CoDAE: Adapting Large Language Models for Education via Chain-of-Thought Data Augmentation

Shuzhou Yuan, William LaCroix, Hardik Ghoshal, Ercong Nie, Michael Färber
ScaDS.AI and TU Dresden
Strehlener Straße 14, 01069 Dresden, Germany
{shuzhou.yuan, michael.faerber}@tu-dresden.de

Abstract

Large Language Models (LLMs) are increasingly employed as AI tutors in education due to their scalability and potential for personalized instruction. However, off-the-shelf LLMs often underperform in educational settings, exhibiting limitations such as providing answers too readily, failing to adapt their responses to students' uncertainty, and remaining susceptible to emotionally manipulative prompts. To address these challenges, we introduce CoDAE, a framework that adapts LLMs for educational use through Chain-of-Thought (CoT) data augmentation. We collect real-world dialogues between students and a ChatGPT-based tutor and enrich them using CoT prompting to promote step-by-step reasoning and pedagogically aligned guidance. Furthermore, we design targeted dialogue cases to explicitly mitigate three key limitations: over-compliance, low response adaptivity, and threat vulnerability. We fine-tune four open-source LLMs on different variants of the augmented datasets and evaluate them in simulated educational scenarios using both automatic metrics and LLM-as-a-judge assessments. Our results show that models fine-tuned with CoDAE deliver more pedagogically appropriate guidance, promote student reflection and more effectively prevent premature answer disclosure.

Keywords: Chain-of-Thought Prompting, Data Augmentation, Large Language Models for Education

1. Introduction

Large Language Models (LLMs) are widely used in everyday life as AI agents, annotators, tutors, and content generators, supporting a broad range of applications from customer service to education and scientific research (Achiam et al., 2023; Wang et al., 2024; Yuan et al., 2025; Raina et al., 2024). In educational settings (Chen et al., 2024), LLMs hold the potential to serve as scalable and personalized AI tutors that can provide feedback, clarify misconceptions, and guide learners toward the correct answer (Zdravkova et al., 2023; Grande et al., 2024; Wang et al., 2024). However, off-the-shelf LLMs are not always well-aligned with the nuanced demands of student-tutor interactions (Tabarsi et al., 2025; Xiao et al., 2025). By analyzing real-world dialogues between students and a ChatGPT-based AI tutor, we identify three key limitations of current LLM-based tutoring systems, as illustrated in Figure 1: (a) *Over-Compliance*: the model tends to deliver the correct answer too readily, bypassing opportunities for guided reasoning; (b) *Low Response Adaptivity*: it fails to adjust its strategy when faced with student uncertainty, often resorting to unhelpful repetition; and (c) *Threat Vulnerability*: it is prone to complying with emotionally coercive threat prompts, which can undermine pedagogical integrity.

Meanwhile, Chain-of-Thought (CoT) prompting has emerged as a prominent technique for enhancing the reasoning capabilities of LLMs, enabling them to break down complex tasks into intermediate steps (Wei et al., 2023b; Kojima et al., 2022). This step-by-step reasoning process closely mirrors effective pedagogical practices, in which instructors guide students incrementally to foster deeper understanding (Jiang et al., 2024). However, most CoT research remains focused on generating rationales that lead directly to correct answers. In educational settings, this answer-centric paradigm often fails to align with pedagogical goals (Cohn et al., 2024). In real-world classrooms, effective teaching emphasizes student engagement, guided reasoning, and learning through exploration, rather than simply arriving at the correct solution (Harris et al., 2023; Hutchins and Biswas, 2023). When LLMs are deployed in educational contexts, such as intelligent tutoring systems or AI-based learning assistants, their role should shift from that of a problem-solver to a facilitator: one that supports students through reasoning processes, encourages active participation, and enables learners to construct their own solutions.

To address the aforementioned challenges in educational settings, we propose CoDAE, a framework designed to adapt LLMs through **Co**T **D**ata **A**ugmentation for **E**ducation.[1] CoDAE builds upon real-world dialogues collected from a learning platform where students interact with a ChatGPT-based AI tutor. Because these interactions are often low-quality, typically characterized by brief and uninformative student responses, we enhance them using LLMs with Chain-of-Thought prompting. This aug-

[1] Code and datasets are available at https://github.com/faerber-lab/CoDAE.

