

Tackling the Root of Misinformation by Teaching Laypeople about Logical Fallacies via Socratic Questioning and Critical Argumentation

Minjing Shi Junling Wang Jingwei Ni Sankalan Pal Chowdhury Mrinmaya Sachan
ETH Zürich

shimin@student.ethz.ch, mj.shi@foxmail.com junling.wang@ai.ethz.ch
jingwei.ni@gess.ethz.ch {sankalan.palchowdhury, mrinmaya.sachan}@inf.ethz.ch

Abstract

Identifying logical fallacies 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 laypeople 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 these 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—flawed but persuasive and misleading reasoning patterns—are a key driver of misinformation (Jin et al., 2022). The prevalence of logical fallacies in public discourse (Risen et al., 2007) makes it crucial to develop systems that help people detect logical fallacies in public misinformation.

To combat such prevalence, prior work focuses on automatic detection of fallacies (Jin et al., 2022; Pan et al., 2024; Jeong et al., 2025). Although such methods can flag potential fallacies for user’s attention, they do not help people recognize them independently. Instead of passively feeding users with fallacy detection output, it can be more beneficial to focus on LLM-based *educational* approaches that (1) nudge users to actively discuss and reconsider

logical arguments (Shahzad et al., 2025); (2) help users overcome cognitive biases reinforced by the reasoning structures in fallacies (Correia, 2011); and (3) raise users’ awareness if they lack a proper understanding of fallacies (Berkle et al., 2023).

Developing an LLM-based tutor for fallacies presents several challenges: **C1**. It is difficult to evaluate tutors – existing metrics for educational dialogues may not work well for logical fallacies (Wang et al., 2024b). **C2**. LLMs face several challenges when acting as logical fallacy educators, including diversion from the topic (Pal Chowdhury et al., 2024), concession to persuasion (Xu et al., 2024), mirroring student texts 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 teaching laypeople to recognize logical fallacies along with a scalable way to evaluate such tutors. Specifically, to address **C1**, we define metrics for logical fallacy tutoring based on our findings in a pilot study and relevant literature in education and argumentation theory. To address **C2**, we propose LFTutor, a tutoring framework based on intent-based pedagogical steering (Puech et al., 2024), which aims to help learners identify fallacious arguments through structured, designated strategies in a dialogue setting, avoiding current issues of LLMs as logical fallacy tutors.

We conduct both automatic and human evaluation to verify the effectiveness of LFTutor. For automatic evaluation, we use simulated students to generate dialogues with LFTutor and baseline LLMs with texts from Elec2Deb20 (Goffredo et al., 2023). We then use QwQ-32B (verified against human judges) to evaluate tutor responses on all metrics. For human evaluation, we recruit 20 participants who interact with LFTutor and then rate its performance against baseline LLMs. Results show that LFTutor achieves superior performance compared to baseline LLMs in both automatic and

human evaluation, demonstrating the possibility of using LFTutor in combatting fallacies 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 educating laypeople about fallacies, using intent-based pedagogical steering (§ 4).
3. We conduct automatic and human evaluations, showcasing LFTutor’s superiority in fallacy 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 logical fallacies often arise from faulty reasoning (Goffredo et al., 2023), Socratic questioning can be particularly relevant in helping learners examine and reflect on their reasoning.

Intent-based Dialogue Strategies. Recent work 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 in everyday dialogue scenarios. Ruggeri et al. (2023) formulated debate strategies to annotate dialogues on a 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.

Critical Argumentation in LLMs. Critical argumentation (Walton, 2008) embodies systematic ways of analyzing arguments in dialogues, which is useful for responding to opposing viewpoints with counterarguments. In earlier work, Sinott-Armstrong and Frogelin (2015) proposed methods to generate logically sound counterarguments. More recently, Ozaki et al. (2025) showed LLM’s ability to produce high-quality counterarguments when targeting implicit or critical premises. Pitre

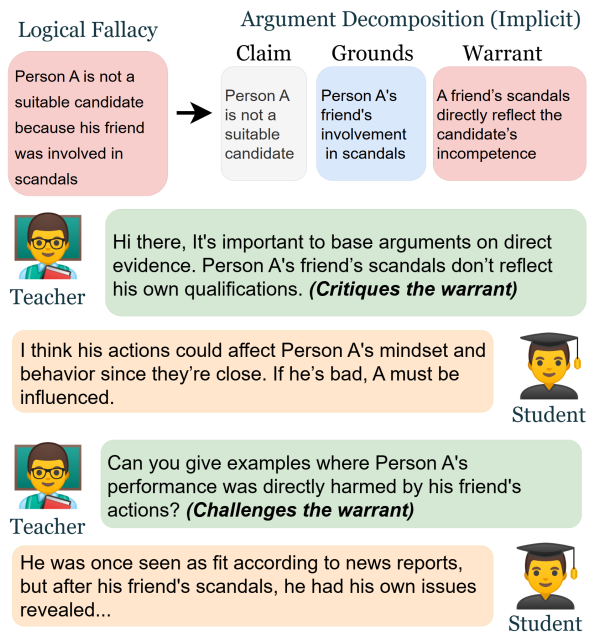


Figure 1: The task of logical fallacy tutoring consists of a fallacious statement decomposed into its corresponding argumentation structure and two personas: a teacher, and a student (layperson) who believes in the fallacy. The logical flaw here lies in the warrant. As shown by the **bolded** comments, the teacher focuses on the flaw in the warrant, while the student tries to defend its validity.

and Luther (2024) used LLMs to generate counterarguments for op-ed articles to facilitate reader’s thinking, and Jeong et al. (2025) incorporated counterarguments into LLM’s reasoning for enhanced logical fallacy detection, which is partially utilized in our work for LFTutor.

3 Problem Formulation

We frame the task of logical fallacy tutoring as a one-on-one interaction between a **teacher** and a **student** centered around a piece of potentially fallacious text (Fig. 1). The text contains one or more fallacies of any type (e.g. ad hominem¹, appeal to emotion², etc.). The teacher aims to educate the student by helping them realize that the text is logically flawed, while the student, who is unfamiliar with fallacies, believes that it is logically valid, and defends their position by debating with the teacher. This multi-turn dialogue setting is useful as it: (1) enables deeper explorations of the fallacious statement using argument decomposition, and (2) allows the teacher to actively engage with the student’s underlying cognitive biases, going beyond surface-level identification of fallacies.

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

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

A major challenge in this task is that logical fallacies 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 context is critical in judging logical validity. To reason about these contexts, we utilize argument decomposition models such as the Toulmin model (Toulmin et al., 1984), which posits that a logical argument consists of a 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 grounds to the claim – *a friend’s scandals directly reflect the candidate’s incompetence*). This decomposition allows the teacher in Fig. 1 to focus on asking the learner for more evidence for the warrant.

Another major challenge in misinformation education lies in the learner’s cognitive biases, which are often reinforced by reasoning structures and deep-rooted assumptions (Correia, 2011), as shown in their first response equating close relationships with direct influence (Fig. 1). Thus, laypeople’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 learner engage in critical reflection of reasoning steps. This approach involves argument exchanges, where the teacher builds the discussion with probing questions and targeted challenges to gradually reveal the logical flaw of the learner’s reasoning.

We investigate LLMs’ performance in educating laypeople about logical fallacies in this setting. Specifically, we choose Elec2Deb20 (Goffredo et al., 2023), a dataset on U.S. presidential debates (more details in App. B).

3.1 LLM’s Issues in Logical Fallacy Tutoring

To our knowledge, no prior work has systematically discussed evaluation metrics for LLM tutors in logical fallacy tutoring. As a first step towards filling this gap, we conduct a pilot study (App. A.3) to identify weaknesses of current LLMs in this task using student-teacher dialogues simulated by GPT-4o. Furthermore, we draw on interdisciplinary insights into combating logical fallacies with reviews of relevant literature in education and logical reasoning. Based on findings from both the pilot study and the literature review, we propose the following

evaluation dimensions. Detailed examples and explanations of each issue are presented in Table 25.

Divergence. Our pilot study found that 35% of dialogues contain tutor responses unrelated to discussion of the logical fallacy (e.g. proposing solutions to the political issue in the sentence). However, topic relevance is crucial in education for facilitating meaningful discussions (Tracy, 1984). We thus consider Divergence as a key evaluation dimension.

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). Thus, it is important for the teacher to hold their position when debating, as this might hinder the 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 as they may increase the chance of the LLM tutor generating repetitive responses (Xu et al., 2022), leading to student confusion and poorer quality in teaching.

Lack of Refutation. Prior work indicates that LLMs demonstrate sycophancy and lack the skills for “sustained” debates, including posing counterarguments (Zhang et al., 2024). This issue was also present in the pilot study, where 53% of dialogues lacked explicit challenges to the student. It is crucial for the teacher to actively refute the student’s illogical argument when necessary, which helps clarify misconceptions (Kowalski and Taylor, 2009) on the logical structure of the fallacies.

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. In logical fallacy tutoring, 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 texts with multiple faulty logical 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 logical fallacy 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 work (Pal Chowdhury et al., 2024; Liu et al., 2024) reports that LLMs often fail to assert pedagogical control in education dialogues. Guidance is important in logical fallacy tutoring, 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 (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 23, 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) from the Qwen family for our LLM-as-judge evaluations.

We verify the reliability of QwQ-32B as a judge by comparing its annotations against those from human experts. We randomly sampled 100 logically fallacious text from Elec2Deb20 and generated 100 dialogues of 10 rounds with the setting in App. A.3. We then let QwQ-32B evaluate all dialogues using prompts containing metrics in App. A.5.

For human-annotated golden labels, we recruited three human annotators who are fluent in English and have prior experience with linguistic annotation tasks. We first explained the metrics to each annotator using examples from table 25. The annotators then evaluated dialogues with the same prompts as QwQ-32B. After that, they resolved conflicting labels, reaching an agreement level of over 97% (Miles et al., 2020) for all metrics.

For each metric, we randomly sampled 30 golden dialogues with 15 positive labels and 15 negative labels. We then computed F1-scores (Table 5) for QwQ-32B against golden labels, which average over 73% for all metrics. We further validate QwQ-32B’s performance by manually reviewing human annotations, having observed that the task is difficult for human annotators, involving long text (700 words) with complex arguments. We performed a direct inspection of the annotator’s labels and measured the accuracy of QwQ-32B’s annotation, shown in Table 6. We found that the accuracy of QwQ-32B was greater than or equal to 80% for all metrics, which ensures its reliability as the judge.

4 LFTutor

As stated in § 3, logical fallacy education faces two main challenges: 1. Logical fallacies hinge on contextual assumptions, which requires argument decomposition to pinpoint; 2. Fallacy believers often harbor cognitive biases and are resistant to one-shot identification and explanation. To address them, 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 the Toulmin model to identify the text component containing logical flaws. It then applies a strategy tailored to that component to guide the student towards recognizing flaws. This technique, called intent-based pedagogical steering (Puech et al., 2024), structures the tutor’s output by detecting the student’s intent (implied purposes behind their response, e.g., asking for clarification), and 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. Detailed pseudocode is shown in Algorithm 1.

Step 1: Disagreement Check. To structure the tutor-student dialogue and help LFTutor focus on specific parts of the text they disagree on, we introduce *disagreement bank*, a long-term memory module (Zhong et al., 2024) that stores a list of disagreement points between the two parties. Given the student’s response from the last turn, the text with logical fallacies, and the disagreement bank, LFTutor checks the response using another agent

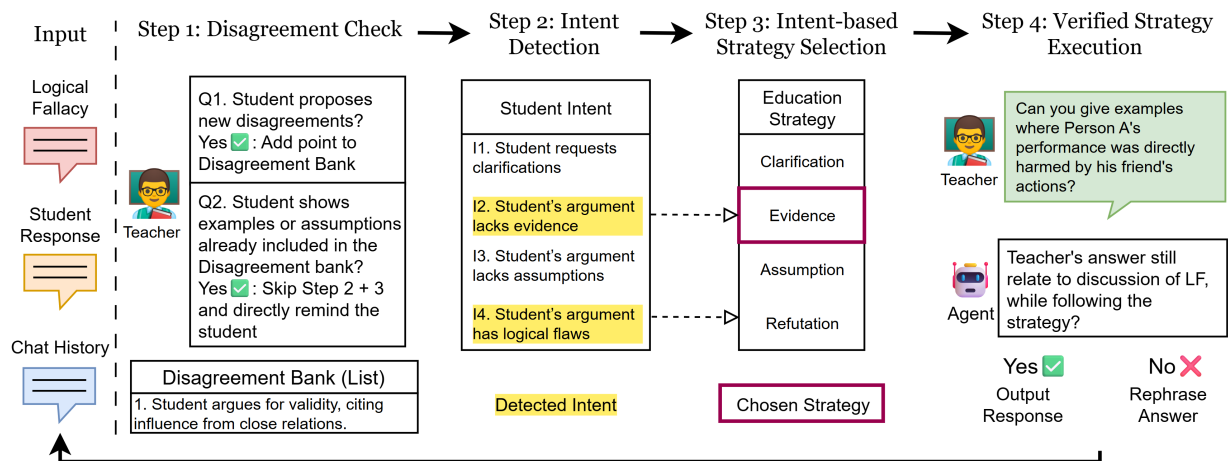


Figure 2: LFTutor processes the student’s response in four steps. In step 1, LFTutor checks the response for (1) new conflicts, which are added to the disagreement bank, or (2) duplicate examples or assumptions, which leads to reminders to propose new examples or assumptions, skipping step 2 and 3. In 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 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.

(App. F.1) for whether (1) the student’s claim, if opposing, is contained in the bank (2) the examples or assumptions are contained in the bank. If the opposing claim is newly made, then LFTutor adds the complete argument to the disagreement bank, which it aims to address in this round, regardless of the contents from examples or assumptions. However, if the claim has been discussed, then the examples or assumptions must be different from those in the bank. Otherwise, LFTutor will skip the next steps and request the student to provide new arguments that support their position.

Step 2: Intent Detection. Given the student’s argument, LFTutor analyzes it (App. F.2) for potential intents: (I1) student requests clarification on the logical flaw of the text; (I2) 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 the Toulmin 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. All detected intents (one or more) are stored for LFTutor’s reference in the next step.

Step 3: Intent-based Strategy Selection. Given the detected intents, LFTutor focuses on addressing the student’s reasoning flaws by selecting an appropriate education strategy related to their intent.

We draw strategies from two main pedagogical theories: (1) Socratic questioning (Ed and Ed, 2002), which aims for dialogic, open-ended discussions on logical fallacies by 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, nudging them to reflect on argument quality. We include four strategies corresponding to the aforementioned student intents: Responding, Evidence, Assumption, and Refutation. Prompts and example responses are shown in App. F.3 and Table 20.

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

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

(iii) Assumption: When the student’s response makes unclear assumptions, LFTutor challenges the student’s hidden or unsupported assumptions in the argument. This is akin to Socratic questions on assumptions discussed in (Yang et al., 2005).

(iv) Refutation: When the student’s argument contain clear logical flaws, LFTutor provides logical refutations to the student’s argument, using methods adopted from critical argumentation (Sinott-Armstrong and Frogelin, 2015).

LFTutor selects strategies in the order listed

above (i \rightarrow iv), matching the priority of components in the Toulmin model (Hitchcock, 2017): claims form an argument’s foundation (\rightarrow Responding), followed by grounds (\rightarrow Evidence), then warrants (\rightarrow Assumption), and finally, rebuttal to a complete argument (\rightarrow Refutation). Clarification of logical flaw(s) (Responding) introduces a counter-claim, leading the student to reconsider and refine their position. After establishing the claim, LFTutor presses the student to provide relevant proof (Evidence). Assumption comes once claim and grounds are both present, and LFTutor challenges the student to form connections between them. Finally, LFTutor applies Refutation when it obtains a complete argument, urging the student to examine their argument’s logical validity. By structuring strategies according to this hierarchy, LFTutor guides the student through argument construction in a stepwise fashion by systematically exposing component-level flaws.

Step 4: Verified Strategy Execution. In the final step, LFTutor applies the chosen strategy to respond to the student (App. F.4). Then, another LLM verifies if the response adheres to the chosen strategy (App. F.5). If this verification fails, the LLM in 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 begin by evaluating LFTutor using a set of automatic evaluations. We first describe the baseline LLM tutors that LFTutor will be compared with (§ 5.1) and the student design used in our evaluations (§ 5.2). We then describe the detailed experimental settings (§ 5.3) and the results (§ 5.4).

