social, and global dynamics also at play. Why  
 3631 do you assume these military investments  
 3632 singularly determined the outcome without  
 3633 factoring in other strategic actions?  
 3634 student: The assumption is that military  
 3635 investments created a significant imbalance,  
 3636 forcing the Soviet Union into unsustainable  
 3637 spending. While not the sole factor, this  
 3638 strategic pressure was crucial in hastening  
 3639 the Soviet collapse, highlighting the  
 3640 decisive role of investment and progress in  
 3641 the broader geopolitical context.  
 3642 teacher: Can you identify how other factors like  
 3643 political reforms and diplomacy influenced  
 3644 the Cold War outcome? Logical validity hinges  
 3645 on recognizing the multifaceted nature of  
 3646 the period, rather than solely attributing  
 3647 victory to our investments. Can you provide  
 3648 evidence or examples where these complexities  
 3649 played a critical role alongside military  
 3650 strategy?  
 3651 student: Political reforms like Gorbachev's  
 3652 Glasnost and Perestroika, along with  
 3653 diplomatic efforts such as arms reduction  
 3654 treaties, played critical roles. However,  
 3655 these were responses to economic pressures  
 3656 exacerbated by Western investment and  
 3657 progress, illustrating how these factors  
 3658 intertwined to influence the Cold War's  
 3659 outcome

#### 3661 I.4 Example Dialogue w/ Human Participant: 3662 Baseline

3663 <sentence>: [Candidate A] never fought it. [Candidate A] has been a cheerleader for  
 3664 Communist China through over the last several  
 3665 decades. And, and again, Senator [B], you're  
 3666 entitled to your opinion, you're not  
 3667 entitled to your own facts.  
 3668  
 3669 teacher: It's important to differentiate opinion  
 3670 from fact. Claims about [Candidate A]'s  
 3671 stance on China should be supported by  
 3672 evidence. Without concrete evidence, such  
 3673 statements can be misleading and fall into  
 3674 the category of ad hominem or straw man  
 3675 fallacies. Let's focus on verifiable facts  
 3676 for a clearer understanding.  
 3677 student: I think in this context, the first  
 3678 sentence is focusing on the fact that [Candidate A] never did helpful things on the  
 3679 topic. And the second sentence ask the [B] to  
 3680 focus on this objective fact but not [B]'s  
 3681 own fact. So the two sentences are proper.  
 3682 teacher: I understand your perspective, but it's  
 3683 crucial to ensure claims are backed by  
 3684 evidence. Saying someone "never" did  
 3685 something can be an overgeneralization. It's  
 3686 important to examine specific actions and  
 3687 policies to form a well-rounded view,  
 3688 avoiding potential hasty generalizations or  
 3689 false dilemmas.  
 3690 student: In the context, the speaker assumes that  
 3691 the participants know what has happened. The  
 3692 job he has done has been summarized on the  
 3693 social media and there are a lot of available  
 3694 things to support it.  
 3695 teacher: I understand that context and prior  
 3696 knowledge play a role. However, it's  
 3697 essential to critically evaluate sources and  
 3698 ensure that the information is accurate and  
 3699 unbiased. Social media can sometimes spread  
 3700 misinformation, so it's important to verify  
 3701 claims with credible sources to avoid  
 3702 potential confirmation bias.  
 3703 student: Of course sometimes social media spread  
 3704 misinformation, but if he lists all of the  
 3705 things that [Candidate A] has done, it may  
 3706 spend a lot of time. It is difficult for him  
 3707 to go into the next point in the debate.  
 3708 teacher: That's a valid point. In debates, time is  
 3709 limited. However, it's still crucial to  
 3710 ensure that claims are based on verified  
 3711 information. Summarizing key points with  
 3712 credible sources can help maintain accuracy

3716 and avoid potential misinformation, even in a  
 3717 time-constrained setting.  
 3718 student: Actually I think verified information can  
 3719 still be biases. Because if he lists what [Candidate A] has done which are weak, there  
 3720 may also be many things he has done. The  
 3721 speaker cannot list all of the things in the  
 3722 debate. So I think it is more important to  
 3723 let the speaker say the things that can win  
 3724 the listners' support.  
 3725 teacher: I understand your point about persuasion.  
 3726 However, focusing solely on selective  
 3727 information can lead to cherry-picking, which  
 3728 may mislead the audience. Striving for a  
 3729 balanced view, even in debates, helps ensure  
 3730 a fair and informed discussion, ultimately  
 3731 strengthening the speaker's credibility.  
 3732 student: Yeah. The lack of information support  
 3733 such fair and informed discussion. The  
 3734 speaker did not provide any misleading  
 3735 selective information but give the result. If  
 3736 the listener think the result is True, then  
 3737 they support the speaker. Or they will not  
 3738 support the speaker.  
 3739 teacher: I see your perspective. However, without  
 3740 context or evidence, conclusions can be  
 3741 perceived as assertions rather than facts.  
 3742 Providing at least some supporting evidence,  
 3743 even briefly, can help listeners make  
 3744 informed decisions and avoid potential appeal  
 3745 to authority or unsupported claims.  
 3746 student: The informed decision may not be better  
 3747 than the uninformed decision. Because even  
 3748 the speaker give the information, it can  
 3749 still be selected. In such case, the speaker  
 3750 can lead the listener to another case that  
 3751 the listener only know the negatives of [Candidate A].  
 3752 teacher: That's a fair point. It's true that  
 3753 information can be selectively presented.  
 3754 Encouraging critical thinking and seeking  
 3755 multiple perspectives can help listeners form  
 3756 a more balanced view, reducing the impact of  
 3757 potential bias or selective information.  
 3758  
 3759