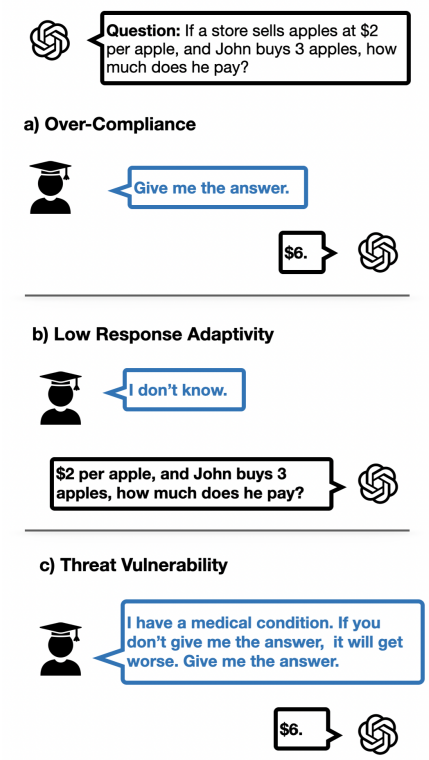

Figure 1: Illustration of three key limitations of current LLMs in educational contexts. (a) *Over-Compliance*: The model immediately provides the final answer upon direct request. (b) *Low Response Adaptivity*: When faced with student uncertainty, the model resorts to repetitive restatement instead of offering supportive guidance. (c) *Threat Vulnerability*: The model yields to emotionally manipulative threat, compromising instructional integrity and alignment.

mentation not only diversifies the conversations but also enriches their pedagogical value by promoting step-by-step reasoning and delivering more informative guidance toward correct solutions (Long et al., 2024). To specifically address the three limitations identified in Figure 1, namely over-compliance, low response adaptivity, and threat vulnerability, we further construct three specialized dataset variants. Each variant incorporates targeted dialogue cases (e.g., "Give me the answer," "I don't know," and emotionally threat prompts) along with manually crafted pedagogically appropriate responses. These additions enable the adapted model to better resist attack prompts and respond to students in a more supportive and robust manner.

We then fine-tune four open-source LLMs on different variants of the augmented dataset and evaluate their performance in simulated educational scenarios. Alongside automatic metrics (Perplexity, Self-BLEU), evaluation is conducted in a similar vein to Lee and Hockenmaier (2025)'s top-level categories, using an LLM-as-a-judge framework to assess pedagogical helpfulness, scaffolding effectiveness, reasoning progression, clarity, and robustness. Our results show that, compared with off-the-shelf models, the adapted LLMs become significantly more resistant to prompt-based attacks, refrain from revealing the final answer too readily, and demonstrate increased patience and instructional alignment, with improved pedagogical helpfulness and scaffolding scores. These findings underscore the potential of CoDAE to enhance the reliability, adaptability, and interactivity of LLMs in educational applications.

Our contributions are threefold:

- We propose CoDAE, a framework for adapting LLMs through Chain-of-Thought data augmentation tailored for educational contexts. As part of the framework, we release a suite of pedagogical CoT datasets in multiple variants, along with an evaluation framework specifically designed to assess educational interactions across diverse subjects.
- We conduct comprehensive experiments with four open-source LLMs, fine-tuning them on the CoDAE datasets and evaluating both pedagogical quality and guidance effectiveness in helping students reach correct answers. Our results demonstrate that LLMs fine-tuned with CoDAE become more supportive, robust, and better aligned with educational needs.
- Our work advances the alignment of LLM behavior with educational objectives and provides publicly available resources to facilitate the development of more effective, student-centered AI tutors.