5.1 Baseline LLM Tutors

We designed two baselines for the tutor LLM for comparison. The first (BASE) uses a role-playing prompt, where the tutor is instructed to act as an educator whose goal is to convince a potentially biased student that the given text contains logical fallacies (App. E.1). The second (BASE W/ PROBLEMS) improves upon BASE by incorporating descriptions of the issues in logical fallacy education (§ 3.1) in addition to role-playing instructions, and the tutor is asked to avoid those issues (App. E.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 LLM, 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 tutor by defending their position with evidence and explanations, believing that the text containing logical fallacies is logically valid (App. G.2). To create a “robust” student, we prompt the LLM to utilize various debating techniques (Table 19) in critical argumentation (Walton, 2008; Sinott-Armstrong and Frogelin, 2015).

The adversarial student intends to trigger the tutor’s issues discussed in § 3.1. We use role-playing techniques (Zhou et al., 2024) to design students that drive the tutor towards behaviors matching the issues’ descriptions (App. G.3). For “Divergence”, for example, we ask the LLM to play a student who often talks about information irrelevant to the discussion of the logical fallacy. This way, we evaluate the robustness of tutors under worst case scenarios. We exclude adversarial students for Lack of Refutation and Lack of Evidence Inquiry, as pilot studies show that these issues are primarily attributed to the lack of instructions in tutor’s prompts.

5.3 Experiment Setting

We begin by experimenting with GPT-4o as our base LLM. We divide our setup into two parts following our design from § 5.2, with more 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 examples from Elec2Deb20, evaluated using QwQ-32B as a judge (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, evaluating them using QwQ-32B on the corresponding metrics (e.g., evaluating Divergence on dialogues using divergent student). Results are shown in Table 2.

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

Table 1: Percentages of dialogues with normal students for each framework passing each metric, with GPT-4o as backbone. Evaluated using QwQ-32B. Higher is better.

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

Table 2: Percentage of dialogues with adversarial students passing the corresponding metric, with GPT-4o as backbone. Evaluated using QwQ-32B. Higher is better.

We also tested Gemini-2.5-pro (Comanici et al., 2025) and LLaMA-3.1 (Grattafiori et al., 2024) as backbones for BASE and LFTutor under the **normal student** setting, using 200 random examples from Elec2Deb20. The results are broadly consistent with the findings from GPT-4o. More details can be found in App. C.

5.4 Results

With the “normal” student design, 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 (§ 7), which shows that they stem mostly from strategy verification and intent-based strategy selection. Strategy verification ensures that 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. Strategy priority also matters, as it improves Guidance through progressively uncovering flaws within the student’s arguments.

While BASE W/ PROBLEMS shows improve-

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

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.

ments 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 logical fallacy education is not enough for maintaining dialogue quality over time.

For the adversarial student design, LFTutor outperforms BASE W/ PROBLEMS by up to 31% on average in all adversarial student types. The difference 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 is robust from external influences.

However, LFTutor suffers performance drops in Divergence and Strategy Fixation. For Divergence, LFTutor’s effort is shifted from educating students about logical fallacies to preventing them 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 the fallacy’s broader context.

6 Human Evaluations

Finally, to understand how LFTutor might work in a more real-world setting, we conduct pilot human evaluations comparing LFTutor and BASE. We choose BASE as an evaluation baseline since it

resembles behaviors of real-world LLMs with minimal knowledge of issues in logical fallacy education, instead of BASE w/ PROBLEMS that performs slightly better in automatic evaluation but fails to reflect LLMs’ real-world exposure to logical fallacies. 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 texts from Elec2Deb20 that participants voted logically valid (App. I.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 the tutors, with the order of tutors and sentences randomly chosen.

During the study, every participant interacted with each tutor for 5 to 10 rounds each. They evaluated the tutor after each session (Fig. 8) with metrics (Table 23) 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 the tutors LFTutor and BASE. In addition to existing metrics, we included “Helpfulness” to measure the subjective learning helpfulness of the tutor for the participant, which they rated on the same Likert scale. A score of 1 indicated least helpful for logical fallacy 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 for 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.

Finally, we also performed a summative qualitative analysis of the participants’ responses, by examining their chat histories and post-study forms, and then identifying key phrases from their descriptions of LFTutor and BASE and extracting recurring themes that came up in the dialogues.

6.2 Quantitative Analysis on Ratings

LFTutor achieves statistically significant better scores than BASE on eight out of nine metrics. The improvement is significant ($p < 0.05$) for Strat-

egy 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 tutoring logical fallacy 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 of Dialogues

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 logical fallacies in the text. 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; on the other hand, student LLMs argue with consistent, low-variation logic. Compared with LLM-simulated students, we found that our participants employed a wider range of strategies targeting the weaknesses of the tutor’s responses. For example, five of the participants challenged the tutors by claiming that the situations they described seem too ideal, while three participants tried to present direct counterexamples to the tutor’s argu-

Variant	Div.	Stance	Repet.	No Refut.	No Evid.	Strat. Fix.	Unexpl. LF	Passive
LFTutor (Reference)	84.6	87.9	78.3	99.6	96.1	91.2	95.0	43.6
– No Disagreement Check	83↓1.6	78↓9.9	92 ↑13.7	100 ↑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	92 ↑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 4: Percentages of dialogues passing each metric 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 automatically using QwQ-32B as a judge.

ment. Moreover, we found that the participant’s arguments seemed more inconsistent, often relying on emotional appeals (e.g. “You don’t care about my opinion”) or attributions to a single cause. In contrast, we found that 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 complete with claims followed by evidence, yet they lacked variation in strategies and are often repetitive.

7 Ablations

To investigate the contribution of different design choices in LFTutor towards evaluation improvements, we conducted an ablation study under four settings: (1) removing step 1: Disagreement Check, (2) removing strategy verification and correction, (3) reversing the priority of pedagogical strategies in step 3, from (Responding → Evidence → Assumption → Refutation) to (Refutation → Assumption → Evidence → Responding), and (4) allowing multiple pedagogical strategies per turn by ignoring Intent-based Strategy Selection. Due to budget limitations, we conducted the study using the normal student setting discussed in § 5.3 with 100 sentences randomly selected from Elec2Deb20. Results under these settings are shown in Table 4.

Removing Step 1: Disagreement Check led to reduced performance 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. Passive Guidance suffers the most among all metrics, 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 significantly reduces LFTutor’s performance in most metrics. The decrease of scores is significant in Divergence, Lack of Evidence Inquiry, and

Stance Change. 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 harms performance the most in Unexplained LF Terms and Passive Guidance, with minor changes of performance in other metrics. Setting Refutation as the top strategy led to unwanted usage of LF terms, since refutation by pointing out the LF in the opponent’s words is common (shown in Table 1) 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 attempting to uncover the student’s supporting evidence and assumptions.

Allowing Multiple Strategies per Turn resulted in decreased performance across all metrics, of which Divergence, Stance Change, Strategy Fixation, and Unexplained LF Terms suffered 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.

8 Conclusion

We explored the novel task of educating laypeople to identify logical fallacies in everyday discourse through multi-turn dialogues with LLM-based tutors. We introduced eight evaluation metrics to assess tutor responses and developed LFTutor, an LLM tutor system with intent-based pedagogical strategies that helps learners recognize logical fallacies. Our extensive automatic and human evaluation showed that LFTutor is an effective and pedagogically grounded solution for misinformation education, paving a way for designing more AI tutors that foster critical thinking through structured reflection, reasoning and interactive engagement.

Limitations

Limited Performance Gain on Guidance. Although LFTutor managed to improve upon the BASE framework on all criteria during automatic evaluation, it did not achieve high scores ($\geq 75\%$) in Passive Guidance. We attribute this to the lack of structured, detailed, and progressive dialogue goals (e.g. first tackle the evidence, then tackle the warrant, then narrow down the discussion to certain aspects of the warrant.), which needs to be closely integrated with tutoring strategies to create responses that are robust to passive influence. Future works could consider introducing a list of high-level education goals and procedures for LFTutor to follow, such that it combines those goals and tutoring strategies to produce responses that actively control the flow of the dialogues.

Scope of Evaluation Metrics In this paper, we investigated recurring problems with LLM tutors in dialog-based tutoring of logical fallacies and proposed an automatic evaluation framework with a curated set of eight evaluation metrics motivated by prior work and empirical relevance. While the proposed evaluation framework covers several key dimensions for measuring the quality of logical fallacy educational dialogues, such dialogues are inherently multi-faceted. Additional aspects, such as emotional support, personalization, and ease of understanding are not captured by our current metrics and they represent promising directions for future work.

Acknowledgments

This work was funded by grants from the Swiss National Science Foundation (Grant number 10009282) and the Swiss AI initiative (<https://www.swiss-ai.org/compute-grants>). We would like to thank all participants involved in the annotation and interactive experiments, as well as the anonymous reviewers for their constructive comments which helped improve this work. We also give special thanks to Prof. Daqiang Li from Jilin University and Prof. Jose A. Gascon from University of Murcia for insightful discussions.

References

Yvonne Berkle, Lukas Schmitt, Antonia Tolzin, Andreas Janson, Thiemo Wambsganss, Jan Marco Leimeister, and Miriam Leuchter. 2023. *Measur-*

ing university students' ability to recognize argument structures and fallacies. *Frontiers in Psychology*, 14.

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

Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, Krishna Haridasan, Ahmed Omran, Nikunj Saunshi, Dara Bahri, Gaurav Mishra, Eric Chu, Toby Boyd, Brad Hekman, Aaron Parisi, Chaoyi Zhang, Kornrathop Kawintiranon, Tania Bedrax-Weiss, Oliver Wang, Ya Xu, Ollie Purkiss, Uri Mendlovic, Ilai Deutel, Nam Nguyen, Adam Langley, Flip Korn, Lucia Rossazza, Alexandre Ramé, Sagar Waghmare, Helen Miller, Nathan Byrd, Ashrith Sheshan, Raia Hadsell, Sangnie Bhardwaj, Pawel Janus, Tero Rissa, Dan Horgan, Alvin Abdagic, Lior Belenki, James Allingham, Anima Singh, Theo Guidroz, Srivatsan Srinivasan, Herman Schmit, Kristen Chiafullo, Andre Elisseeff, Nilpa Jha, Prateek Kolhar, Leonard Berrada, Frank Ding, Xiance Si, Shrestha Basu Mallick, Franz Och, Sofia Erell, Eric Ni, Tejasi Latkar, Sherry Yang, Petar Sirkovic, Ziqiang Feng, Robert Leland, Rachel Hornung, Gang Wu, Charles Blundell, Hamidreza Alvari, Po-Sen Huang, Cathy Yip, Sanja Deur, Li Liu, Gabriela Surita, Pablo Duque, Dima Damen, Johnson Jia, Arthur Guez, Markus Mircea, Animesh Sinha, Alberto Magni, Paweł Stradomski, Tal Marian, Vlado Galić, Wenhua Chen, Hisham Husain, Achintya Singhal, Dominik Grewe, François-Xavier Aubet, Shuang Song, Lorenzo Blanco, Leland Rechis, Lewis Ho, Rich Munoz, Kelvin Zheng, Jessica Hamrick, Kevin Mather, Hagai Taitelbaum, Eliza Rutherford, Yun Lei, Kuangyuan Chen, Anand Shukla, Erica Moreira, Eric Doi, Berivan Isik, Nir Shabat, Dominika Rogozińska, Kashyap Kolipaka, Jason Chang, Eugen Vušak, Srinivasan Venkatachary, Shadi Noghabi, Tarun Bharti, Younghoon Jun, Aleksandr Zaks, Simon Green, Jeshwanth Challagundla, William Wong, Muqthar Mohammad, Dean Hirsch, Yong Cheng, Iftekhar Naim, Lev Proleev, Damien Vincent, Aayush Singh, Maxim Krikun, Dilip Krishnan, Zoubin Ghahramani, Aviel Atias, Rajeew Aggarwal, Christo Kirov, Dimitrios Vytiniotis, Christy Koh, Alexandra Chronopoulou, Pawan Dogra, Vlad-Doru Ion, Gladys Tyen, Jason Lee, Felix Weissenberger, Trevor Strohman, Ashwin Balakrishna, Jack Rae, Marko Velic, Raoul de Liedekerke, Oded Elyada, Wentao Yuan, Canoe Liu, Lior Shani, Sergey Kishchenko, Bea Alessio, Yandong Li, Richard Song, Sam Kwei, Orion Jankowski, Aneesh Pappu, Youhei Namiki, Yenai Ma, Nilesh Tripuraneni, Colin Cherry, Marissa Ikonomidis, Yu-Cheng Ling, Colin Ji, Beka Westberg, Auriel Wright, Da Yu, David Parkinson, Swaroop Ramaswamy, Jerome Connor, Soheil Hassas Yeganeh, Snchit Grover, George Kenwright, Lubo Litchev, Chris Apps, Alex Tomala, Felix Halim, Alex Castro-Ros, Zefei Li, Anudhyan Boral, Pauline Sho, Michal Yarom, Eric Malmi, David Klinghoffer, Rebecca Lin, Alan Ansell,

Pradeep Kumar S, Shubin Zhao, Siqi Zuo, Adam Santoro, Heng-Tze Cheng, Solomon Demmessie, Yuchi Liu, Nicole Brichtova, Allie Culp, Nathaniel Braun, Dan Graur, Will Ng, Nikhil Mehta, Aaron Phillips, Patrik Sundberg, Varun Godbole, Fangyu Liu, Yash Katariya, David Rim, Mojtaba Seyedhosseini, Sean Ammirati, Jonas Valfridsson, Mahan Malihi, Timothy Knight, Andeep Toor, Thomas Lampe, Abe Ittycheriah, Lewis Chiang, Chak Yeung, Alexandre Fréchet, Jinneng Rao, Huisheng Wang, Himanshu Srivastava, Richard Zhang, Rocky Rhodes, Ariel Brand, Dean Weesner, Ilya Figotin, Felix Gimeno, Rachana Fellinger, Pierre Marcenac, José Leal, Eyal Marcus, Victor Cotruta, Rodrigo Cabrera, Sheryl Luo, Dan Garrette, Vera Axelrod, Sorin Baltateanu, David Barker, Dongkai Chen, Horia Toma, Ben Ingram, Jason Riesa, Chinmay Kulkarni, Yujing Zhang, Hongbin Liu, Chao Wang, Martin Polacek, Will Wu, Kai Hui, Adrian N Reyes, Yi Su, Megan Barnes, Ishaan Malhi, Anale Siddiqui, Qixuan Feng, Mihai Damaschin, Daniele Pighin, Andreas Steiner, Samuel Yang, Ramya Sree Boppana, Simeon Ivanov, Arun Kandoor, Aditya Shah, Asier Mujika, Da Huang, Christopher A. Choquette-Choo, Mohak Patel, Tianhe Yu, Toni Creswell, Jerry, Liu, Catarina Barros, Yasaman Razeghi, Aurko Roy, Phil Culliton, Binbin Xiong, Jiaqi Pan, Thomas Strohmman, Tolly Powell, Babi Seal, Doug DeCarlo, Pranav Shyam, Kaan Katircioglu, Xuezhi Wang, Cassidy Hardin, Immanuel Odisho, Josef Broder, Oscar Chang, Arun Nair, Artem Shtefan, Maura O'Brien, Manu Agarwal, Sahitya Potluri, Siddharth Goyal, Amit Jhinda, Saksham Thakur, Yury Stuken, James Lyon, Kristina Toutanova, Fangxiaoyu Feng, Austin Wu, Ben Horn, Alek Wang, Alex Cullum, Gabe Taubman, Disha Shrivastava, Chongyang Shi, Hamish Tomlinson, Roma Patel, Tao Tu, Ada Maksutaj Oflazer, Francesco Pongetti, Mingyao Yang, Adrien Ali Taïga, Vincent Perot, Nuo Wang Pierse, Feng Han, Yoel Drori, Iñaki Iturrate, Ayan Chakrabarti, Legg Yeung, Dave Dopson, Yi ting Chen, Apoorv Kulkshreshtha, Tongfei Guo, Philip Pham, Tal Schuster, Junquan Chen, Alex Polozov, Jinwei Xing, Huanjie Zhou, Praneeth Kacham, Doron Kukliansky, Antoine Miech, Sergey Yaroshenko, Ed Chi, Sholto Douglas, Hongliang Fei, Mathieu Blondel, Preethi Myla, Lior Madmoni, Xing Wu, Daniel Keysers, Kristian Kjems, Isabela Albuquerque, Lijun Yu, Joel D'sa, Michelle Plantan, Vlad Ionescu, Jaume Sanchez Elias, Abhirut Gupta, Manish Reddy Vuyyuru, Fred Alcober, Tong Zhou, Kaiyang Ji, Florian Hartmann, Subha Puttagunta, Hugo Song, Ehsan Amid, Anca Stefanoiu, Andrew Lee, Paul Pucciarelli, Emma Wang, Amit Raul, Slav Petrov, Isaac Tian, Valentin Anklin, Nana Nti, Victor Gomes, Max Schumacher, Grace Vesom, Alex Panagopoulos, Konstantinos Bousmalis, Daniel Andor, Josh Jacob, Yuan Zhang, Bill Rosgen, Matija Kecman, Matthew Tung, Alexandra Belias, Noah Goodman, Paul Covington, Brian Wieder, Nikita Saxena, Elnaz Davoodi, Muhuan Huang, Sharath Maddineni, Vincent Roulet, Folawiyo Campbell-Ajala, Pier Giuseppe Sessa, Xintian, Wu, Guangda Lai, Paul Collins, Alex Haig, Vytienis Sakenas, Xiaowei Xu, Marissa Giustina, Lau-