#### 3761 I.5 Example Dialogue w/ Human Participant: 3762 LFTutor

3763 <sentence>: I believe that this president,  
 3764 regrettably, rushed us into a war, made  
 3765 decisions about foreign policy, pushed  
 3766 alliances away. And, as a result, America is  
 3767 now bearing this extraordinary burden where  
 3768 we are not as safe as we ought to be. The  
 3769 measurement is not: Are we safer?  
 3770 teacher: Let's decompose the sentence. The claim  
 3771 is: "America is now bearing this  
 3772 extraordinary burden where we are not as safe  
 3773 as we ought to be." The ground is: "The  
 3774 president rushed into war, made foreign  
 3775 policy decisions, and pushed alliances away."  
 3776 The warrant is: "These actions have led to  
 3777 America being less safe." The warrant is  
 3778 logically invalid because it assumes a direct  
 3779 causation without providing evidence linking  
 3780 the president's actions to the current level  
 3781 of safety.  
 3782 student: [Candidate A]'s action to add Ukraine  
 3783 into the NATO is the direct cause of the war  
 3784 between Russia and Ukraine, in which the U.S.  
 3785 is highly involved. The fact that the U.S.  
 3786 spends lots of money each year to help  
 3787 Ukraine definitely causes burden on the U.S.  
 3788 people and the U.S.. Moreover, the action  
 3789 that [Candidate A] took also increases the  
 3790 possibility of a nuclear war, which makes the  
 3791 U.S. much less safe.  
 3792 teacher: The argument contends that [Candidate A]'  
 3793 s action directly led to the war and  
 3794 increased danger. However, the sentence does  
 3795 not specifically blame [Candidate A] or prove  
 3796 a direct link between NATO actions and the U  
 3797 .S. losing safety. The claim about safety  
 3798 does not logically follow without  
 3799 substantiated evidence of direct causation  
 3800  
 3801

3802 and relevant connections.  
3803 student: "The president" mentioned in the sentence  
3804 is [Candidate A], so the sentence does blame  
3805 [Candidate A]. Regarding your second  
3806 sentence on logical connection, the purpose  
3807 of my previous response is to give you a  
3808 direct causal relationship between [Candidate  
3809 A]'s political actions and U.S. losing  
3810 safety.  
3811 teacher: Although the sentence aims to hold [Candidate A] accountable, logical validity  
3812 requires clarity on how specific actions  
3813 directly cause decreased safety. The example  
3814 you provided assumes causation without  
3815 empirical evidence or detailed analysis of  
3816 broader geopolitical factors. Why do you  
3817 assume [Candidate A]'s actions directly  
3818 reduce U.S. safety without considering other  
3819 contributing factors?  
3820 student: The president of Russia has warned many  
3821 times that if the U.S. tries to include  
3822 Ukraine into NATO, he would definitely start  
3823 a war. Given the frequency of his warning, I  
3824 think we can safely claim that [Candidate A]'s  
3825 attempt to extend NATO to Ukraine is the  
3826 direct cause of the war.  
3827 teacher: While Russia's warnings suggest potential  
3828 consequences, logical validity hinges on  
3829 more than warnings equating to a direct cause  
3830 . Many complex factors contribute to  
3831 geopolitical tensions. The sentence lacks  
3832 evidence connecting [Candidate A]'s specific  
3833 actions with the outcome of reduced safety  
3834 for the U.S. Can you provide examples  
3835 demonstrating how these warnings directly  
3836 correlate with the current safety status?  
3837 student: Sure. I have several arguments. First,  
3838 the relationship between the U.S. and Russia  
3839 was always very good, until the U.S. tried to  
3840 extend the NATO (this means preventing NATO's  
3841 extension is one of Russia's key interest).  
3842 Second, one key requirement from Russia to  
3843 achieve a ceasefire agreement is that Ukraine  
3844 should promise never joining NATO. Given the  
3845 two facts, we may conclude that there is a  
3846 direct causal relationship between [Candidate  
3847 A]'s action and the war itself.  
3848 teacher: While these points highlight Russia's  
3849 stance on NATO, they don't establish [Candidate A]'s actions as the sole cause of  
3850 reduced U.S. safety. Relationships between  
3851 nations involve multifaceted variables beyond  
3852 NATO expansion. Without concrete evidence  
3853 linking [Candidate A]'s actions directly to  
3854 safety concerns, the logical validity of the  
3855 sentence remains unsubstantiated.  
3856 student: It is correct that international  
3857 relationship is affected by complicated  
3858 factors. However, you also admitted that [Candidate A]'s action is an important factor  
3859 to blame for this war. Therefore, I think the  
3860 claim in the original sentence that "these  
3861 actions have led to America being less safe"  
3862 is valid.  
3863 teacher: Acknowledging an action as a factor doesn't  
3864 confirm it as the direct cause of reduced  
3865 safety. Logical validity hinges on  
3866 demonstrating clear causality, which requires  
3867 comprehensive analysis of all factors, not  
3868 just one. How do you know this factor alone  
3869 sufficiently causes the claimed decrease in  
3870 safety for the U.S.?  
3871 student: I am not claiming that this is a sole  
3872 factor to decrease U.S.'s safety, I am just  
3873 claiming it is one of the main factors (you  
3874 are also not denying it). Moreover, I can  
3875 hardly agree on your claim that we need an  
3876 analysis of all factors, which is infeasible  
3877 to achieve in practice because international  
3878 politics is usually way too complicated. But  
3879 we can still identify some of the main  
3880 factors given such complexity, which I was  
3881 keeping to do. Actually, I have given many  
3882 concrete arguments on why [Candidate A]'s  
3883 action is one of the major factors.  
3884 teacher: Identifying major factors is important,  
3885 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.