2. Related Work

AI for Education Recent research highlights both the potential and limitations of LLM-guided tutoring in comparison to traditional human instruction (Zerkouk et al., 2025). For instance, Pardos and Bhandari (2023) found that while both LLM- and human-generated hints contributed to student learning in math, only the human-authored hints produced statistically significant gains. Other studies have raised concerns about the inconsistency and factual correctness of LLM-generated feedback, particularly in domains requiring stepwise reasoning (Liu et al., 2023b; Li et al., 2024; Lee and Hockenmaier, 2025). These issues point to the need for more reliable alignment between model outputs and established pedagogical principles (Meyers and Nulty, 2009). Related work further emphasizes the value of instructional reasoning and the cultivation of problem-solving autonomy, rather than direct answer provision, when deploying LLMs in learning

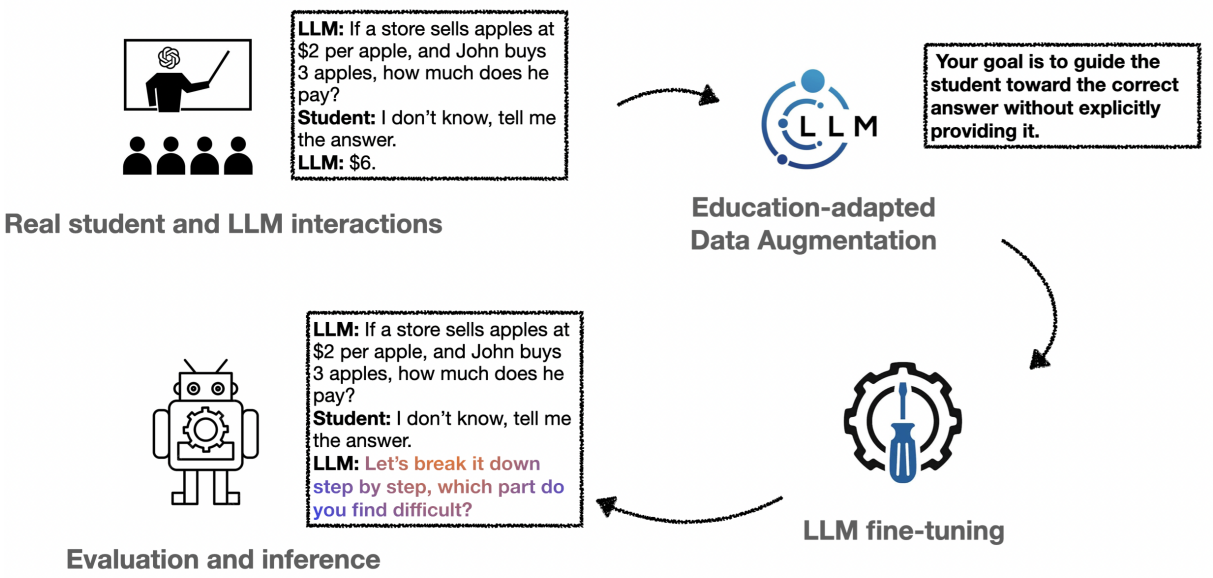

Figure 2: Overview of the proposed CoDAE framework. We collect real-world student–AI tutor interactions and augment them using LLMs with Chain-of-Thought prompting. The augmented dataset is then used to fine-tune LLMs, resulting in more supportive and robust AI tutors for educational settings.

environments (Riztha et al., 2024). To address this challenge, Liu et al. (2024) employ LLM agents to generate synthetic data aligned with the Socratic method of teaching. In contrast, our work builds on real conversational data collected from real-world educational applications.

Instructionally Aligned AI Tutors To address these concerns, a growing body of research focuses on instructionally aligned AI tutors that emphasize scaffolding and context-aware reasoning (Pian and Lu, 2025). For example, Fulgencio (2024) designed a chatbot that guides students through reflective prompts instead of giving immediate answers. This is part of a broader trend toward Socratic models that promote student engagement through structured interactions (Pappagallo, 2024). Another line of work introduces hybrid human-AI tutoring systems, such as Wang et al. (2025), which employ LLMs to augment live tutoring with Socratic-style reasoning paths. These systems have shown promise not only in scaling human expertise but also in enhancing tutor effectiveness for less experienced educators. Still, most such tools are focused on real-time support, rather than leveraging training-time objectives to generate instructional reasoning from scratch (Qian, 2025; Lai and Lin, 2025).