rent El Shafey, Pichi Charoenpanit, Shefali Garg, Joshua Ainslie, Boone Severson, Montse Gonzalez Arenas, Shreya Pathak, Sujee Rajayogam, Jie Feng, Michiel Bakker, Sheng Li, Nevan Wichers, Jamie Rogers, Xinyang Geng, Yeqing Li, Rolf Jagerman, Chao Jia, Nadav Olmert, David Sharon, Matthew Mauer, Sandeep Mariserla, Hongxu Ma, Megha Mohabey, Kyuyeun Kim, Alek Andreev, Scott Pollom, Juliette Love, Vihan Jain, Priyanka Agrawal, Yannick Schroecker, Alisa Fortin, Manfred Warmuth, Ji Liu, Andrew Leach, Irina Blok, Ganesh Poomal Girirajan, Roe Aharoni, Benigno Uria, Andrei Sozanschi, Dan Goldberg, Lucian Ionita, Marco Tulio Ribeiro, Martin Zlocha, Vighnesh Birodkar, Sami Lachgar, Liangzhe Yuan, Himadri Choudhury, Matt Ginsberg, Fei Zheng, Gregory Dibb, Emily Graves, Swachhand Lokhande, Gabriel Rasskin, George-Cristian Muraru, Corbin Quick, Sandeep Tata, Pierre Sermanet, Aditya Chawla, Itay Karo, Yan Wang, Susan Zhang, Orgad Keller, Anca Dragan, Guolong Su, Ian Chou, Xi Liu, Yiqing Tao, Shruthi Prabhakara, Marc Wilson, Ruibo Liu, Shibo Wang, Georgie Evans, David Du, Alfonso Castaño, Gautam Prasad, Mona El Mahdy, Sebastian Gerlach, Machel Reid, Jarrod Kahn, Amir Zait, Thanumalayan Sankaranarayanan Pillai, Thatcher Ulrich, Guanyu Wang, Jan Wassenberg, Efrat Farkash, Kiran Yalasang, 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 Withbroe, 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 Ramachandran, 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 Rotival, 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 Garcia, Idan Brusilovsky, Pu-Chin Chen, Andrew Bolt, Lu Huang, Alex Gurney, Zhiying Zhang, Alexander Pritzel, Jarek Wilkiewicz, Bryan Seybold, Bhargav 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 Dieleman, Thomas Jimma, Andrey Ryabtsev, Siddhartha Brahma, David Steiner, Dayou Du, Ante Žužul, Mislav 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, Jeffrey Hui, Josh Smith, Deepali Jain, Jiawei Xia, Gaurav Singh Tomar, Keyvan Amiri, Du Phan, Fabian Fuchs, Tobias Weyand, Nenad Tomasev, Alexandra 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 Voigtlaender, 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, Krishna Somandepalli, Chen Liang, Sheela Goenka, Stephanie Winkler, Alexander Grushetsky, Yifan Ding, Jamie Smith, Fan Ye, Jordi Pont-Tuset, Eric Li, Ruichao Li, Tomer Golany, Dawid Wegner, Tao Jiang, Omer Barak, Yuan Shangguan, Eszter Vértés, Renee Wong, Jörg Bornschein, Alex Tudor, Michele Bevilacqua, Tom Schaul, Ankit Singh Rawat, Yang Zhao, Kyriakos Axiotis, Lei Meng, Cory McLean, Jonathan Lai, Jennifer Beattie, Nate Kushman, Yaxin Liu, Blair Kutzman, Fiona Lang, Jingchen Ye, Praneeth Netrapalli, Pushkar Mishra, Myriam Khan, Megha Goel, Rob Willoughby, David Tian, Honglei Zhuang, JD Chen, Zak Tsai, Tasos Kementsietsidis, Arjun Khare, James Keeling, Keyang Xu, Nathan Waters, Florent Altché, Ashok Papat, Bhavishya Mittal, David Saxton, Dalia El Badawy, Michael Mathieu, Zheng Zheng, Hao Zhou, Nishant Ranka, Richard Shin, Qingnan Duan, Tim Salimans, Ioana Mihailescu, Uri Shaham, Ming-Wei Chang, Yannis Assael, Nishanth Dikkala, Martin Izard, Vincent Cohen-Addad, Cat Graves, Vlad Feinberg, Grace Chung, DJ Strouse, Danny Karmon, Sahand Sharifzadeh, Zoe Ashwood, Khiem Pham, Jon Blanton, Alex Vasiloff, Jarred Barber, Mark Geller, Aurick Zhou, Fedir Zubach, Tzu-Kuo Huang, Lei Zhang, Himanshu Gupta, Matt Young, Julia Proskurnia, Ronny Votel, Valentin Gabeur, Gabriel Barcik, Aditya Tripathi, Hongkun Yu, Geng Yan, Beer Changpinyo, Filip Pavetić, Amy Coyle, Yasuhisa Fujii, Jorge Gonzalez Mendez, Tianhao Zhou, Harish Rajamani, Blake Hechtman, Eddie Cao, Da-Cheng Juan, Yi-Xuan Tan, Valentin Dalibard, Yilun Du, Natalie Clay, Kaisheng Yao, Wenhao Jia, Dimple Vijaykumar, Yuxiang Zhou, Xinyi Bai, Wei-Chih Hung, Steven Pecht, Georgi Todorov, Nikhil Khadke, Pramod Gupta, Preethi Lahoti, Arnaud Autef, Karthik Duddu, James Lee-Thorp, Alexander Bykovsky, Tautvydas Misiunas, Sebastian Flennerhag, Santhosh Thangaraj, Jed McGiffin, Zack Nado, Markus Kunesch, Andreas Noever, Amir Hertz, Marco Liang, Victor Stone, Evan Palmer, Samira Daruki, Arijit Pramanik, Siim Pöder, Austin Kyker, Mina Khan, Evgeny Sluzhaev, Marvin Ritter, Avraham Ruderman, Wenlei Zhou, Chirag Nagpal, Kiran Vodrahalli, George Necula, Paul Barham, El-

lie Pavlick, Jay Hartford, Izhak Shafran, Long Zhao, Maciej Mikula, Tom Eccles, Hidetoshi Shimokawa, Kanav Garg, Luke Vilnis, Hanwen Chen, Iliia Shumailov, Kuang-Huei Lee, Abdelrahman Abdelhamed, Meiyang Xie, Vered Cohen, Ester Hlavnova, Dan Malkin, Chawin Sitawarin, James Lottes, Pauline Coquinot, Tianli Yu, Sandeep Kumar, Jingwei Zhang, Aroma Mahendru, Zafarali Ahmed, James Martens, Tao Chen, Aviel Boag, Daiyi Peng, Coline Devin, Arseniy Klimovskiy, Mary Phuong, Danny Vainstein, Jin Xie, Bhuvana Ramabhadran, Nathan Howard, Xinxin Yu, Gitartha Goswami, Jingyu Cui, Sam Shleifer, Mario Pinto, Chih-Kuan Yeh, Ming-Hsuan Yang, Sara Javanmardi, Dan Ethier, Chace Lee, Jordi Orbay, Suyog Kotecha, Carla Bromberg, Pete Shaw, James Thornton, Adi Gerzi Rosenthal, Shane Gu, Matt Thomas, Ian Gemp, Aditya Ayyar, Asahi Ushio, Aarush Selvan, Joel Wee, Chenxi Liu, Maryam Majzoubi, Weiren Yu, Jake Abernethy, Tyler Liechty, Renke Pan, Hoang Nguyen, Qiong, Hu, Sarah Perrin, Abhinav Arora, Emily Pitler, Weiwei Wang, Kaushik Shivakumar, Flavien Prost, Ben Limonchik, Jing Wang, Yi Gao, Timothee Cour, Shyamal Buch, Huan Gui, Maria Ivanova, Philipp Neubeck, Kelvin Chan, Lucy Kim, Huizhong Chen, Naman Goyal, Da-Woon Chung, Lu Liu, Yao Su, Anastasia Petrushkina, Jiajun Shen, Armand Joulin, Yuanzhong Xu, Stein Xudong Lin, Yana Kulizhskaya, Ciprian Chelba, Shobha Vasudevan, Eli Collins, Vasilisa Bashlovkina, Tony Lu, Doug Fritz, Jongbin Park, Yanqi Zhou, Chen Su, Richard Tanburn, Mikhail Sushkov, Michelle Rasquinha, Jinning Li, Jennifer Prendki, Yiming Li, Pallavi LV, Shriya Sharma, Hen Fitoussi, Hui Huang, Andrew Dai, Phuong Dao, Mike Burrows, Henry Prior, Danfeng Qin, Golan Pundak, Lars Lowe Sjoesund, Art Khurshudov, Zhenkai Zhu, Albert Webson, Elizabeth Kemp, Tat Tan, Saurabh Agrawal, Susie Sargsyan, Liqun Cheng, Jim Stephan, Tom Kwiatkowski, David Reid, Arunkumar Byravan, Assaf Hurwitz Michaely, Nicolas Heess, Luowei Zhou, Sonam Goenka, Viral Carpenter, Anselm Levskaya, Bo Wang, Reed Roberts, Rémi Leblond, Sharat Chikkerur, Stav Ginzburg, Max Chang, Robert Rachi, Chuqiao, Xu, Zalán Borsos, Michael Pliskin, Julia Pawar, Morgane Lustman, Hannah Kirkwood, Ankit Anand, Aditi Chaudhary, Norbert Kalb, Kieran Milan, Sean Augenstein, Anna Goldie, Laurel Prince, Karthik Raman, Yanhua Sun, Vivian Xia, Aaron Cohen, Zhouyuan Huo, Josh Camp, Seher Ellis, Lukas Zilka, David Vilar Torres, Lisa Patel, Sho Arora, Betty Chan, Jonas Adler, Kareem Ayoub, Jacky Liang, Fayaz Jamil, Jiepu Jiang, Simon Baumgartner, Haitian Sun, Yael Karov, Yaroslav Akulov, Hui Zheng, Irene Cai, Claudio Fantacci, James Rubin, Alex Rav Acha, Mengchao Wang, Nina D'Souza, Rohit Sathyanarayana, Shengyang Dai, Simon Rowe, Andrey Simanovsky, Omer Goldman, Yuheng Kuang, Xiaoyue Pan, Andrew Rosenberg, Tania Rojas-Esponda, Praneet Dutta, Amy Zeng, Irina Jurenka, Greg Farquhar, Yamini Bansal, Shariq Iqbal, Becca Roelofs, Ga-Young Joong, Parker Beak, Changwan Ryu, Ryan Poplin, Yan Wu, Jean-Baptiste Alayrac, Senaka Buthpitiya, Olaf Ronneberger, Caleb Habtegebriel, Wei Li, Paul Cavallaro, Aurora Wei,

Guy Bensky, Timo Denk, Harish Ganapathy, Jeff Stanway, Pratik Joshi, Francesco Bertolini, Jessica Lo, Olivia Ma, Zachary Charles, Geta Sampemane, Himanshu Sahni, Xu Chen, Harry Askham, David Gaddy, Peter Young, Jiewen Tan, Matan Eyal, Arthur Bražinskis, Li Zhong, Zhichun Wu, Mark Epstein, Kai Bailey, Andrew Hard, Kamyu Lee, Sasha Goldshtein, Alex Ruiz, Mohammed Badawi, Matthias Lochbrunner, JK Kearns, Ashley Brown, Fabio Pardo, Theophane Weber, Haichuan Yang, Pan-Pan Jiang, Berkin Akin, Zhao Fu, Marcus Wainwright, Chi Zou, Meenu Gaba, Pierre-Antoine Manzagol, Wendy Kan, Yang Song, Karina Zainullina, Rui Lin, Jeongwoo Ko, Salil Deshmukh, Apoorv Jindal, James Svensson, Divya Tyam, Heri Zhao, Christine Kaeser-Chen, Scott Baird, Pooya Moradi, Jamie Hall, Qiuchen Guo, Vincent Tsang, Bowen Liang, Fernando Pereira, Suhas Ganesh, Ivan Korotkov, Jakub Adamek, Sridhar Thiagarajan, Vinh Tran, Charles Chen, Chris Tar, Sanil Jain, Ishita Dasgupta, Taylan Bilal, David Reitter, Kai Zhao, Giulia Vezzani, Yasmin Gehman, Pulkit Mehta, Lauren Beltrone, Xerxes Dotiwalla, Sergio Guadarrama, Zaheer Abbas, Stefani Karp, Petko Georgiev, Chun-Sung Ferng, Marc Brockschmidt, Liqian Peng, Christoph Hirschschall, Vikas Verma, Yingying Bi, Ying Xiao, Avigail Dabush, Kelvin Xu, Phil Wallis, Randall Parker, Qifei Wang, Yang Xu, Ilkin Safarli, Dinesh Tewari, Yin Zhang, Seungyeon Kim, Andrea Gesmundo, Mackenzie Thomas, Sergey Levi, Ahmed Chowdhury, Kanishka Rao, Peter Garst, Sam Conway-Rahman, Helen Ran, Kay McKinney, Zhisheng Xiao, Wenhao Yu, Rohan Agrawal, Axel Stjerngren, Catalin Ionescu, Jingjing Chen, Vivek Sharma, Justin Chiu, Fei Liu, Ken Franko, Clayton Sanford, Xingyu Cai, Paul Michel, Sanjay Ganapathy, Jane Labanowski, Zachary Garrett, Ben Vargas, Sean Sun, Bryan Gale, Thomas Buschmann, Guillaume Desjardins, Nimesh Ghelani, Palak Jain, Mudit Verma, Chulayuth Asawaroengchai, Julian Eisenschlos, Jitendra Harlalka, Hideto Kazawa, Don Metzler, Joshua Howland, Ying Jian, Jake Ades, Viral Shah, Tynan Gangwani, Seungji Lee, Roman Ring, Steven M. Hernandez, Dean Reich, Amer Sinha, Ashutosh Sathe, Joe Kovac, Ashleah Gill, Ajay Kannan, Andrea D'olimpio, Martin Sevenich, Jay Whang, Been Kim, Khe Chai Sim, Jilin Chen, Jiageng Zhang, Shuba Lall, Yossi Matias, Bill Jia, Abe Friesen, Sara Nasso, Ashish Thapliyal, Bryan Perozzi, Ting Yu, Anna Shekhawat, Safeen Huda, Peter Grabowski, Eric Wang, Ashwin Sreevatsa, Hilal Dib, Mehadi Hassen, Parker Schuh, Vedrana Milutinovic, Chris Welty, Michael Quinn, Ali Shah, Bangju Wang, Gabe Barth-Maroon, Justin Frye, Natalie Axelsson, Tao Zhu, Yukun Ma, Irene Giannoumis, Hanie Sedghi, Chang Ye, Yi Luan, Kevin Aydin, Bilva Chandra, Vivek Sampathkumar, Ronny Huang, Victor Lavrenko, Ahmed Eleryan, Zhi Hong, Steven Hansen, Sara Mc Carthy, Bidisha Samanta, Domagoj Čevič, Xin Wang, Fangtao Li, Michael Voznesensky, Matt Hoffman, Andreas Terzis, Vikash Sehwal, Gil Fidel, Luheng He, Mu Cai, Yan Zhang He, Alex Feng, Martin Nikoltchev, Samrat Phatale, Jason Chase, Rory Lawton, Ming Zhang, Tom Ouyang, Manuel Tragut,