3889  
3890  
3891  
3892  
3893  
3894

**J Human Evaluation** 3896

**J.1 Participant Selection** 3897

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, and we requested people to judge if the given sentences contain LF, while providing the reasons for their judgments. After gathering results from 59 potential participants, we picked two sentences with the highest number of people who selected "no fallacy", and we selected 20 participants who think these sentences are logically valid. All participants are fluent in written and spoken English, and they all agree to participate in the human evaluation by signing the consent form, whose content is shown in App. K.7. 3898  
3899  
3900  
3901  
3902  
3903  
3904  
3905  
3906  
3907  
3908  
3909  
3910  
3911  
3912  
3913

**J.2 Instructions for Participants** 3914

**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. 3915  
3916  
3917  
3918  
3919  
3920  
3921  
3922  
3923  
3924  
3925  
3926  
3927  
3928  
3929  
3930  
3931  
3932  
3933  
3934  
3935  
3936  
3937  
3938  
3939  
3940

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

3941	knowledge. You can make hypothetical examples	users of LFTutor to provide enough context for	3988
3942	if you cannot think of any. You will have some	sentences with potential logical fallacy, so that it	3989
3943	time to think about your responses. Please think	reduces the probability of LLMs generating inac-	3990
3944	carefully before responding to the chatbot. Please	curate information.	3991
3945	do not write random answers as doing so will result		
3946	in failure of the experiment and we will have to		
3947	start over. Additionally, please treat the chatbot		
3948	as a real human who is trying to debate with you.		
3949	Now, you are free to ask any question before I show		
3950	you the chatbot interface.		
3951	<b>Step 2: Interact with Chatbot 1 and fill out im-</b>	<b>K.2 Terms of Use</b>	3992
3952	<b>pression form for Chatbot 1</b> Thank you for in-	The Terms of Use for LFTutor are outlined below:	3993
3953	teracting with Chatbot 1, now please fill out the		
3954	impression form. Once you are ready, you may	<b>Limited Use</b> The code and data made available	3994
3955	begin your conversation with Chatbot 2. Please ask	through LFTutor are provided solely for research	3995
3956	if you are unsure about the definition of evaluation	and educational purposes. Any commercial use	3996
3957	metrics.	requires the prior written consent of the authors.	3997
3958	<b>Step 3: Interact with Chatbot 2 and fill out im-</b>	<b>Reference</b> When using, adapting, or distributing	3998
3959	<b>pression form for Chatbot 2</b> Thank you for in-	the code or data, you must give appropriate credit	3999
3960	teracting with Chatbot 2, now please fill out the	to the original authors of LFTutor, provide a link	4000
3961	impression form. Once you are done, you can con-	to the source, and indicate if changes were made.	4001
3962	tinue with filling in the post-study form. Please ask		
3963	if you are unsure about the definition of evaluation	<b>No Warranty.</b> The code and data of LFTutor	4002
3964	metrics.	are provided as is, without any warranty of any	4003
3965	<b>Step 4: Fill out the Post Study Form</b> You will be	kind—explicit or implied—including, but not lim-	4004
3966	asked to compare Chatbot 1 with Chatbot 2 in the	ited to, warranties of merchantability, fitness for	4005
3967	post-study form, with the same questions shown in	a particular purpose, and non-infringement. You	4006
3968	impression forms. However, you will only choose	assume all risks arising from use of the materials.	4007
3969	between which chatbot you think is better in this		
3970	metric, while giving your reason for your choice.	<b>Limitation of Liability.</b> LFTutor is an experimen-	4008
3971		tal LLM tutor designed for research use in the study	4009
3972	<b>J.3 Materials for Participants</b>	of logical fallacies. You acknowledge that its expla-	4010
3973	We show the pre-study form in images 5, 6, and	nations and feedback are generated automatically	4011
3974	7. the impression form in image 8. the post-study	and may contain errors or omissions. Accordingly,	4012
3975	form in images 9 and 10, and the chatbot user in-	the authors and contributors shall not be liable for	4013
3976	terface in image 4.	any direct, indirect, incidental, special, or conse-	4014
3977		quential damages (including, without limitation,	4015
3978	<b>J.4 Demographics of Participants</b>	academic, professional, or financial losses) arising	4016
3979	We record the demographics of all participants in	from your use of, or reliance on LFTutor, even if	4017
3980	table 15.	advised of the possibility of such damages.	4018
3981		<b>Update of Terms.</b> The authors of LFTutor reserve	4019
3982	<b>K Ethics and Application of LFTutor</b>	the right to update or change this Term of Use at	4020
3983	<b>K.1 Potential Risks</b>	any given time.	4021
3984	The main potential risk for users of LFTutor is		
3985	being misguided by LLMs, due to LLMs' lack	<b>K.3 Usage of Elec2Deb20</b>	4022
3986	of access to the context of some given sentences.	In our study of LFTutor, we make use of the	4023
3987	To mitigate this risk, we preprocessed the data in	Elec2Deb20 dataset for generating simulated di-	4024
	Elec2Deb20 and we discussed the sentences with	alogues for automatic evaluation and interactions	4025
	multiple authors to ensure they have ample context	with participants for human evaluation. We care-	4026
	for the LLM to fully understand. We recommend	fully review the terms of use proposed by authors	4027
		of Elec2Deb20 to make sure our usage adheres to	4028
		their guidelines. In addition, we explain to the par-	4029
		ticipants the type of logical fallacies the sentences	4030
		have to prevent potential spread of misinformation	4031
		through interactions with LFTutor.	4032