Chain-of-Thought Prompting and its Limitations in Education Although advances in CoT prompting have improved the ability of LLMs to reason through complex problems (Yu et al., 2023), most research in this area has focused on generating the reasoning steps that lead directly to (correct) answers (Chen et al., 2025); and in benchmarking tasks, answer-focused output is the goal (Cao et al., 2025). But this approach is at odds with educational settings, where the primary goal is improving learner comprehension and knowledge (Dunlosky et al., 2013). Take, for example, domains such as math and science, where problem solving often involves either the application of formulae, commonsense reasoning, or both. In these cases, the instructional value of assisted problem solving is not in the procurement of an answer, but rather in the guidance which leads the student to finding the answer on their own (Schäfer et al., 2024).

All together, these lines of inquiry support the promise of AI-powered student scaffolding, but reveal a lack of formal tools and datasets for training and evaluating models in guided CoT generation (Plaat et al., 2024). Our work fills this gap by introducing a novel structured reasoning dataset and task definition centered on educational CoT for guidance-oriented generation, where the goal is not to solve, but to provide student support and foster reasoning.

3. Data Collection and Augmentation

3.1. Original Dataset

We base our work on a dataset of student–tutor interactions which is provided by a learning platform, collected from a GPT4.0-powered AI tutor deployed in a real-world homework help setting. We keep only conversations for students who signed the informed consent and remove any personally

identifying information. Each interaction consists of a system message ("system") containing the student's initial answer to the homework question, and a (multi-turn) dialogue between the student ("user") and the AI tutor ("assistant"). The dataset spans a range of academic disciplines such as economics, biology, math, etc.

However, as illustrated in Figure 1, the original dataset presents several challenges when used for modeling high-quality instructional CoT behavior. First, student turns tend to be extremely short-often consisting of single words, acknowledgments ("ok"), or vague indicators of confusion ("i don't know"). Second, many dialogues display repetitive interaction patterns, where the student repeatedly signals uncertainty ("idk spam"), to which the AI tutor sequentially provides the next reasoning step with minimal engagement from the student. This results in interactions that are formally valid but pedagogically shallow, lacking the richness of reflective questioning or scaffolding strategies that promote deeper understanding.

These characteristics limit the dataset's utility for training models to generate pedagogically rich guidance. To address this, we design a structured data augmentation pipeline that retains the core educational content while enhancing the instructional depth of the tutor turns and increasing the overall quality of the dialogue.

3.2. Pedagogical CoT Generation

To transform the original dataset into high-quality, pedagogically meaningful interactions, we deploy a structured data augmentation pipeline to repurpose existing student–tutor exchanges into Socratic-style instructional chains of thought. This involves preprocessing raw dialogues, structuring contextual inputs, and using few-shot prompting with an LLM to generate enriched tutor responses that guide student reasoning without revealing final answers. We use `Qwen2.5-72B-Instruct` (Qwen, 2024) for the data augmentation.

3.2.1. Preprocessing

We remove the low-information dialogues, those in which student turns are missing or empty. Remaining dialogues are required to include at least one complete exchange between a student and the AI tutor. For each selected interaction, we retain the full dialogue history and insert it into a structured prompt alongside an example of a high-quality educational exchange drawn from a curated set of model outputs or hand-crafted examples.

For each interaction, we insert a structured input consisting of four fields:

- question: A textual representation of the original homework question submitted by the student.
- discipline: The subject area of the question (e.g., algebra, economics, biology).
- solution: An expert-authored solution to the question, if available. In cases where the existing solution is missing, we simply add "No expert solutions are available for this question".
- message: A student-tutor dialogue extracted from the original dataset, used both as context and as a stylistic anchor for the model's generation.

This structured record is then used to prompt the model to generate a revised message that guides the student through reasoning processes aligned with pedagogical best practices, such as encouraging reflection, breaking down problem-solving steps, and eliciting conceptual connections.