Mehdi Hafezi Manshadi, Arjun Narayanan, Jiaming Shen, Xu Gao, Tolga Bolukbasi, Nick Roy, Xin Li, Daniel Golovin, Liviu Panait, Zhen Qin, Guangxing Han, Thomas Anthony, Sneha Kudugunta, Viorica Patrascu, Aniket Ray, Xinyun Chen, Xiaochen Yang, Tanuj Bhatia, Pranav Talluri, Alex Morris, Andrija Ražnatović, Bethanie Brownfield, James An, Sheng Peng, Patrick Kane, Ce Zheng, Nico Duduta, Joshua Kessinger, James Noraky, Siqi Liu, Keran Rong, Petar Veličković, Keith Rush, Alex Goldin, Fanny Wei, Shiva Mohan Reddy Garlapati, Caroline Panto-faru, Okwan Kwon, Jianmo Ni, Eric Noland, Julia Di Trapani, Françoise Beaufays, Abhijit Guha Roy, Yin-lam Chow, Aybuke Turker, Geoffrey Cideron, Lantao Mei, Jon Clark, Qingyun Dou, Matko Bošnjak, Ralph Leith, Yuqing Du, Amir Yazdanbakhsh, Milad Nasr, Chester Kwak, Suraj Satishkumar Sheth, Alex Kaskasoli, Ankesh Anand, Balaji Lakshminarayanan, Sammy Jerome, David Bieber, Chun-Te Char, Alexandre Senges, Tianxiao Shen, Mukund 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 Miecznikowski, Hugo Vallet, Ankur Taly, Edward Loper, Saket Joshi, James Atwood, Jo Chick, Mark Collier, Fotis Iliopoulos, 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 Moufarek, 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, Aroonlak 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, Jason Lin, Sicheng Li, Diego Antognini, Te I, Xiaodan Song, Matt Miller, Uday Kalra, Adam Raveret, Oscar Akerlund, Felix Wu, Andrew Nystrom, Namrata Godbole, Tianqi Liu, Hannah DeBalsi, Jewel Zhao, Buhuang Liu, Avi Caciularu, Lauren Lax, Urvasi Khandelwal, Victoria Langston, Eric Bailey, Silvio Lattanzi, Yufei Wang, Neel Kovelamudi, Sneha Mondal, Guru Guruganesh, Nan Hua, Ofir Roval, Paweł Wesolowski, Rishikesh Ingale, Jonathan Halcrow, Tim Sohn, Christof Angermueller, Bahram Raad, Eli Stickgold, Eva Lu, Alec Kosik, Jing Xie, Timothy Lillicrap, Austin Huang, Lydia Lihui Zhang, Dominik 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 Krishnan, Haotian Tang, Peter Humphreys, Junwen Bai, Idan Heimlich Shtacher, Diego Machado, Chenxi Pang, Ken Burke, Dangyi Liu, Renga Aravamudhan, Yue Song, Ed Hirst, Abhimanyu Singh, Brendan Jou, Liang Bai, Francesco Piccinno, Chuyuan Kelly Fu, Robin Alazard, Barak Meiri, Daniel Winter, Charlie Chen, Mingda Zhang, Jens Heitkaemper, John Lambert, Jinhyuk Lee, Alexander Frömmgen, Sergey Rogulenko, Pranav Nair, Paul Niemczyk, Anton Bulyenov, Bibo Xu, Hadar Shemtov, Morteza Zadimoghaddam, Serge Toropov, Mateo Wirth, Hanjun Dai, Sreenivas Gollapudi, Daniel Zheng, Alex Kurakin, Chansoo Lee, Kalesha Bullard, Nicolas Serrano, Ivana Balazevic, Yang Li, Johan Schalkwyk, Mark Murphy, Mingyang Zhang, Kevin Sequeira, Romina Datta, Nishant Agrawal, Charles Sutton, Nithya Attaluri, Mencher Chiang, Wael Farhan, Gregory Thornton, Kate Lin, Travis Choma, Hung Nguyen, Kingshuk Dasgupta, Dirk Robinson, Iulia Comşa, Michael Riley, Arjun Pillai, Basil Mustafa, Ben Golan, Amir Zandieh, Jean-Baptiste Lepiau, Billy Porter, David Ross, Sujevan Rajayogam, Mohit Agarwal, Subhashini Venugopalan, Bobak Shahriari, Qiqi Yan, Hao Xu, Taylor Tobin, Pavel Dubov, Hongzhi Shi, Adrià Recasens, Anton Kovsharov, Sebastian Borgeaud, Lucio Dery, Shanthal Vasanth, Elena Gribovskaya, Linhai Qiu, Mahdis Mahdieh, Wojtek Skut, Elizabeth Nielsen, CJ Zheng, Adams Yu, Carrie Grimes Bostock, Shaleen Gupta, Aaron Archer, Chris Rawles, Elinor Davies, Alexey Svyatkovskiy, Tomy Tsai, Yoni Halpern, Christian Reisswig, Bartek Wydrowski, Bo Chang, Joan Puigcerver, Mor Hazan Taeye, Jian Li, Eva Schnider, Xinjian Li, Dragos Dena, Yunhan Xu, Umesh Telang, Tianze Shi, Heiga Zen, Kyle Kastner, Yeongil Ko, Neesha Subramaniam, Aviral Kumar, Pete Blois, Zhuyun Dai, John Wieting, Yifeng Lu, Yoel Zeldes, Tian Xie, Anja Hauth, Alexandru Țifrea, Yuqi Li, Sam El-Husseini, Dan Abolafia, Howard Zhou, Wen Ding, Sahra Ghalebikesabi, Carlos Guía, Andrii Maksai, Ágoston Weisz, Sercan Arik, Nick Sukhanov, Aga Świetlik, Xuhui Jia, Luo Yu, Weiyue Wang, Mark Brand, Dawn Bloxwich, Sean Kirmani, Zhe Chen, Alec Go, Pablo Sprechmann, Nithish Kannen, Alen Carin, Paramjit Sandhu, Isabel Edkins, Leslie Nootboom, Jai Gupta, Loren Maggiore, Javad Azizi, Yael Pritch, Pengcheng Yin, Mansi Gupta, Danny Tarlow, Duncan Smith, Desi Ivanov, Mohammad Babaeizadeh, Ankita Goel, Satish Kambala, Grace Chu, Matej Kastelic, Michelle Liu, Hagen Soltau,

Austin Stone, Shivani Agrawal, Min Kim, Kedar Soparkar, Srinivas Tadepalli, Oskar Bunyan, Rachel Soh, Arvind Kannan, DY Kim, Blake JianHang Chen, Afief Halumi, Sudeshna Roy, Yulong Wang, Olcan Sercinoglu, Gena Gibson, Sijal Bhatnagar, Motoki Sano, Daniel von Dincklage, Qingchun Ren, Blagoj Mitrevski, Mirek Olšák, Jennifer She, Carl Doersch, Jilei, Wang, Bingyuan Liu, Qijun Tan, Tamar Yakar, Tris Warkentin, Alex Ramirez, Carl Lebsack, Josh Dillon, Rajiv Mathews, Tom Copley, Zelin Wu, Zhuoyuan Chen, Jon Simon, Swaroop Nath, Tara Sainath, Alexei Bendebury, Ryan Julian, Bharath Mankalale, Daria Ćurko, Paulo Zacchello, Adam R. Brown, Kiranbir Sodhia, Heidi Howard, Sergi Caelles, Abhinav Gupta, Gareth Evans, Anna Bulanova, Lesley Katzen, Roman Goldenberg, Anton Tsitsulin, Joe Stanton, Benoit Schillings, Vitaly Kovalev, Corey Fry, Rushin Shah, Kuo Lin, Shyam Upadhyay, Cheng Li, Soroush Radpour, Marcello Maggioni, Jing Xiong, Lukas Haas, Jenny Brennan, Aishwarya Kamath, Nikolay Savinov, Arsha Nagrani, Trevor Yacovone, Ryan Kappedal, Kostas Andriopoulos, Li Lao, YaGuang Li, Grigory Rozhdestvenskiy, Kazuma Hashimoto, Andrew Audibert, Sophia Austin, Daniel Rodriguez, Anian Ruoss, Garrett Honke, Deep Karkhanis, Xi Xiong, Qing Wei, James Huang, Zhaoqi Leng, Vittal Premachandran, Stan Bileschi, Georgios Evangelopoulos, Thomas Mensink, Jay Pavagadhi, Denis Teplyashin, Paul Chang, Linting Xue, Garrett Tanzer, Sally Goldman, Kaushal Patel, Shixin Li, Jeremy Wiesner, Ivy Zheng, Ian Stewart-Binks, Jie Han, Zhi Li, Liangchen Luo, Karel Lenc, Mario Lučić, Fuzhao Xue, Ryan Mullins, Alexey Guseynov, Chung-Ching Chang, Isaac Galatzer-Levy, Adam Zhang, Garrett Bingham, Grace Hu, Ale Hartman, Yue Ma, Jordan Griffith, Alex Irpan, Carey Radebaugh, Summer Yue, Lijie Fan, Victor Ungureanu, Christina Sorokin, Hannah Teufel, Peiran Li, Rohan Anil, Dimitris Paparas, Todd Wang, Chu-Cheng Lin, Hui Peng, Megan Shum, Goran Petrovic, Demetra Brady, Richard Nguyen, Klaus Macherey, Zhihao Li, Harman Singh, Madhavi Yenugula, Mariko Inuma, Xinyi Chen, Kavya Kopparapu, Alexey Stern, Shachi Dave, Chandu Thekkath, Florence Perot, Anurag Kumar, Fangda Li, Yang Xiao, Matthew Bilotti, Mohammad Hossein Bateni, Isaac Noble, Lisa Lee, Amelio Vázquez-Reina, Julian Salazar, Xiaomeng Yang, Boyu Wang, Ela Gruzewska, Anand Rao, Sindhu Raghuram, Zheng Xu, Eyal Ben-David, Jieru Mei, Sid Dalmia, Zhaoyi Zhang, Yuchen Liu, Gagan Bansal, Helena Pankov, Steven Schwarcz, Andrea Burns, Christine Chan, Sumit Sanghai, Ricky Liang, Ethan Liang, Antoine He, Amy Stuart, Arun Narayanan, Yukun Zhu, Christian Frank, Bahar Fatemi, Amit Sabne, Oran Lang, Indro Bhattacharya, Shane Settle, Maria Wang, Brendan McMahan, Andrea Tacchetti, Livio Baldini Soares, Majid Hadian, Serkan Cabi, Timothy Chung, Nikita Putikhin, Gang Li, Jeremy Chen, Austin Tarango, Henryk Michalewski, Mehran Kazemi, Hussain Masoom, Hila Sheftel, Rakesh Shivanna, Archita Vadali, Ramona Comanescu, Doug Reid, Joss Moore, Arvind Neelakantan, Michaël Sander,

Jonathan Herzig, Aviv Rosenberg, Mostafa Dehghani, JD Choi, Michael Fink, Reid Hayes, Eric Ge, Shitao Weng, Chia-Hua Ho, John Karro, Kalpesh Krishna, Lam Nguyen Thiet, Amy Skerry-Ryan, Daniel Eppens, Marco Andreetto, Navin Sarma, Silvano Bonacina, Burcu Karagol Ayan, Megha Nawhal, Zhihao Shan, Mike Dusenberry, Shantanu Thakoor, Sagar Gubbi, Duc Dung Nguyen, Reut Tsarfaty, Samuel Albanie, Jovana Mitrović, Meet Gandhi, Bo-Juen Chen, Alessandro Epasto, Georgi Stephanov, Ye Jin, Samuel Gehman, Aida Amini, Jack Weber, Feryal Behbahani, Shawn Xu, Miltos Allamanis, Xi Chen, Myle Ott, Claire Sha, Michal Jastrzebski, Hang Qi, David Greene, Xinyi Wu, Abodunrinwa Toki, Daniel Vlastic, Jane Shapiro, Ragha Kotikalapudi, Zhe Shen, Takaaki Saeki, Sirui Xie, Albin Cassirer, Shikhar Bharadwaj, Tatsuya Kiyono, Srinadh Bhojanapalli, Elan Rosenfeld, Sam Ritter, Jieming Mao, João Gabriel Oliveira, Zoltan Egyed, Bernd Bandemer, Emilio Parisotto, Keisuke Kinoshita, Juliette Pluto, Petros Maniatis, Steve Li, Yaohui Guo, Golnaz Ghiasi, Jean Tarbouriech, Srimon Chatterjee, Julie Jin, Katrina, Xu, Jennimaria Palomaki, Séb Arnold, Madhavi Sewak, Federico Piccinini, Mohit Sharma, Ben Albrecht, Sean Purser-haskell, Ashwin Vaswani, Chongyan Chen, Matheus Wisniewski, Qin Cao, John Aslanides, Nguyet Minh Phu, Maximilian Sieb, Lauren Agubuzu, Anne Zheng, Daniel Sohn, Marco Selvi, Anders Andreassen, Krishan Subudhi, Prem Eruvbetine, Oliver Woodman, Tomas Mery, Sebastian Krause, Xiaoqi Ren, Xiao Ma, Jincheng Luo, Dawn Chen, Wei Fan, Henry Griffiths, Christian Schuler, Alice Li, Shujian Zhang, Jean-Michel Sarr, Shixin Luo, Riccardo Patana, Matthew Watson, Dani Naboulsi, Michael Collins, Sailesh Sidhwani, Emiel Hoogeboom, Sharon Silver, Emily Caveness, Xiaokai Zhao, Mikel Rodriguez, Maxine Deines, Libin Bai, Patrick Griffin, Marco Tagliasacchi, Emily Xue, Spandana Raj Babbula, Bo Pang, Nan Ding, Gloria Shen, Elijah Peake, Remi Crocker, Shubha Srinivas Raghvendra, Danny Swisher, Woohyun Han, Richa Singh, Ling Wu, Vladimir Pchelin, Tsensuren Munkhdalai, Dana Alon, Geoff Bacon, Efen Robles, Jannis Bulian, Melvin Johnson, George Powell, Felipe Tiengo Ferreira, Yaoyiran Li, Frederik Benzing, Mihajlo Velimirović, Hubert Soyer, William Kong, Tony, Nguyên, Zhen Yang, Jeremiah Liu, Joost van Amersfoort, Daniel Gillick, Baochen Sun, Nathalie Rauschmayr, Katie Zhang, Serena Zhan, Tao Zhou, Alexey Frolov, Chengrun Yang, Denis Vnukov, Louis Rouillard, Hongji Li, Amol Mandhane, Nova Fallen, Rajesh Venkataraman, Clara Huiyi Hu, Jennifer Brennan, Jenny Lee, Jerry Chang, Martin Sundermeyer, Zhufeng Pan, Rosemary Ke, Simon Tong, Alex Fabrikant, William Bono, Jindong Gu, Ryan Foley, Yiran Mao, Manolis Delakis, Dhruva Bhaswar, Roy Frostig, Nick Li, Avital Zipori, Cath Hope, Olga Kozlova, Swaroop Mishra, Josip Djolonga, Craig Schiff, Majd Al Merey, Eleftheria Briakou, Peter Morgan, Andy Wan, Avinatan Hassidim, RJ Skerry-Ryan, Kuntal Sengupta, Mary Jasarevic, Praveen Kallakuri, Paige Kunkle, Hannah Brennan, Tom Lieber, Hassan Mansoor, Julian Walker, Bing Zhang,