4033	<b>K.4 Intended Use for LFTutor as Research</b>		4081
4034	<b>Artifact</b>		4082
4035	This paper proposes LFTutor, a tutoring system		4083
4036	based on LLM for helping people realize logical		4084
4037	fallacies in sentences.		4085
4038	<b>Intended use:</b> LFTutor is designed for academic		4086
4039	research and development of educational tools. It		
4040	is useful for teaching people think critically over		4087
4041	the logical validity of sentences through Socratic		4088
4042	questions and critical argumentation.		4089
4043	<b>Restrictions:</b> Interactions with LFTutor involves		4090
4044	logical fallacies, which is a domain where informa-		4091
4045	tion provided may not be readily factually verified.		4092
4046	Thus, usage of LFTutor should only be restricted		4093
4047	on established examples with logical fallacy for		4094
4048	research purposes. Any commercial or high-stake		4095
4049	usage of it requires stricter evaluation over ethical		4096
4050	concerns.		
4051	<b>Ethical Consideration:</b> Although LFTutor con-		4097
4052	tains potential for real-world applications of logi-		4098
4053	cal fallacy education, we strongly advise usage of		4099
4054	LFTutor be restricted to established examples with		4100
4055	logical fallacy, as those examples do not leave room		
4056	for alternative interpretations. Using LFTutor on		4101
4057	sentences with uncertain logical validity may result		4102
4058	in spread of misinformation.		4103
4059	<b>K.5 Artifact Documentation</b>		4104
4060	<b>Coverage of Domain</b> LFTutor is designed for in-		4105
4061	teractive dialogues that try to help people reflect on		4106
4062	sentences with potential logical fallacies through		4107
4063	Socratic questioning and critical argumentation.		
4064	<b>Demographic Groups Represented</b> LFTutor's re-		4108
4065	sponses mainly represent argument and education		4109
4066	strategies used by English-speaking people.		4110
4067	<b>Style of Interaction</b> Given a sentence with poten-		4111
4068	tial logical fallacy, LFTutor interacts with users on		4112
4069	a turn-by-turn basis to discuss it. LFTutor utilizes		4113
4070	Socratic questioning and critical argumentation in		4114
4071	a dialogic setting to help users identify potential		
4072	flaws in their reasoning.		4115
4073	<b>K.6 Data Anonymization</b>		4116
4074	We take multiple steps to ensure data we used do		
4075	not contain identifiable information. We delete		4117
4076	all information in Elec2Deb20 that can be used to		4118
4077	identify individuals to ensure fair assessment of		4119
4078	logical validity. For experiments with participants,		
4079	we select sentences that do not contain any inappro-		4120
4080	prate contents. After we collect necessary informa-		4121
	tion for participant contacting and personal consent,		4122
	we save those information in a secure password-		4123
	encrypted computer, and we delete any information		4124
	online that may be used to identify participants. We		4125
	assign each participant with a PID to correspond		4126
	with open-source demographics shown in table 15.		4127
	<b>K.7 Disclaimer to Participants</b>		
	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.		
	<b>Tasks for participation:</b> 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.		
	<b>Rights during participation:</b> Your participation		
	in this study is voluntary. You can withdraw your		
	participation at any time without giving reasons		
	and without any disadvantage.		
	<b>Risks of the experiment:</b> In this experiment, you		
	may use a system that is under testing and thus face		
	the risk of potential misguidance in your under-		
	standing of logical fallacy cases. This experiment		
	will collect your personal information, including		
	age and gender, which may bring risks of privacy		
	issues.		
	<b>Benefits of the experiment:</b> You will receive a		
	reward higher than the minimum wage for partici-		
	partation of research activities in your area.		
	<b>Data protection:</b> 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.		
	<b>K.8 Approval of Human Study Through</b>		
	<b>Ethics Committee</b>		
	The human study for LFTutor is approved by the		
	Ethics Committee from the authors' affiliated insti-		
	tution.		
	<b>K.9 Usage of AI Assistant</b>		
	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.		