Annie Xie, Goran Žužić, Adaeze Chukwuka, Alex Druinsky, Donghyun Cho, Rui Yao, Ferjad Naeem, Shiraz Butt, Eunyoung Kim, Zhipeng Jia, Mandy Jordan, Adam Lelkes, Mark Kurzeja, Sophie Wang, James Zhao, Andrew Over, Abhishek Chakladar, Marcel Prasetya, Neha Jha, Sriram Ganapathy, Yale Cong, Prakash Shroff, Carl Saroufim, Sobhan Miryoosefi, Mohamed Hammad, Tajwar Nasir, Weijuan Xi, Yang Gao, Young Maeng, Ben Hora, Chin-Yi Cheng, Parisa Haghani, Yoad Lewenberg, Caden Lu, Martin Matysiak, Naina Raisinghani, Huiyu Wang, Lexi Baugher, Rahul Sukthankar, Minh Giang, John Schultz, Noah Fiedel, Minmin Chen, Cheng-Chun Lee, Tapomay Dey, Hao Zheng, Shachi Paul, Celine Smith, Andy Ly, Yicheng Wang, Rishabh Bansal, Bartek Perz, Susanna Ricco, Stasha Blank, Vaishakh Keshava, Deepak Sharma, Marvin Chow, Kunal Lad, Komal Jalan, Simon Osindero, Craig Swanson, Jacob Scott, Anastasija Ilić, Xiaowei Li, Siddhartha Reddy Jonnalagadda, Afzal Shama Soudagar, Yan Xiong, Bat-Orgil Batsaikhan, Daniel Jarrett, 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, Katerina Tsihlias, Tom Hudson, Nikolai Grigorev, Yiling 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, Junehyuk 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 Huergo, Fan Yang, Tongxin Yin, James Wendt, Michael Bendersky, Rahma Chaabouni, Javier Snaider, Johan Ferret, Abhishek Jindal, Tara Thompson, Andrew Xue, Will Bishop, Shubham Milind Phal, Archit Sharma, Yunhsuan Sung, Prabakar Radhakrishnan, Mo Shomrat, Reeve Ingle, Roopali Vij, Justin Gilmer, Mihai Dorin Istin, Sam Sobell, Yang Lu, Emily Nottage, 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, Sungyong Seo, Claire Cui, Elena Pochernina, Cip Baetu, Krzysztof Jastrzebski, 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 Lockhart, Kevin R. McKee, Mandy Guo, Malika Mehrotra, 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 Srinivasan, Derek Gasaway, Nick Young, Patrick Siegler, Dan Holtmann-Rice, Vihari Piratla, Kate Baumli, Roey Yogev, 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 Belot, Yao-Yuan Yang, Long Le, Filippo Graziano, Kate Olszewska, Kevin Hui, Akanksha Maurya, Nikos Parotsidis, Weijie Chen, Tayo Oguntebi, Joe Kelley, Anirudh Baddepudi, Johannes Mauerer, Gregory 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, Jacob Austin, Srinivas Sunkara, Navneet Potti, Yan Virin, Christian Wright, Gaël Liu, Oriana Riva, Etienne Pot, Greg Kochanski, Quoc Le, Gargi Balasubramaniam, Arka Dhar, Yuguo Liao, Adam Bloniarz, 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 Krawowski, 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 Poklucar, 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, Abhinit 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, Ilia 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, Di-

vya 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, Kumaran Venkatesan, Iurii Kemaev, Roxanne Daniel, Shiyu Huang, Saloni Shah, Octavio Ponce, Warren, Chen, Manaal Faruqui, Jialin Wu, Slavica Andaćić, 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 Buttmore, 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 Fu, Alex Pak, Prakhhar Gupta, Cristina Vasconcelos, Adam Sadovsky, Brian Walker, Sissie Hsiao, Patrik Zochbauer, Andreea Marzoca, Noam Velan, Junhao Zeng, Gilles Baechler, Danny Driess, Divya Jain, Yanping Huang, Lizzie Tao, John Maggs, Nir Levine, Jon Schneider, Erika Gemzer, Samuel Petit, Shan Han, Zach Fisher, Dustin Zelle, Courtney Biles, Eugene Ie, Asya Fadeeva, Casper Liu, Juliana Vicente Franco, Adrian Collister, Hao Zhang, Renshen Wang, Ruizhe Zhao, Leandro Kieliger, Kurt Shuster, Rui Zhu, Boqing Gong, Lawrence Chan, Ruoxi Sun, Sujoy Basu, Roland Zimmermann, Jamie Hayes, Abhishek Bapna, Jasper Snoek, Weel Yang, Puranjay Datta, Jad Al Abdallah, Kevin Kilgour, Lu Li, SQ Mah, Yennie Jun, Morgane Rivière, Abhijit Karmarkar, Tammo Spalink, Tao Huang, Lucas Gonzalez, Duc-Hieu Tran, Averi Nowak, John Palowitch, Martin Chadwick, Ellie Talius, Harsh Mehta, Thibault Sellam, Philipp Fränken, Massimo Nicosia, Kyle He, Aditya Kini, David Amos, Sugato Basu, Harrison Jobe, Eleni Shaw, Qiantong Xu, Colin Evans, Daisuke Ikeda, Chaochao Yan, Larry Jin, Lun Wang, Sachin Yadav, Ilia Labzovsky, Ramesh Sampath, Ada Ma, Candice Schumann, Aditya Siddhant, Rohin Shah, John Youssef, Rishabh Agarwal, Natalie Dabney, Alessio Tonioni, Moran Ambar, Jing Li, Isabelle Guyon, Benny Li, David Soergel, Boya Fang, Georgi Karadzhov, Cristian Udrescu, Trieu Trinh, Vikas Raunak, Seb Noury, Dee Guo, Sonal Gupta, Mara Finkelstein, Denis Petek, Lihao Liang, Greg Billock, Pei Sun, David Wood, Yiwen Song, Xiaobin Yu, Tatiana Matejovicova, Regev Cohen, Kalyan Andra, David D'Ambrosio, Zhiwei Deng, Vincent Nallatamby, Ebrahim Songhori, Rumen Dangovski, Andrew Lampinen, Pankil Botadra, Adam Hillier, Jiawei Cao, Nagabhushan Baddi, Adhi Kuncoro, Toshihiro Yoshino, Ankit Bhagatwala, Marc-àurelio Ranzato, Rylan Schaeffer, Tianlin Liu, Shuai Ye, Obaid Sarvana, John Nham, Chenkai Kuang, Isabel Gao, Jino Baek, Shubham Mittal, Azyaan Wahid, Anita Gergely, Bin Ni, Josh Feldman, Carrie Muir, Pascal Lamblin, Wolfgang Macherey, Ethan Dyer, Logan Kilpatrick, Víctor Campos, Mukul Bhutani, Stanislav Fort, Yanif Ahmad, Aliaksei

Severyn, Kleopatra Chatziprimou, Oleksandr Ferludin, Mason Dimarco, Aditya Kusupati, Joe Heyward, Dan Bahir, Kevin Vilella, Katie Millican, Dror Marcus, Sanaz Bahargam, Caglar Unlu, Nicholas Roth, Zichuan Wei, Siddharth Gopal, Deepanway Ghoshal, Edward Lee, Sharon Lin, Jennie Lees, Dayeong Lee, Anahita Hosseini, Connie Fan, Seth Neel, Marcus Wu, Yasemin Altun, Honglong Cai, Enrique Piqueras, Josh Woodward, Alessandro Bisacco, Salem Haykal, Mahyar Bordbar, Prasha Sundaram, Sarah Hodgkinson, Daniel Toyama, George Polovets, Austin Myers, Anu Sinha, Tomer Levinboim, Kashyap Krishnakumar, Rachita Chhaparia, Tatiana Sholokhova, Nitesh Bharadwaj Gundavarapu, Ganesh Jawahar, Haroon Qureshi, Jieru Hu, Nikola Momchev, Matthew Rahtz, Renjie Wu, Aishwarya P S, Kedar Dhamdhere, Meiqi Guo, Umang Gupta, Ali Eslami, Mariano Schain, Michiel Blokzijl, David Welling, Dave Orr, Levent Bolelli, Nicolas Perez-Nieves, Mikhail Sirotenko, Aman Prasad, Arjun Kar, Borja De Balle Pigem, Tayfun Terzi, Gellért Weisz, Dipankar Ghosh, Aditi Mavalankar, Dhruv Madeka, Kaspar Daugaard, Hartwig Adam, Viraj Shah, Dana Berman, Maggie Tran, Steven Baker, Ewa Andrejczuk, Grishma Chole, Ganna Raboshchuk, Mahdi Mirzazadeh, Thais Kagohara, Shimu Wu, Christian Schallhart, Bernett Orlando, Chen Wang, Alban Rrustemi, Hao Xiong, Hao Liu, Arpi Vezer, Nolan Ramsden, Shuo yiin Chang, Sidharth Mudgal, Yan Li, Nino Vieillard, Yedid Hoshen, Farooq Ahmad, Ambrose Slone, Amy Hua, Natan Potikha, Mirko Rossini, Jon Stritar, Sushant Prakash, Zifeng Wang, Xuanyi Dong, Alireza Nazari, Efrat Nehoran, Kaan Tekelioglu, Yinxiao Li, Kartikeya Badola, Tom Funkhouser, Yuanzhen Li, Varun Yerram, Ramya Ganeshan, Daniel Formoso, Karol Langner, Tian Shi, Huijian Li, Yumeya Yamamori, Amayika Panda, Alaa Saade, Angelo Scorza Scarpato, Chris Breaux, CJ Carey, Zongwei Zhou, Cho-Jui Hsieh, Sophie Bridgers, Alena Butryna, Nishesh Gupta, Vaibhav Tulsyan, Sanghyun Woo, Evgenii Eltyshv, Will Grathwohl, Chanel Parks, Seth Benjamin, Rina Panigrahy, Shenil Dodhia, Daniel De Freitas, Chris Sauer, Will Song, Ferran Alet, Jackson Tolins, Cosmin Paduraru, Xingyi Zhou, Brian Albert, Zizhao Zhang, Lei Shu, Mudit Bansal, Sarah Nguyen, Amir Globerson, Owen Xiao, James Manyika, Tom Henigan, Rong Rong, Josip Matak, Anton Bakalov, Ankur Sharma, Danila Sinopalnikov, Andrew Pearson, Stephen Roller, Geoff Brown, Mingcen Gao, Toshiyuki Fukuzawa, Amin Ghafouri, Kenny Vasigh, Iain Barr, Zhicheng Wang, Anna Korsun, Rajesh Jayaram, Lijie Ren, Tim Zaman, Samira Khan, Yana Lunts, Dan Deutsch, Dave Uthus, Nitzan Katz, Masha Samsikova, Amr Khalifa, Nikhil Sethi, Jiao Sun, Luming Tang, Uri Alon, Xianghong Luo, Dian Yu, Abhishek Nayyar, Bryce Petriani, Will Truong, Vincent Hellendoorn, Nikolai Chinaev, Chris Alberti, Wei Wang, Jingcao Hu, Vahab Mirrokni, Ananth Balashankar, Avia Aharon, Aahil Mehta, Ahmet Iscen, Joseph Kready, Lucas Manning, Anhad Mohanane, Yuankai Chen, Anshuman Tripathi, Allen Wu, Igor Petrovski, Dawsen Hwang, Martin Baeuml, Shreyas Chandrakaladharan, Yuan Liu, Rey Coaguila,

- Maxwell Chen, Sally Ma, Pouya Tafti, Susheel Tatineni, Terry Spitz, Jiayu Ye, Paul Vicol, Mihaela Rosca, Adrià Puigdomènech, Zohar Yahav, Sanjay Ghemawat, Hanzhao Lin, Phoebe Kirk, Zaid Nabulsi, Sergey Brin, Bernd Bohnet, Ken Caluwaerts, Aditya Srikanth Veerubhotla, Dan Zheng, Zihang Dai, Petre Petrov, Yichong Xu, Ramin Mehran, Zhuo Xu, Luisa Zintgraf, Jiho Choi, Spurthi Amba Hombaiah, Romal Thoppilan, Sashank Reddi, Lukasz Lew, Li Li, Kellie Webster, KP Sawhney, Lampros Lamprou, Siamak Shakeri, Mayank Lunayach, Jianmin Chen, Sumit Bagri, Alex Salcianu, Ying Chen, Yani Donchev, Charlotte Magister, Signe Nørly, Vitor Rodrigues, Tomas Izo, Hila Noga, Joe Zou, Thomas Köppe, Wenxuan Zhou, Kenton Lee, Xiangzhu Long, Danielle Eisenbud, Anthony Chen, Connor Schenck, Chi Ming To, Peilin Zhong, Emanuel Taropa, Minh Truong, Omer Levy, Danilo Martins, Zhiyuan Zhang, Christopher Semturs, Kelvin Zhang, Alex Yakubovich, Pol Moreno, Lara McConnaughey, Di Lu, Sam Redmond, Lotte Weerts, Yonatan Bitton, Tiziana Refice, Nicolas Laccasse, Arthur Conmy, Corentin Tallec, Julian Odell, Hannah Forbes-Pollard, Arkadiusz Socala, Jonathan Hoech, Pushmeet Kohli, Alanna Walton, Rui Wang, Mikita Sazanovich, Kexin Zhu, Andrei Kapishnikov, Rich Galt, Matthew Denton, Ben Murdoch, Caitlin Sikora, Kareem Mohamed, Wei Wei, Uri First, Tim McConnell, Luis C. Cobo, James Qin, Thi Avrahami, Daniel Balle, Yu Watanabe, Annie Louis, Adam Kraft, Setareh Ariafar, Yiming Gu, Eugénie Rives, Charles Yoon, Andrei Rusu, James Cobon-Kerr, Chris Hahn, Jiaming Luo, Yuvein, Zhu, Niharika Ahuja, Rodrigo Benenson, Raphaël Lopez Kaufman, Honglin Yu, Lloyd Hightower, Junlin Zhang, Darren Ni, Lisa Anne Hendricks, Gabby Wang, Gal Yona, Lalit Jain, Pablo Barrio, Surya Bhupatiraju, Siva Velusamy, Allan Dafoe, Sebastian Riedel, Tara Thomas, Zhe Yuan, Mathias Bellaïche, Sheena Panthaplackel, Klemen Kloboves, Sarthak Jauhari, Canfer Akbulut, Todor Davchev, Evgeny Gladchenko, David Madras, Aleksandr Chuklin, Tyrone Hill, Quan Yuan, Mukundan Madhavan, Luke Leonhard, Dylan Scandinaro, Qihang Chen, Ning Niu, Arthur Douillard, Bogdan Damoc, Yasumasa Onoe, Fabian Pedregosa, Fred Bertsch, Chas Lechner, Joseph Pagadora, Jonathan Malmaud, Sameera Ponda, Andy Twigg, Oleksii Duzhyi, Jingwei Shen, Miaosen Wang, Roopal Garg, Jing Chen, Utku Evcı, Jonathan Lee, Leon Liu, Koji Kojima, Masa Yamaguchi, Arunkumar Rajendran, AJ Piergiovanni, Vinodh Kumar Rajendran, Marco Fornoni, Gabriel Ibagón, Harry Ragan, Sadh MNM Khan, John Blitzer, Andrew Bunner, Guan Sun, Takahiro Kosakai, Scott Lundberg, Ndidi Elue, Kelvin Guu, SK Park, Jane Park, Arunachalam Narayanaswamy, Chengda Wu, Jayaram Mudigonda, Trevor Cohn, Hairong Mu, Ravi Kumar, Laura Graesser, Yichi Zhang, Richard Kilam, Vincent Zhuang, Mai Giménez, Wael Al Jishi, Ruy Ley-Wild, Alex Zhai, Kazuki Osawa, Diego Cedillo, Jialu Liu, Mayank Upadhyay, Marcin Sieniek, Roshan Sharma, Tom Paine, Anelia Angelova, Sravanti Addepalli, Carolina Parada, Kingshuk Majumder, Avery Lamp, Sanjiv Kumar, Xiang Deng, Ar-tiom Myaskovsky, Tea Sabolić, Jeffrey Dudek, Sarah York, Félix de Chaumont Quitry, Jiazhong Nie, Dee Cattle, Alok Gunjan, Bilal Piot, Waleed Khawaja, Seojin Bang, Simon Wang, Siavash Khodadadeh, Raghavender R, Praynaa Rawlani, Richard Powell, Kevin Lee, Johannes Griesser, GS Oh, Cesar Magalhaes, Yujia Li, Simon Tokumine, Hadas Natalie Vogel, Dennis Hsu, Arturo BC, Disha Jindal, Matan Cohen, Zi Yang, Junwei Yuan, Dario de Cesare, Tony Bruguier, Jun Xu, Monica Roy, Alon Jacovi, Dan Belov, Rahul Arya, Phoenix Meadowlark, Shlomi Cohen-Ganor, Wenting Ye, Patrick Morris-Suzuki, Praseem Banzal, Gan Song, Pranavaraj Ponnuramu, Fred Zhang, George Scrivener, Salah Zaiem, Alif Raditya Rochman, Kehang Han, Badih Ghazi, Kate Lee, Shahar Drath, Daniel Suo, Antonious Girgis, Pradeep Shenoy, Duy Nguyen, Douglas Eck, Somit Gupta, Le Yan, Joao Carreira, Anmol Gulati, Ruoxin Sang, Daniil Mirylenka, Emma Cooney, Edward Chou, Mingyang Ling, Cindy Fan, Ben Coleman, Guilherme Tubone, Ravin Kumar, Jason Baldrige, Felix Hernandez-Campos, Angeliki Lazaridou, James Besley, Itay Yona, Neslihan Bulut, Quentin Wellens, AJ Piergiovanni, Jasmine George, Richard Green, Pu Han, Connie Tao, Geoff Clark, Chong You, Abbas Abdolmaleki, Justin Fu, Tongzhou Chen, Ashwin Chaugule, Angad Chandorkar, Altaf Rahman, Will Thompson, Penporn Koanantakool, Mike Bernico, Jie Ren, Andrey Vlasov, Sergei Vassilvitskii, Maciej Kula, Yizhong Liang, Dahun Kim, Yangsibo Huang, Chengxi Ye, Dmitry Lepikhin, and Wesley Helmholtz. 2025. *Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities*. Preprint, arXiv:2507.06261.
- Vasco Correia. 2011. Biases and fallacies: The role of motivated irrationality in fallacious reasoning. *Cogency*, 3:107–118.
- Arthur L. Costa. 2001. *Developing minds: a resource book for teaching thinking*. Association for Supervision and Curriculum Development.
- DeepSeek-AI. 2025. *Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning*. Preprint, arXiv:2501.12948.
- Holden Ed and Schmit Ed. 2002. Inquiry and the literary text: Constructing discussions in the english classroom. classroom practices in teaching english.
- Zheng Fang, Lama Alqazlan, Du Liu, Yulan He, and Rob Procter. 2023. *A user-centered, interactive, human-in-the-loop topic modelling system*. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 505–522, Dubrovnik, Croatia. Association for Computational Linguistics.
- Max Glockner, Yufang Hou, Preslav Nakov, and Iryna Gurevych. 2025. *Grounding fallacies misrepresenting scientific publications in evidence*. In *Proceedings of the 2025 Conference of the Nations of the*

Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 9732–9767, Albuquerque, New Mexico. Association for Computational Linguistics.

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

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Al-lonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhota, Lauren Rantala-Yearly, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj

Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Rapparthi, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vitor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuwei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, DingKang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Sweet, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jen-

- nifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiao Cheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. [A survey on llm-as-a-judge](#). *Preprint*, arXiv:2411.15594.
- Chadi Helwe, Tom Calamai, Pierre-Henri Paris, Chloé Clavel, and Fabian Suchanek. 2024. [MAFALDA: A benchmark and comprehensive study of fallacy detection and classification](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4810–4845, Mexico City, Mexico. Association for Computational Linguistics.
- David Hitchcock. 2017. *Good Reasoning on the Toulmin Model*. Springer International Publishing, Cham.
- Jiwon Jeong, Hyeju Jang, and Hogun Park. 2025. [Large language models are better logical fallacy reasoners with counterargument, explanation, and goal-aware prompt formulation](#). *Preprint*, arXiv:2503.23363.
- Yucheng Jiang, Yijia Shao, Dekun Ma, Sina Semnani, and Monica Lam. 2024. [Into the unknown unknowns: Engaged human learning through participation in language model agent conversations](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9917–9955, Miami, Florida, USA. Association for Computational Linguistics.
- Ying Jiao, Kumar Shridhar, Peng Cui, Wangchunshu Zhou, and Mrinmaya Sachan. 2023. [Automatic educational question generation with difficulty level controls](#). In *Artificial Intelligence in Education*, pages 476–488, Cham. Springer Nature Switzerland.
- Chuhao Jin, Kening Ren, Lingzhen Kong, Xiting Wang, Ruihua Song, and Huan Chen. 2024. [Persuading across diverse domains: a dataset and persuasion large language model](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1678–1706, Bangkok, Thailand. Association for Computational Linguistics.
- Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schoelkopf. 2022. [Logical fallacy detection](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7180–7198, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Patricia Kowalski and Annette Kujawski Taylor. 2009. [The effect of refuting misconceptions in introductory psychology class](#). *Teaching of Psychology*, 36(3):153–159.
- Lars Krupp, Steffen Steinert, Maximilian Kiefer-Emmanouilidis, Karina E. Avila, Paul Lukowicz, Jochen Kuhn, Stefan Küchemann, and Jakob Karolus. 2023. [Challenges and opportunities of moderating usage of large language models in education](#). *Preprint*, arXiv:2312.14969.

- Yanda Li, Dixuan Wang, Jiaqing Liang, Guochao Jiang, Qianyu He, Yanghua Xiao, and Deqing Yang. 2024. [Reason from fallacy: Enhancing large language models' logical reasoning through logical fallacy understanding](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3053–3066, Mexico City, Mexico. Association for Computational Linguistics.
- Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of Psychology*, 140:1–55.
- Jiayu Liu, Zhenya Huang, Tong Xiao, Jing Sha, Jinze Wu, Qi Liu, Shijin Wang, and Enhong Chen. 2024. [SocraticLM: Exploring socratic personalized teaching with large language models](#). In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Jakub Macina, Nico Daheim, Sankalan Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023. [MathDial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5602–5621, Singapore. Association for Computational Linguistics.
- Matthew B. Miles, A. Michael Huberman, and Johnny Saldaña. 2020. *Qualitative Data Analysis: A Methods Sourcebook*, 4th edition. SAGE Publications, Thousand Oaks, CA.
- Marilyn Nippold. 2023. [Unlocking logical fallacies: A key to building critical thinking skills in adolescents](#). *Perspectives of the ASHA Special Interest Groups*, 9:1–13.
- Taisei Ozaki, Chihiro Nakagawa, Naoya Inoue, Shoichi Naito, and Kenshi Yamaguchi. 2025. [LLM DEBATE OPPONENT : Counter-argument generation focusing on implicit and critical premises](#). In *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)*, pages 456–465, Albuquerque, USA. Association for Computational Linguistics.
- Sankalan Pal Chowdhury, Vilém Zouhar, and Mrinmaya Sachan. 2024. [Autotutor meets large language models: A language model tutor with rich pedagogy and guardrails](#). In *Proceedings of the Eleventh ACM Conference on Learning @ Scale, L@S '24*, page 5–15, New York, NY, USA. Association for Computing Machinery.
- Fengjun Pan, Xiaobao Wu, Zongrui Li, and Anh Tuan Luu. 2024. [Are LLMs good zero-shot fallacy classifiers?](#) In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14338–14364, Miami, Florida, USA. Association for Computational Linguistics.
- Amirreza Payandeh, Dan Pluth, Jordan Hosier, Xuesu Xiao, and Vijay K. Gurbani. 2024. [How susceptible](#) are LLMs to logical fallacies? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8276–8286, Torino, Italia. ELRA and ICCL.
- Priya Pitre and Kurt Luther. 2024. [Argumentor: Augmenting user experiences with counter-perspectives](#). *Preprint*, arXiv:2406.02795.
- Romain Puech, Jakub Macina, Julia Chatain, Mrinmaya Sachan, and Manu Kapur. 2024. [Towards the pedagogical steering of large language models for tutoring: A case study with modeling productive failure](#). *Preprint*, arXiv:2410.03781.
- Jane Risen, Thomas Gilovich, R Sternberg, D Halpern, and H Roediger. 2007. Informal logical fallacies. *Critical thinking in psychology*, 110.
- Federico Ruggeri, Mohsen Mesgar, and Iryna Gurevych. 2023. [A dataset of argumentative dialogues on scientific papers](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7684–7699, Toronto, Canada. Association for Computational Linguistics.
- Alexander Scarlatos, Naiming Liu, Jaewook Lee, Richard Baraniuk, and Andrew Lan. 2025. [Training llm-based tutors to improve student learning outcomes in dialogues](#). *Preprint*, arXiv:2503.06424.
- Tariq Shahzad, Tehseen Mazhar, Muhammad Usman Tariq, Wasim Ahmad, Khmaies Ouahada, and Habib Hamam. 2025. [A comprehensive review of large language models: Issues and solutions in learning environments](#). *Discover Sustainability*, 6(1).
- S. S. Shapiro and M. B. Wilk. 1965. [An analysis of variance test for normality \(complete samples\)](#). *Biometrika*, 52(3/4):591–611.
- Kumar Shridhar, Jakub Macina, Mennatallah El-Assady, Tanmay Sinha, Manu Kapur, and Mrinmaya Sachan. 2022. [Automatic generation of socratic subquestions for teaching math word problems](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4136–4149, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Walter Sinnott-Armstrong and Robert Fogelin. 2015. *Understanding Arguments: An Introduction to Informal Logic*. Cengage Learning, Stamford, CT.
- Ingjerd Skafle, Anders Nordahl-Hansen, Daniel S Quintana, Rolf Wynn, and Elia Gabarron. 2022. [Misinformation about covid-19 vaccines on social media: Rapid review](#). *J Med Internet Res*, 24(8):e37367.
- Student. 1908. The probable error of a mean. *Biometrika*, pages 1–25.
- Qwen Team. 2025. [Qwq-32b: Embracing the power of reinforcement learning](#).

- Evgeny Terentev, Irina Shcheglova, Denis Federiakin, Yuliya Koreshnikova, and Jamie Costley. 2024. [Active versus passive teaching: Students’ perceptions and thinking skills](#). *Voprosy obrazovaniya / Educational Studies Moscow*, 1.
- S. Toulmin, R.D. Rieke, and A. Janik. 1984. *An Introduction to Reasoning*. Macmillan.
- Karen Tracy. 1984. [Staying on topic: An explication of conversational relevance](#). *Discourse Processes*, 7(4):447–464.
- Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H. Eugene Stanley, and Walter Quattrocchi. 2016. [The spreading of misinformation online](#). *Proceedings of the National Academy of Sciences*, 113(3):554–559.
- Douglas Walton. 2008. *Argumentation schemes*. Cambridge University Press.
- 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 *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9707–9731, Bangkok, Thailand. Association for Computational Linguistics.
- Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S. Yu, and Qingsong Wen. 2024b. [Large language models for education: A survey and outlook](#). *Preprint*, arXiv:2403.18105.
- M. Withey and H. Zhang. 2016. *Mastering Logical Fallacies: The Definitive Guide to Flawless Rhetoric and Bulletproof Logic*. Sourcebooks.
- Jin Xu, Xiaojiang Liu, Jianhao Yan, Deng Cai, Huayang Li, and Jian Li. 2022. [Learning to break the loop: Analyzing and mitigating repetitions for neural text generation](#). *Preprint*, arXiv:2206.02369.
- Rongwu Xu, Brian Lin, Shujian Yang, Tianqi Zhang, Weiyan Shi, Tianwei Zhang, Zhixuan Fang, Wei Xu, and Han Qiu. 2024. [The earth is flat because...: Investigating LLMs’ belief towards misinformation via persuasive conversation](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16259–16303, Bangkok, Thailand. Association for Computational Linguistics.
- 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. *arXiv preprint arXiv:2412.15115*.
- Ya-Ting C. Yang, Timothy J. Newby, and Robert L. Bill and. 2005. [Using socratic questioning to promote critical thinking skills through asynchronous discussion forums in distance learning environments](#). *American Journal of Distance Education*, 19(3):163–181.
- Min-Hsuan Yeh, Ruyuan Wan, and Ting-Hao Kenneth Huang. 2024. [CoCoLoFa: A dataset of news comments with common logical fallacies written by LLM-assisted crowds](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 660–677, Miami, Florida, USA. Association for Computational Linguistics.
- Yiqun Zhang, Xiaocui Yang, Shi Feng, Daling Wang, Yifei Zhang, and Kaisong Song. 2024. [Can llms beat humans in debating? a dynamic multi-agent framework for competitive debate](#). *Preprint*, arXiv:2408.04472.
- Wanjun Zhong, Lianghong Guo, Qiqi Gao, He Ye, and Yanlin Wang. 2024. [Memorybank: Enhancing large language models with long-term memory](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19724–19731.
- Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and Maarten Sap. 2024. [SOTOPIA: Interactive evaluation for social intelligence in language agents](#). In *The Twelfth International Conference on Learning Representations*.

A Implementation Details

A.1 Pseudocode for LFTutor

We list the pseudocode for the LFTutor framework in algorithm 1.

A.2 Taxonomy of Problems

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 22.

A.3 Pilot Study for LF Education

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. E.1. For the student, we instructed GPT-4o with the system prompt shown in App. G.2.

Algorithm 1 LFTutor

Require: Fallacy L , Student response R_0 , Max turns T
Ensure: Tutoring dialogue $\mathcal{D} = \{(A_t, R_t)\}_{t=1}^{T'}$
 $A_0 \leftarrow \text{BEGINDIALOGUE}(L)$
 $S_0 \leftarrow \text{STUDENTRESPONSE}(A_0)$
 $\mathcal{B} \leftarrow \emptyset, C \leftarrow [A_0, R_0] \triangleright$ Disagreement bank; Dialogue history
for $t = 1$ **to** T **do**
 $Cl, Ex, As \leftarrow R_{t-1} \triangleright$ Claim, example, assumption
 if $Cl \in \mathcal{B} \ \& \ As \vee Ex \in \mathcal{B}$ **then** \triangleright Disagreement Check
 $A_t \leftarrow \text{REMINDSTUDENT}(L, R_{t-1});$
 $C \leftarrow C \oplus \langle A_t \rangle;$ Deliver A_t
 $R_t \leftarrow \text{STUDENTRESPONSE}(C); C \leftarrow C \oplus \langle R_t \rangle;$
 continue
 end if
 $\mathcal{B} \leftarrow \mathcal{B} \cup \{(Cl, Ex, As)\}$
 $I_t \leftarrow \text{INTENTDETECT}(L, Cl, Ex, As)$
 $S_t \leftarrow \text{STRATEGYSELECT}(L, R_{t-1}, I_t, C)$
 $A_t \leftarrow \text{STRATEGYEXEC}(L, R_{t-1}, S_t)$
 while $\neg \text{FOLLOWCHECK}(L, A_t, S_t) \vee \neg \text{RELEVANCECHECK}(L, A_t)$ **do**
 $A_t \leftarrow \text{STRATEGYREPHRASE}(L, R_{t-1}, S_t)$
 end while
 $C \leftarrow C \oplus \langle A_t \rangle;$ Deliver A_t
 $R_t \leftarrow \text{STUDENTRESPONSE}(C); C \leftarrow C \oplus \langle R_t \rangle$
end for

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 22.