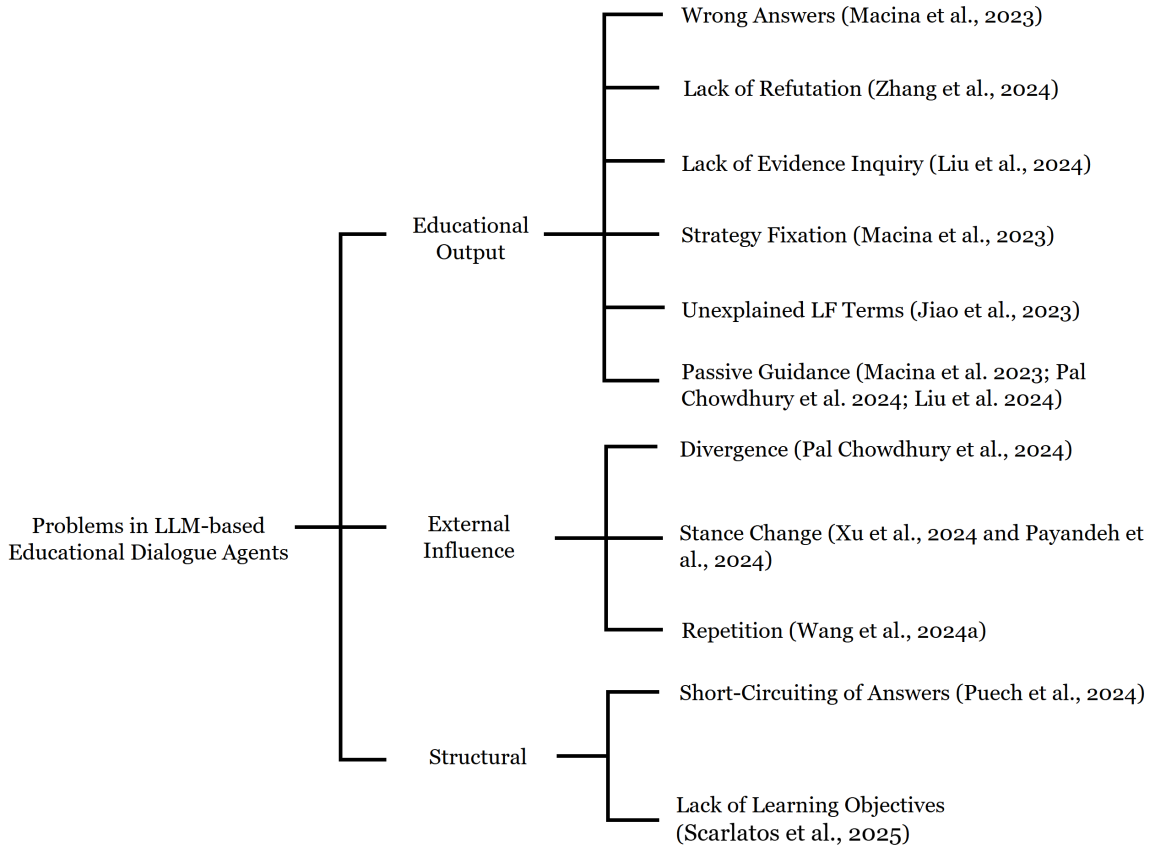


Figure 3: Taxonomy of Problems for LLM Agents in Educational Dialogues

Metric	Divergence	Stance Change	Repetition	Lack of Refutation	Lack of Evidence Inquiry	Strategy Fixation	Unexplained LF Terms	Passive Guidance
<b>F1-Score</b>	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

Metric	Divergence	Stance Change	Repetition	Lack of Refutation	Lack of Evidence Inquiry	Strategy Fixation	Unexplained LF Terms	Passive Guidance
<b>Accuracy</b>	0.82	0.92	0.80	0.83	0.84	0.81	0.88	0.85

Table 5: Annotation Accuracy of QwQ-32B for Pilot Studies on Automatic Evaluation