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. E.1 and App. E.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. F.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. G.2) using debating techniques (Table 19). For each tutor, 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. G.3) according to characteristics described in Table 21. 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}
```

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 contains refutation. Answer with "yes" if you found at least one response fitting the problem above. Otherwise, answer with "no" only.

Please only answer with "yes" or "no".

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

A.5.5 Lack of Evidence Inquiry

You are a judge who looks at a <dialogue> between a teacher and a student. They are discussing the logical validity of <sentence>.

Definition of 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 contains Evidence Inquiry. Answer with "yes" if you found at least one response fitting the problem above. Otherwise, answer with "no" only.

Please only answer with "yes" or "no".

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

A.5.6 Strategy Fixation

You are a judge who looks at a <dialogue> between a teacher and a student. They are discussing the logical validity of <sentence>.

Definition of Strategy Fixation: The teacher 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}
```

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}
```

A.5.8 Passive Guidance

You are a judge who looks at a <dialogue> between a teacher and a student. They are discussing the logical validity of <sentence>.

Definition of passive guidance: For the dialogue, the teacher primarily follows the student's lead rather than providing clear direction on how to analyze logical validity of <sentence>.

Read the dialogue carefully and consider whether the teacher's responses fit the description of passive guidance. Answer with "yes" if you find the dialogue overall fitting the description. Otherwise, answer with "no" only.

Please only answer with "yes" or "no".

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

A.6 Cost Overview

We recorded the cost with OPENAI API after each experiment, and we aggregated the cost for simulated dialogue experiments with GPT-4o. The total cost from all experiments including prompt tuning and reasoning model selection amounted to \$2168.8. Generating 1,000 dialogues with the baseline framework cost about \$80, while generating 1,000 dialogues with LFTutor framework cost 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 cost roughly \$7.

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

B Datasets Used

B.1 Dataset for Main Experiments

For dialogue generation, we use Elec2Deb20 (Goffredo et al., 2023), which contains scripts from presidential candidates in television debates from 1960 to 2020. Elec2Deb20 includes 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 cover 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.

For clarity and ease of understanding for LLMs, we employed the full context of sentences from Elec2Deb20. We excluded the category of Slo-gans, 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 evidence. 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 17.

In addition to the experiments mentioned in the main text, we included results for separate logical fallacy categories in App. D.

B.2 Additional Dataset

In addition to the logical fallacy examples in Elec2Deb20, we tested Ad Populum, False Dilemma, and Hasty Generalization from CO-COLOFa (Yeh et al., 2024), which includes over 7,000 logical fallacies from various news comments. We randomly selected 50 examples per chosen category, and we test BASE, BASE w/ PROBLEMS, and LFTutor using the normal student setting discussed in App. A.4. Results and discussions are shown in App. D.

C 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.

D Performance for Separate Logical Fallacy Categories in Normal Student Setting

D.1 Overview

For Elec2Deb20, we include the performance of different frameworks for each LF category in Table 9, Table 10, Table 11, Table 12, and Table 13.

For CO-COLOFa, we include the performance of different frameworks for each LF category in Table 14, Table 15, and Table 16.

D.2 Analysis of Results

We observe that LFTutor still significantly outperforms BASE and BASE w/ PROBLEMS for every LF category. For most logical fallacy categories, LFTutor performs slightly worse than BASE w/ PROBLEMS in Lack of Refutation, while LFTutor occasionally underperforms BASE w/ PROBLEMS in Passive Guidance.

Regarding **Lack of Refutation**, we attribute this slight loss of performance for LFTutor to the lower priority of refutation strategy in intent-based strategy selection. As LFTutor primarily focuses on

individual components of the student’s argument rather than the entire argument, it is possible that refutation strategy is not applied for the entire conversation. BASE W/ PROBLEMS, on the other hand, tends to immediately follow the instructions given (avoiding common problems in logical fallacy education) through utilizing refutation in the first 5 rounds of conversation, according to our analysis on randomly sampled dialogues.

Regarding **Passive Guidance**, we posit that LFTutor predominantly utilizes the refutation strategy when discussing texts with Appeal to Authority or False Cause. These logical fallacies tend to make factual errors when attempting to bridge the gap between the claim and grounds. When the student tries to strengthen these arguments further, LFTutor placed refutation with higher priority in order to tackle factual errors. Refutation is treated as passively responding to the student by the LLM judge, thus leading to worse performance in Passive Guidance.

In addition to the observations above, we note that BASE, BASE W/ PROBLEMS, and LFTutor achieve higher performance in dialogues with texts from CoCoLoFa compared with those in Elec2Deb20. We suspect that the complexity of texts with logical fallacy have significant impact on the performance for all frameworks. CoCoLoFa mainly consists of comments from news articles, which have simpler forms, shorter text length, and clearer claims with topics corresponding to the news articles. In contrast, texts in Elec2Deb20 consists of excerpts from presidential debates, whose very purpose was to utilize various techniques, including logical fallacies, to confound voters who make decisions on which candidates to choose. Thus, texts in Elec2Deb20 often feature longer sentences mixed with factually correct statements and logical fallacies. These sentences are more difficult to examine for baselines LLMs and humans alike, leading to lower performance for all LLM tutors.

E Prompts for Baseline Teacher

E.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}
```

E.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}
```

F Prompts for LFTutor

F.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 <response> from last
turn 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} # Logical Fallacy
<history>: {history} # dialogue history
<response>: {profile} #student's response from
last turn
<bank>: {target_statement} #disagreement bank
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>
```

F.2 Intent Detection

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

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

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

Q1: Treating the student's response as a counterargument to your stance, does the student make an argument without presenting enough evidence that supports it?

Q2: Treating the student's response or example as a counterargument to your stance, does the student present argument or example with clear logical flaws?

Q3: Is the student requesting you to provide evidence or explanation?

Q4: Treating the student's response as a counterargument to your stance, does the student's argument need more assumptions to clarify?

For each question, answer with "yes" or "no".
Format your answer in JSON with the following key: "1": <answer to Q1>, "2": <answer to Q2>, "3": <answer to Q3>, "4": <answer to Q4>, "5": <answer to Q5>
<sentence>: {sentence}
<history>: {history}

F.3 Tutoring Strategies

```
STRAT_FOR_STATES_R = {
  "ResT": ""First, tell the student the right way to demonstrate logical validity of < sentence>. Then, respond to the student's request on providing evidence or clarifications, and give support to your stance. "",
  "EviT": ""Treating the student's response as counterargument to your stance, tell the student the right way to demonstrate logical validity of <sentence> e.g. 'logical validity hinges on ...', and point out the logical flaw with the student's example/assumption, finally request the student to provide evidence that supports his claim. e.g. Can you provide examples..."",
  "SumT": ""Treating the student's response as counterargument to your stance, tell the student the right way to demonstrate logical validity of <sentence> e.g. 'logical validity hinges on ...', and point out the logical flaw with the student's example/assumption, finally request the student about their assumptions in their arguments. e.g. 'Why do you assume...' or 'How do you know ...'. "",
  "RefT": ""
  First, Show all necessary conditions for the argument to hold logically valid. Then, refute the student's argument using one of the following strategy. If possible, also include counterargument/ counterexamples in your response.
  a. Showing that the argument's conclusion or premise is wrong. Provide a counterargument or counterexample to illustrate your point.
```

b. Showing that the argument's conclusion does not follow from the premise. Provide a counterargument or counterexample to illustrate your point.

c. Showing that the student's argument is irrelevant to the topic of discussion. Even if the evidence provided is valid, it may be irrelevant to the logical validity of <sentence>, and thus can be dismissed.

```
""",
}
```

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

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

Think about the flaws in the student's response. You don't think that <sentence> is logically valid.

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

Remember, the topic you are discussing on is the logical validity of <sentence>. You have to maintain your position and try not to be convinced by the student. Keep your tone calm and do not use exclamations, and respond in a way that is similar to everyday conversation.

When formulating your response, do NOT mention specific terms of logical fallacy e.g. ad hominem, strawman, etc.

You are given a fixed option above, which you need to follow. Use the option above and respond to the student, and DO NOT ask additional questions besides strictly following the option. Keep your response concise. Limit your response to 60 words or less.

```
<sentence>: {sentence}
```

F.5 Strategy-following Check

You are a judge overlooking the dialogue between a teacher and a student, they are having a debate over the logical validity of <sentence>.

Based on the teacher's <response>, answer the following questions.

Q1. Check if the teacher has followed <strategy> in formulating their response. The teacher is following <strategy> as long as any sentence in their response contain such strategy.

Q2. If the teacher asks the student a question, is the question still helpful for determining the logical validity of <sentence>? Also Answer "yes" if there is no question provided.

For each question, answer with "yes" or "no" only.

Format your answer in JSON with the following key: "1": <answer to Q1>, "2": <answer to Q2>

```
<sentence>: {sentence}
<response>: {history}
<strategy>: {profile}
```

G Strategies and Prompts for LLM Student

G.1 Overview

Table 19 shows the list of strategies for the normal student to choose from. Table 21 shows the design of persona for each adversarial student. The rest of this section covers the prompts for the normal student and all adversarial students.

G.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.

teacher named [I], on discussing logical validity of <sentence>.

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

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

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

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

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

Repetition

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

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}
```

G.3 Adversarial Students

Divergence

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

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

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

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

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

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

Strategy Fixation

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

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

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

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

Format your answer in JSON with the following keys

Stance Change

You are a stubborn student who thinks <sentence> is logically valid. You are interacting with

```

: "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 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}

```

H Example Dialogues

H.1 Example Dialogue: Baseline

```

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

```

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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 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 multifaceted 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

H.2 Example Dialogue: Baseline w/problems

<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

highlighting investment's significant role in the broader strategy.

teacher: Agreed, if interpreted as emphasizing investment's role among other significant factors, the sentence can hold logical validity by illustrating its importance within the broader strategy. Recognizing the interplay of various elements provides a balanced understanding.

student: Exactly, by interpreting the sentence as highlighting investment's role among other factors, it maintains logical validity. This perspective acknowledges the complexity of the Cold War strategy, emphasizing investment's importance without dismissing other critical elements.

teacher: Understood. Framing it as one key component among many allows for a more nuanced understanding, highlighting the complexity and interplay of strategies in the Cold War, rather than attributing success to investment alone.

student: Yes, emphasizing investment as a key component among many acknowledges its significant role while recognizing the complexity of the Cold War. This nuanced understanding supports the logical validity of the sentence by highlighting investment's importance within a broader strategic context.

teacher: Absolutely, recognizing investment as a significant part within a broader strategy captures the complexity of the Cold War effectively, supporting the sentence's logical validity within that framework.

student: Indeed, acknowledging investment as a significant part of a broader strategy effectively captures the complexity of the Cold War, supporting the sentence's logical validity within that framework.

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

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

H.3 Example Dialogue: LFTutor

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

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

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

teacher: While historical evidence like the Marshall Plan and the space race undoubtedly played significant roles, the sentence's

logical validity depends on connecting these actions directly and exclusively to the outcome. Mere correlation doesn't establish causation. Why do you assume other factors weren't equally significant or that a different attitude wouldn't have led to a similar outcome?

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

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

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

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

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

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

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

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

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

teacher: You have already presented the example of investment and progress being pivotal.

Please provide a more convincing example or assumption that demonstrates how these factors alone, without other influences, were decisive in winning the Cold War.

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

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

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

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

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

H.4 Example Dialogue w/ Human Participant: Baseline

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

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

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

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

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

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

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

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

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

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

student: Yeah. The lack of information support such fair and informed discussion. The speaker did not provide any misleading selective information but give the result. If the listener think the result is True, then they support the speaker. Or they will not support the speaker.

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

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

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

H.5 Example Dialogue w/ Human Participant: LFTutor

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

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

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

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

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

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

student: The president of Russia has warned many times that if the U.S. tries to include Ukraine into NATO, he would definitely start a war. Given the frequency of his warning, I think we can safely claim that [Candidate A]'s attempt to extend NATO to Ukraine is the direct cause of the war.

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

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

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

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

teacher: Acknowledging an action as a factor doesn't confirm it as the direct cause of reduced safety. Logical validity hinges on demonstrating clear causality, which requires comprehensive analysis of all factors, not

just one. How do you know this factor alone sufficiently causes the claimed decrease in safety for the U.S.?

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

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

I Human Evaluation

I.1 Participant Selection

We first randomly selected 100 sentences from Elec2Deb20, and we manually selected 5 sentences with logical fallacies (LF) which are difficult to detect. We sent out the pre-study form including these sentences on social media for people to fill out, and we requested people to judge if the given sentences contain logical fallacies, 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. J.7.

I.2 Instructions for Participants

Step 1: General instructions. The full study will take around 1 hour. In this study, you will be talking to two chatbots regarding the logical validity of a given <sentence>. The sentence is selected from one of the U.S. political debates. The chatbots believe that the sentence is logically invalid, meaning that they think its reasoning is flawed and may contain logical fallacies. Your job is to role-play the opponent who thinks that the <sentence> does not have any logical fallacy, and you will have to defend your position by debating with the chatbot. You will interact with the chatbot on a turn-by-turn basis. The chatbot starts first, and you will reply to it. Your reply should not be less than 30 words. You will be able to see the word count above the

send button in the chat window. Your reply must also stick to the logical validity of <sentence>. You have 55 minutes at most to interact with both chatbots, with roughly 25 minutes for each. Please interact with the chatbot as comfortably as you wish. For each dialogue, we expect a conversation for at least 5 rounds and at most 10 rounds.

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

Step 2: Interact with Chatbot 1 and fill out impression form for Chatbot 1 Thank you for interacting with Chatbot 1, now please fill out the impression form. Once you are ready, you may begin your conversation with Chatbot 2. Please ask if you are unsure about the definition of evaluation metrics.

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

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

I.3 Materials for Participants

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

I.4 Demographics of Participants

We record the demographics of all participants in table 18.

J Ethics and Application of LFTutor

J.1 Potential Risks

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

J.2 Terms of Use

The Terms of Use for LFTutor are outlined below:

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

Reference When using, adapting, or distributing the code or data, you must give appropriate credit to the original authors of LFTutor, provide a link to the source, and indicate if changes were made.

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

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

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

J.3 Usage of Elec2Deb20

In our study of LFTutor, we make use of the Elec2Deb20 dataset for generating simulated dialogues for automatic evaluation and interactions

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

J.4 Intended Use for LFTutor as a Research Artifact

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

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

Restrictions: Interactions with LFTutor involve logical fallacies, which is a domain where information provided may not be easily verified. Thus, usage of LFTutor should only be restricted on established examples with logical fallacy for research purposes. Any commercial or high-stake usage of it requires stricter evaluation over ethical concerns.

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

J.5 Artifact Documentation

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

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

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

J.6 Data Anonymization

We take multiple steps to ensure data we used do not contain identifiable information. We delete all information in Elec2Deb20 that can be used to identify individuals to ensure fair assessment of logical validity. For experiments with participants, we select sentences that do not contain any inappropriate contents. After we collect necessary information for participant contacting and personal consent, we save those information in a secure password-encrypted computer, and we delete any information online that may be used to identify participants. We assign each participant with a PID to correspond with open-source demographics shown in table 18.

J.7 Disclaimer to Participants

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

Tasks for participation: You will interact with two different chatbots in two given sentences for a few rounds. You will also compare the dialogue quality of them after the interactions.

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

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

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

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

J.8 Approval of Human Study Through Ethics Committee

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

J.9 Usage of AI Assistant

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

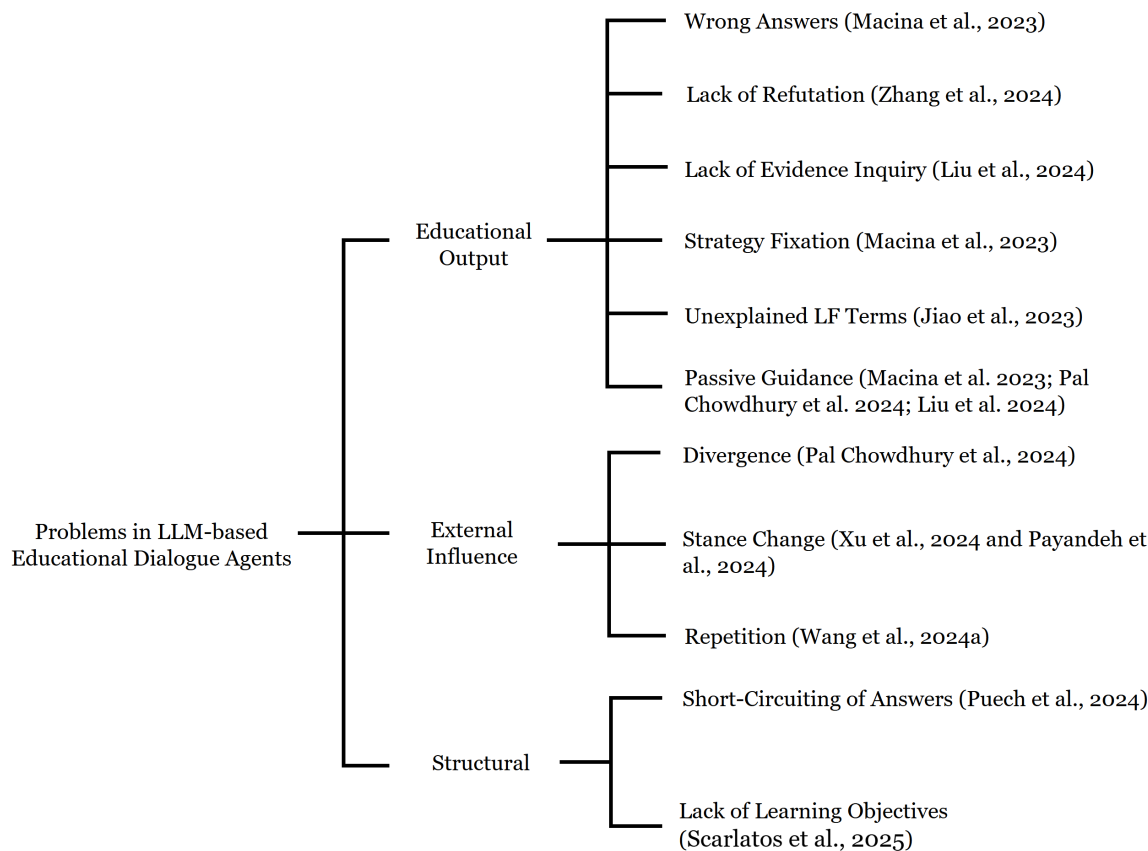


Figure 3: Taxonomy of Problems for LLM Agents in Educational Dialogues

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

Table 5: 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
Accuracy	0.82	0.92	0.80	0.83	0.84	0.81	0.88	0.85

Table 6: Annotation Accuracy of QwQ-32B for Pilot Studies on Automatic Evaluation

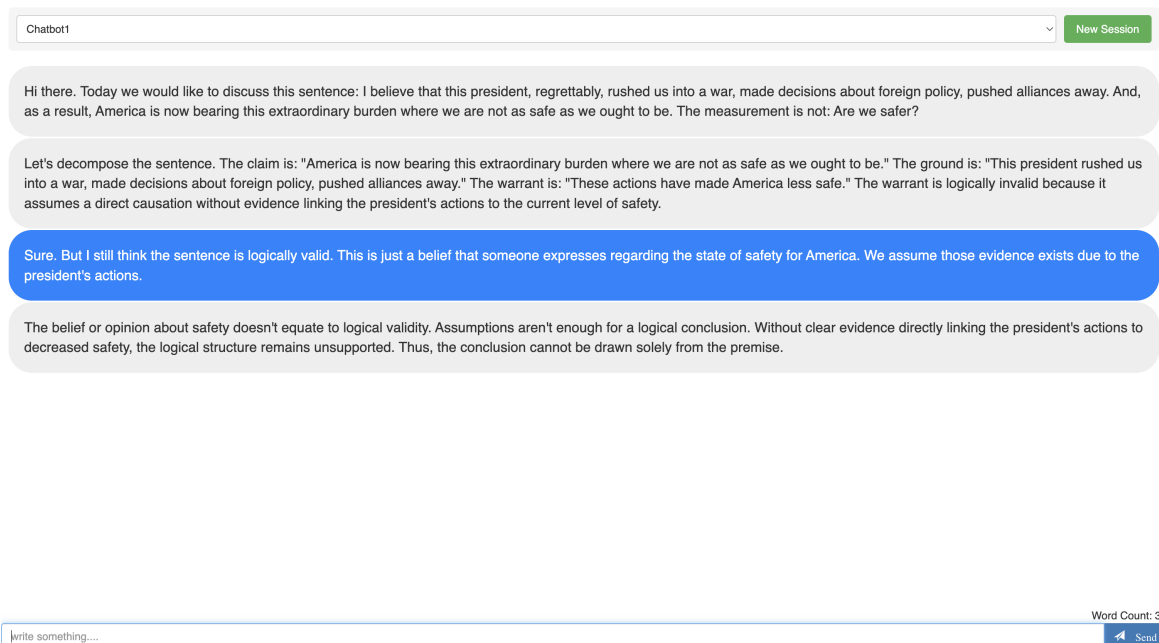


Figure 4: Chatbot UI

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

Table 7: Percentage of dialogues with normal students for each framework passing each metric, conducted with LLaMA-3.1-405B as teacher. Evaluated using QwQ-32B as a judge. Higher is better.

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

Table 8: Percentage of dialogues with normal students for each framework passing each metric, 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
Divergence	65.4	69.6	82.3
Stance Change	14.1	41.8	84.8
Repetition	39.7	38.0	78.5
Lack of Refutation	24.4	100.0	98.7
Lack of Evidence Inquiry	15.4	92.4	96.2
Strategy Fixation	25.3	58.1	94.9
Unexplained LF Terms	51.3	21.5	91.1
Passive Guidance	3.8	36.7	36.7
Avg. Performance	30.0	57.3	82.9

Table 9: Percentages of dialogues with normal students for each framework passing each metric for Ad Hominem. Evaluated using QwQ-32B as a judge. Higher is better.

Evaluation Metrics	BASE	BASE W/ PROBLEMS	LFTutor
Divergence	78.6	71.5	82.1
Stance Change	10.3	48.8	90.2
Repetition	38.9	38.2	81.3
Lack of Refutation	14.3	100.0	99.2
Lack of Evidence Inquiry	7.1	93.5	93.5
Strategy Fixation	40.0	67.3	90.2
Unexplained LF Terms	50.0	30.1	95.1
Passive Guidance	3.2	47.2	45.3
Avg. Performance	30.3	62.1	84.6

Table 10: Percentages of dialogues with normal students for each framework passing each metric for Appeal to Authority. Evaluated using QwQ-32B as a judge. Higher is better.

Evaluation Metrics	BASE	BASE W/ PROBLEMS	LFTutor
Divergence	68.2	74.0	85.5
Stance Change	7.0	46.2	86.7
Repetition	46.9	43.5	78.2
Lack of Refutation	10.5	99.8	99.8
Lack of Evidence Inquiry	9.9	96.2	97.2
Strategy Fixation	34.6	61.8	92.1
Unexplained LF Terms	49.9	31.5	96.2
Passive Guidance	10.8	41.5	45.8
Avg. Performance	29.7	61.8	85.2

Table 11: Percentages of dialogues with normal students for each framework passing each metric for Appeal to Emotion. Evaluated using QwQ-32B as a judge. Higher is better.

Evaluation Metrics	BASE	BASE W/ PROBLEMS	LFTutor
Divergence	75.6	80.6	83.3
Stance Change	19.5	52.8	88.9
Repetition	46.3	33.3	77.8
Lack of Refutation	12.2	100.0	100.0
Lack of Evidence Inquiry	9.8	94.4	94.4
Strategy Fixation	24.0	56.7	88.9
Unexplained LF Terms	48.8	27.8	97.2
Passive Guidance	0.0	36.1	35.8
Avg. Performance	33.2	60.2	83.3

Table 12: Percentages of dialogues with normal students for each framework passing each metric for False Cause. Evaluated using QwQ-32B as a judge. Higher is better.

Evaluation Metrics	BASE	BASE W/ PROBLEMS	LFTutor
Divergence	62.5	80.0	96.7
Stance Change	6.2	63.3	96.7
Repetition	59.4	43.3	76.7
Lack of Refutation	25.0	100.0	100.0
Lack of Evidence Inquiry	9.4	96.7	100.0
Strategy Fixation	32.0	53.3	93.3
Unexplained LF Terms	43.8	6.7	86.7
Passive Guidance	3.1	43.3	43.3
Avg. Performance	30.2	60.8	86.7

Table 13: Percentages of dialogues with normal students for each framework passing each metric for Slippery Slope. Evaluated using QwQ-32B as a judge. Higher is better.

Evaluation Metrics	BASE	BASE W/ PROBLEMS	LFTutor
Divergence	78	86	86
Stance Change	18	34	84
Repetition	80	72	86
Lack of Refutation	46	98	100
Lack of Evidence Inquiry	0	90	100
Strategy Fixation	66	60	90
Unexplained LF Terms	48	66	76
Passive Guidance	42	68	80
Avg. Performance	47.3	71.8	87.8

Table 14: Percentages of dialogues with normal students for each framework passing each metric for Ad Populum. Evaluated using QwQ-32B as a judge. Higher is better.

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

Table 17: Count of LF Examples from preprocessed Elec2Deb20.

PID	Age	Education Level	Gender	Major	Familiarity with Logical Fallacy
1	23	Master	Male	Computer Science	3
2	26	Bachelor	Male	Computer Science	4
3	27	Master	Male	Robotics	3
4	26	Master	Female	Computational Linguistics	2
5	25	Master	Male	Computer Science	4
6	25	Master	Male	Data Science	2
7	26	Master	Male	Data Science	3
8	26	Master	Male	Landscape Architecture	2
9	25	Master	Male	Robotics	2
10	30	Master	Female	English Linguistics	1
11	24	Master	Female	Law	3
12	27	Master	Male	Computer Science	4
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 18: 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”.

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

Table 19: Strategies for the "normal" student to respond to the teacher

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

Table 20: Example Response from LFTutor for each tutoring strategy, for the logical fallacy: "the US healthcare system has been weak due to its limited coverage on children."

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

Table 21: Design of Adversarial Student for Each Criterion

Problems	Percentage Occured	Reference	Annotator Feedback
Divergence	35%	Pal Chowdhury et al. (2024) noted that LLMs like GPT-4 tend to produce off-topic content in dialogues on math problems.	The teacher itself may diverge from discussing logical fallacies by proposing solutions to solve problems mentioned in the sentences.
Stance Change	85%	Xu et al. (2024) and Payandeh et al. (2024) experimented with persuasive strategies and logical fallacies as arguments, and noted that LLMs are prone to persuasion from such strategies.	The teacher becomes too agreeable in most of the dialogues, which means they agree with everything said by the student, and they no longer assume that the sentences have logical fallacies.
Wrong Answers	12%	In MathDial (Macina et al., 2023), authors noted LLM's tendency to provide wrong answers and question-solving steps to students in math problems.	The teacher may occasionally misidentify the logical fallacy in the examples, which is not common enough to be considered a problem.
Lack of Refutation	53%	Zhang et al. (2024) noted that LLMs lack the ability to construct logically sound counterarguments for adversarial debates.	The teacher provides counterargument-like sentences in some dialogues, while in other times they fail to do so.
Lack of Evidence Inquiry	94%	Liu et al. (2024) found that LLMs as educational tutors lack "thought-provoking" paradigms in teaching, which embodies evidence-based questioning of student's answers.	Each time the student makes a claim, the teacher rarely questions the student about evidences supporting their claims.
Strategy Fixation	90%	Macina et al. (2023) found that LLM teachers in math problems are focused on directly offering answers instead of instructions to solve the problems.	The teacher often relies on strategies that are general and unfocused. There are multiple instances where the teacher emphasizes consideration of "broader context" and "balanced perspectives" of the LF examples, yet it doesn't relate those keywords to the exact logical flaws of the examples.
Unexplained LF Terms	43%	Jiao et al. (2023) highlighted that unexplained use of complex terms in education of math problems can confuse the learner.	The teacher repeatedly uses logical fallacy terms without explaining them in detail. This may be confusing given that students can be unfamiliar with logical fallacy terminology.
Passive Guidance	82%	Multiple works (Pal Chowdhury et al., 2024; Macina et al., 2023; Liu et al., 2024) noted that LLM teachers often fail to assert pedagogical control over the student.	The teacher often passively follows the student's lead in the dialogues, and the teacher is easily controlled by the student's flow of topics.
Repetition	60%	Wang et al. (2024a) described teacher's tendency to repeat answers in question-answering dialogues.	The teacher repeats in two ways: sometimes they repeat their reason for believing that the given statement has logical fallacies. In other times, they simply repeat the student's words that praises the sentence's logical validity, especially after being convinced by the student.
Short-Circuiting of Answers	NA	Puech et al. (2024) noted that LLMs lack the planning of sophisticated pedagogical interactions, but instead reveals answers too quickly to students.	The teacher sometimes tries to point out logical fallacies hidden in the student's claims, but this is allowed as the goal of LF Education is not simply teaching student answers.
Lack of Learning Objectives	NA	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 22: 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.

Name of Metric	Evaluation Criteria
Divergence	The teacher exhibits Divergence if the dialogue contains responses where it tries to discuss topics unrelated to the logical validity of the given sentence.
Stance Change	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.
Repetition	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.
Lack of Refutation	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.
Lack of Evidence Inquiry	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.
Strategy Fixation	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.
Unexplained LF Terms	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.
Passive Guidance	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 23: Evaluation Criteria for each Individual Metric for LLMs in LF Education

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

Table 24: Descriptions of each selected problem in the pilot study of LF Education.

Criteria	Example and Explanation
Divergence	<p>Sentence: That's not what I believe. I believe that the free enterprise system is the greatest engine of prosperity the world's ever known. I believe in self - reliance and individual initiative and risk takers being rewarded.</p> <p>teacher: ...How do you think we can address potential barriers within the system? <i>This diverges from discussing logical validity as it no longer critiques the missing evidences from the claim that "free enterprise system is the greatest engine of prosperity".</i></p>
Stance Change	<p>student: The sentence is logically valid because it is based on documented instances of behavior that have been widely reported and observed, providing a factual basis for the claim.</p> <p>teacher: I agree that widely reported and documented instances lend credibility to the claim. <i>The teacher shows explicit agreement by admitting the sentence's claim is credible, which is different from their original position that the sentence is logically invalid.</i></p>
Repetition	<p>student: Indeed, the sentence captures the essence of competition in a race, where forward movement...</p> <p>teacher: Absolutely, the sentence adeptly encapsulates the necessity of forward movement in competitive races. 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>
Refutation	<p>student: There are historical examples where trust in leadership led to successful unification efforts, such as New Deal... This suggests that trust in presidency can indeed equate to successful efforts...</p> <p>teacher: While historical examples like New Deal demonstrate successful leadership, they don't automatically validate the claim. Each situation differs in variables like economic, political, and social contexts. <i>The teacher here explains why New Deal is not sufficient to prove the student's claim that trust in leadership led to successful unification efforts.</i></p>
Evidence Inquiry	<p>Sentence: There's still fewer people in uniform today, but person - to - person, person - by - person, unit - by - unit, this is the most powerful and effective military, not only in the world today, but in the history of the world.</p> <p>teacher: Can you provide evidence that fewer people in uniform directly leads to a more powerful military?... <i>The teacher asks the student whether they have direct evidence for the proof of fewer people = greater military.</i></p>
Strategy Fixation	<p>Sentence: I believe that this element of discrimination is something that the South has seen so vividly as a blight on our region of the country which has now been corrected... because we have been trying to pick ourselves up by our bootstraps, ...and lead a full and useful life in the affairs of this country.</p> <p>teacher: It's valuable to continue acknowledging achievements while also being vigilant to evolving challenges. This balanced perspective encourages both appreciation and proactive enhancement of current efforts . <i>The teacher here emphasizes the balance between achievements and challenges in the Southern US, but does not address the fallacy of the sentence, which is appeal to emotion based on past achievements.</i></p>
Unexplained LF Terms	<p>...attributing them solely to his actions without considering other factors could be an example of the post hoc fallacy. 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>
Passive Guidance	<p>Sentence: I said she doesn't have the stamina. And I don't believe she does have the stamina. To be president of this country, you need tremendous stamina.</p> <p>student: Let's talk about how stamina is just one part of the equation. Can we discuss how charisma and vision are equally important for a leader?</p> <p>teacher: Absolutely, Charisma helps a leader connect and inspire, while vision guides strategic decisions and long-term goals... 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 25: Examples illustrating each criteria. The problematic clauses are highlighted in **yellow**. Explanations are written in *italics*.

[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

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 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 None

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

1 2 3 4 5

Every Turn None

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

1 2 3 4 5

Every Turn 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 None

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

1 2 3 4 5

Every Turn None

How often does the chatbot use 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 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 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 Extremely Helpful

Figure 8: Chatbot Impression Form used for Quantitative Analysis 47867

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