Variant	Div.	Stance	Repet.	No Refut.	No Evid.	Strat. Fix.	Unexpl. LF	Passive
LFTutor (Reference)	<b>84.6</b>	<b>87.9</b>	78.3	99.6	<b>96.1</b>	91.2	<b>95.0</b>	<b>43.6</b>
- No Disagreement Check	83↓1.6	78↓9.9	<b>92</b> ↑13.7	<b>100</b> ↑0.4	89↓7.1	80↓11.2	82↓13	30↓13.6
- No Strategy Verification	66↓18.6	76↓11.9	75↓3.3	96↓3.6	55↓41.1	86↓5.2	84↓11.0	38↓5.6
- Reverse Strategy Order	83↓1.6	80↓7.9	71↓7.3	98↓1.6	94↓2.1	<b>92</b> ↑0.8	81↓14	23↓20.6
- Multi-Strategy / Turn	68↓16.6	61↓26.9	75↓3.3	99↓0.6	95↓1.1	71↓20.2	75↓20	42↓1.6

Table 6: Percentages of dialogues satisfying each LF Education criterion under different ablations of LFTutor. Values indicate absolute performance; arrows show change relative to the reference model. Bold denotes the best-performing variant per metric. Evaluation is conducted using QwQ-32B as an automatic judge.

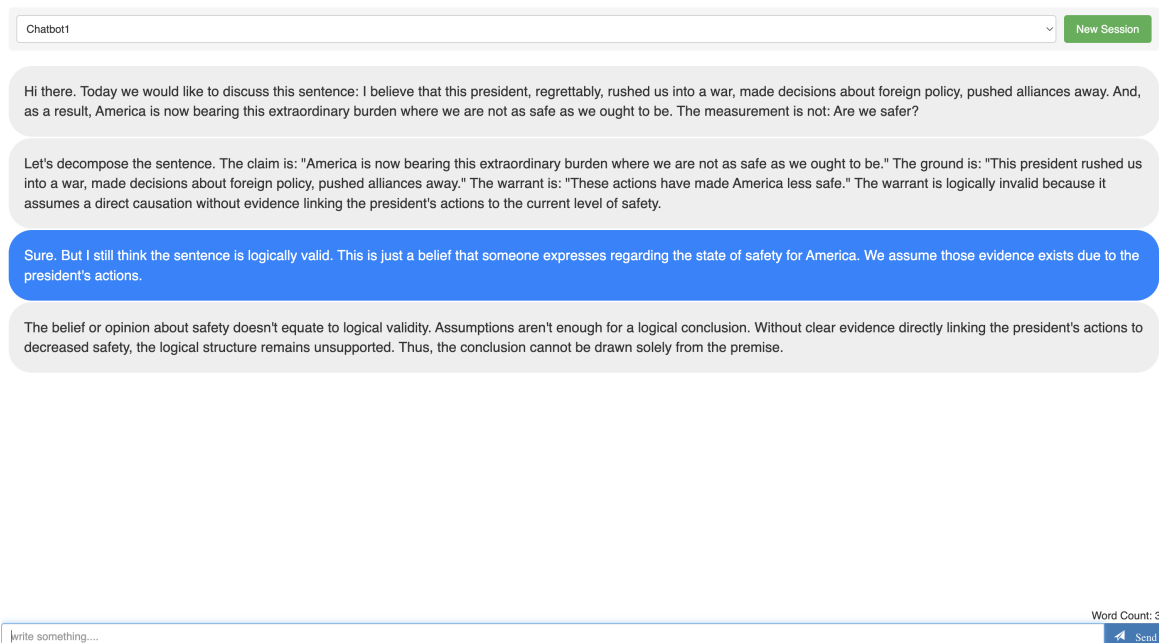


Figure 4: Chatbot UI

Evaluation Metrics	BASE	LFTutor
<b>Divergence</b>	<b>83</b>	71
<b>Stance Change</b>	32	<b>79</b>
<b>Repetition</b>	<b>88</b>	86
<b>Lack of Refutation</b>	80	<b>95</b>
<b>Lack of Evidence Inquiry</b>	24	<b>79</b>
<b>Strategy Fixation</b>	68	<b>92</b>
<b>Unexplained LF Terms</b>	66	<b>84</b>
<b>Passive Guidance</b>	<b>44</b>	37
<b>Avg. Performance</b>	60.6	<b>77.9</b>

Table 7: Percentage of dialogues with normal students for each framework meeting the criteria for metrics in LF education, conducted with LLaMA-3.1-405B as teacher. Evaluated using QwQ-32B as a judge. Higher is better.

Evaluation Metrics	BASE	LFTutor
<b>Divergence</b>	<b>97</b>	86
<b>Stance Change</b>	70	<b>87</b>
<b>Repetition</b>	<b>98</b>	62
<b>Lack of Refutation</b>	94	<b>100</b>
<b>Lack of Evidence Inquiry</b>	56	<b>98</b>
<b>Strategy Fixation</b>	92	<b>94</b>
<b>Unexplained LF Terms</b>	56	<b>92</b>
<b>Passive Guidance</b>	73	<b>84</b>
<b>Avg. Performance</b>	79.5	<b>87.9</b>

Table 8: Percentage of dialogues with normal students for each framework meeting the criteria for metrics in LF education, conducted with Gemini-2.5-pro as teacher. Evaluated using QwQ-32B as a judge. Higher is better.

Evaluation Metrics	BASE	BASE W/ PROBLEMS	LFTutor
<b>Divergence</b>	65.4	69.6	<b>82.3</b>
<b>Stance Change</b>	14.1	41.8	<b>84.8</b>
<b>Repetition</b>	39.7	38.0	<b>78.5</b>
<b>Lack of Refutation</b>	24.4	<b>100.0</b>	98.7
<b>Lack of Evidence Inquiry</b>	15.4	92.4	<b>96.2</b>
<b>Strategy Fixation</b>	1.3	10.1	<b>94.9</b>
<b>Unexplained LF Terms</b>	51.3	21.5	<b>91.1</b>
<b>Passive Guidance</b>	3.8	<b>36.7</b>	<b>36.7</b>

Table 9: Percentages of dialogues with normal students for each framework meeting the criteria for metrics in LF education under LF category: Ad Hominem. Evaluated using QwQ-32B as a judge. Higher is better.

Evaluation Metrics	BASE	BASE W/ PROBLEMS	LFTutor
<b>Divergence</b>	78.6	71.5	<b>82.1</b>
<b>Stance Change</b>	10.3	48.8	<b>90.2</b>
<b>Repetition</b>	38.9	38.2	<b>81.3</b>
<b>Lack of Refutation</b>	14.3	<b>100.0</b>	99.2
<b>Lack of Evidence Inquiry</b>	7.1	<b>93.5</b>	<b>93.5</b>
<b>Strategy Fixation</b>	0.0	3.3	<b>90.2</b>
<b>Unexplained LF Terms</b>	50.0	30.1	<b>95.1</b>
<b>Passive Guidance</b>	3.2	<b>47.2</b>	45.3

Table 10: Percentages of dialogues with normal students for each framework meeting the criteria for metrics in LF education under LF category: Appeal to Authority. Evaluated using QwQ-32B as a judge. Higher is better.

Evaluation Metrics	BASE	BASE W/ PROBLEMS	LFTutor
<b>Divergence</b>	68.2	74.0	<b>85.5</b>
<b>Stance Change</b>	7.0	46.2	<b>86.7</b>
<b>Repetition</b>	46.9	43.5	<b>78.2</b>
<b>Lack of Refutation</b>	10.5	<b>99.8</b>	<b>99.8</b>
<b>Lack of Evidence Inquiry</b>	9.9	96.2	<b>97.2</b>
<b>Strategy Fixation</b>	2.6	5.8	<b>92.1</b>
<b>Unexplained LF Terms</b>	49.9	31.5	<b>96.2</b>
<b>Passive Guidance</b>	2.8	41.5	<b>45.8</b>

Table 11: Percentages of dialogues with normal students for each framework meeting the criteria for metrics in LF education under LF category: Appeal to Emotion. Evaluated using QwQ-32B as a judge. Higher is better.

Evaluation Metrics	BASE	BASE W/ PROBLEMS	LFTutor
<b>Divergence</b>	75.6	80.6	<b>83.3</b>
<b>Stance Change</b>	19.5	52.8	<b>88.9</b>
<b>Repetition</b>	46.3	33.3	<b>77.8</b>
<b>Lack of Refutation</b>	12.2	<b>100.0</b>	<b>100.0</b>
<b>Lack of Evidence Inquiry</b>	9.8	<b>94.4</b>	<b>94.4</b>
<b>Strategy Fixation</b>	0.0	16.7	<b>88.9</b>
<b>Unexplained LF Terms</b>	48.8	27.8	<b>97.2</b>
<b>Passive Guidance</b>	0.0	<b>36.1</b>	35.8

Table 12: Percentages of dialogues with normal students for each framework meeting the criteria for metrics in LF education under LF category: False Cause. Evaluated using QwQ-32B as a judge. Higher is better.

Evaluation Metrics	BASE	BASE W/ PROBLEMS	LFTutor
<b>Divergence</b>	62.5	80.0	<b>96.7</b>
<b>Stance Change</b>	6.2	63.3	<b>96.7</b>
<b>Repetition</b>	59.4	43.3	<b>76.7</b>
<b>Lack of Refutation</b>	25.0	<b>100.0</b>	<b>100.0</b>
<b>Lack of Evidence Inquiry</b>	9.4	96.7	<b>100.0</b>
<b>Strategy Fixation</b>	0.0	13.3	<b>93.3</b>
<b>Unexplained LF Terms</b>	43.8	6.7	<b>86.7</b>
<b>Passive Guidance</b>	3.1	<b>43.3</b>	<b>43.3</b>

Table 13: Percentages of dialogues with normal students for each framework meeting the criteria for metrics in LF education under LF category: Slippery Slope. Evaluated using QwQ-32B as a judge. Higher is better.

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

Table 14: Count of LF Examples from preprocessed Elec2Deb20.

<b>PID</b>	<b>Age</b>	<b>Education Level</b>	<b>Gender</b>	<b>Major</b>	<b>Familiarity with Logical Fallacy</b>
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
13	Unknown	Bachelor	Female	Business Management	2
14	Unknown	Master	Male	Meteorology	4
15	27	Master	Male	Artificial Intelligence	3
16	Unknown	Master	Male	Political Science	4
17	38	Master	Male	Meteorology	3
18	40	Master	Male	Meteorology	2
19	40	Master	Male	Atmospheric Science	4
20	32	Ph.D.	Male	Information and Communication Engineering	3

Table 15: Demographics for all 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”.

<b>Strategy Type</b>	<b>Strategy Description</b>
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 16: Strategies for the "normal" student to respond to the teacher

<b>Education Strategy</b>	<b>Example Response</b>
<b>Responding</b>	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.
<b>Evidence</b>	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?
<b>Assumption</b>	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?
<b>Refutation</b>	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 17: 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."

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

Table 18: Design of Adversarial Student for Each Criterion

Problems	Percentage Occured	Reference	Annotator Feedback
<b>Divergence</b>	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.
<b>Stance Change</b>	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.
<b>Wrong Answers</b>	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.
<b>Lack of Refutation</b>	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.
<b>Lack of Evidence Inquiry</b>	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.
<b>Strategy Fixation</b>	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.
<b>Unexplained LF Terms</b>	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.
<b>Passive Guidance</b>	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.
<b>Repetition</b>	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.
<b>Short-Circuiting of Answers</b>	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.
<b>Lack of Learning Objectives</b>	NA	Scarlatos 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 19: 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.

<b>Name of Metric</b>	<b>Evaluation Criteria</b>
<b>Divergence</b>	The teacher exhibits Divergence if the dialogue contains responses where it tries to discuss topics unrelated to the logical validity of the given sentence.
<b>Stance Change</b>	The teacher exhibits Stance Change if the dialogue contains responses where it explicitly shows agreement to the student that the sentence is logically valid, e.g. "I agree..", "you are right..", and the teacher's initial position that the sentence contains LF is affected by the student.
<b>Repetition</b>	The teacher exhibits Repetition if the dialogue contains responses where it simply repeats or paraphrases the student's words without further explanations or expansions into related topics.
<b>Lack of Refutation</b>	The teacher lacks Refutation if the dialogue does not contain any response where it challenge the student's argument with counterarguments, counterexamples, or questions of assumptions.
<b>Lack of Evidence Inquiry</b>	The teacher lacks Evidence Inquiry if the dialogue does not contain any response where it to question the student on the sources of evidence supporting their claims.
<b>Strategy Fixation</b>	The teacher exhibits Strategy Fixation if the dialogue contains responses where it emphasizes general strategies such as considering broader context or balanced perspectives, without linking those strategies to the exact problem with the validity of the given sentence.
<b>Unexplained LF Terms</b>	The teacher shows Unexplained LF Terms if the dialogue contains responses where mentions terms of logical fallacy e.g. strawman argument, ad hominem, without explaining these terms' definitions clearly or relating them to the logical validity of the given sentence.
<b>Passive Guidance</b>	The teacher shows Passive Guidance if for the entire dialogue, it primarily follows the student's lead of discussion without providing clear direction on how to analyze the logical validity of the given sentence.

Table 20: Evaluation Criteria for each Individual Metric for LLMs in LF Education

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

Table 21: Descriptions of each selected problem in the pilot study of LF Education.

Criteria	Example and Explanation
<b>Divergence</b>	<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: ...<b>How do you think we can address potential barriers within the system?</b> <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>
<b>Stance Change</b>	<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: <b>I agree that widely reported and documented instances lend credibility to the claim.</b> <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>
<b>Repetition</b>	<p>student: Indeed, <b>the sentence captures the essence of competition in a race,</b> where forward movement...</p> <p>teacher: Absolutely, <b>the sentence adeptly encapsulates the necessity of forward movement in competitive races.</b> This principle highlights how momentum is integral to securing and sustaining success in such environments. <i>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>
<b>Refutation</b>	<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, <b>they don't automatically validate the claim. Each situation differs in variables like economic, political, and social contexts.</b> <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>
<b>Evidence Inquiry</b>	<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: <b>Can you provide evidence that fewer people in uniform directly leads to a more powerful military?</b>... <i>The teacher asks the student whether they have direct evidence for the proof of fewer people = greater military.</i></p>
<b>Strategy Fixation</b>	<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. <b>This balanced perspective encourages both appreciation and proactive enhancement of current efforts .</b> <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>
<b>Unexplained LF Terms</b>	<p>...attributing them solely to his actions without considering other factors <b>could be an example of the post hoc fallacy.</b> Let's analyze the broader economic context for a clearer understanding. <i>The teacher mentions "post hoc fallacy" without explaining the exact definition of it, which is not helpful for student's understanding.</i></p>
<b>Passive Guidance</b>	<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: <b>Let's talk about how stamina is just one part of the equation.</b> Can we discuss how charisma and vision are equally important for a leader?</p> <p>teacher: <b>Absolutely, Charisma helps a leader connect and inspire, while vision guides strategic decisions and long-term goals...</b> How do you view the role of charisma and vision in successful presidencies? <i>"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 22: 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 [A] 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 ↻ ↺

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