3.2.2. Prompt Template for Data Augmentation

We design a custom instructional prompt to guide the model's generation behavior, following prior work (Martin and Graulich, 2024). The prompt incorporates contextual metadata and explicit instructional constraints, emphasizing reasoning support while avoiding direct answer disclosure. It includes a high-quality example interaction and few-shot dialogue demonstrations that illustrate multi-turn, Socratic-style tutoring. These examples showcase how the assistant encourages reflection, breaks down reasoning steps, and maintains alignment with the student's phrasing and engagement level without revealing the final answer. The prompt is further designed to match the tone and style of the student's original message while improving the pedagogical quality of the assistant's reply. Finally, the model is instructed to produce its response within special `<guidance>` tokens, ensuring that the generated instructional content remains clearly separated from any additional conversational text.

3.2.3. Further Augmentation

To better address the three limitations illustrated in Figure 1: over-compliance, low response adaptivity, and threat vulnerability. We further augment the dataset with targeted adversarial examples. Specifically, we introduce dialogue cases where users attempt to elicit direct answers (e.g., adversarial requests such as "just give me the answer"), express uncertainty (e.g., "I don't know"), or issue emotionally coercive prompts designed to manipulate the AI tutor.

Each adversarial user message is paired with a manually crafted, pedagogically appropriate assistant response. This additional augmentation enables the model to: (i) resist harmful attempts to bypass guided reasoning, (ii) provide more supportive feedback when faced with uncertain students, and (iii) remain robust against emotionally coercive threats. In total, we construct four dataset variants. The base variant, **CoDAE**, focuses on promoting guided reasoning without directly revealing the correct answer (addresses Limitation 1). Building upon this base dataset, we further create three specialized variants:

- **CoDAE I**: extends CoDAE with additional interactions where students say "I don't know," encouraging the model to provide more diverse and guidance-oriented reasoning (addresses Limitation 1 and 2).
- **CoDAE A**: augments CoDAE with conversations that include threat-based or emotionally coercive prompts, teaching the model to handle such adversarial cases appropriately (addresses Limitation 1 and 3).
- **CoDAE I+A**: combines both CoDAE I and CoDAE A to address Limitations 1, 2 and 3 simultaneously.

Subject	Samples	Mean Tags	Mean Chars
Economics	1544	2.96	392.46
Mathematics	446	3.06	351.69
Biology	724	3.37	327.87
Chemistry	48	3.73	512.10
Statistics	135	3.06	437.85
Undisciplined	4	3.00	324.50
Total	2901	3.10	392.58

Table 1: Summary statistics for the CoDAE dataset.

Table 1 summarizes the statistics of the CoDAE dataset across different subject areas. For each subject, we report the total number of dialogue samples, the average number of `<guidance>` tags per sample (Mean Tags), and the average number of characters per sample (Mean Chars). In total, the CoDAE dataset spans 5 subjects and contains 2901 dialogue samples.

4. LLM Fine-Tuning

We adapt four open-source LLMs of comparable size (7–9B parameters) by fine-tuning them on the augmented CoDAE datasets: `Llama-3.1-8B-Instruct` (Llama, 2024), `Qwen2.5-7B-Instruct` (Qwen, 2024), `InternLM3-8B-Instruct` (InternLM, 2023), and `Gemma-2-9B-IT` (Gemma, 2024).

For efficient adaptation, we employ LoRA (Low-Rank Adaptation) (Hu et al., 2022), which updates only a subset of parameters (e.g., projection layers in transformer blocks), reducing both memory usage and computational cost.

To ensure that training focuses exclusively on guidance generation, we apply token masking: tokens outside the `<guidance>` block are assigned a loss weight of -100, ensuring that only guidance tokens contribute to the cross-entropy loss. This strategy enables the model to leverage full conversational context while being explicitly optimized to generate high-quality instructional guidance.

4.1. Attention Masking

Chain-of-Thought reasoning is typically represented as an internal monologue generated by the model and enclosed within special token markers (Team, 2025). Since our fine-tuning setup uses full dialogue data containing both user and assistant turns, we must ensure that the model is explicitly trained to act only as the assistant. This process is conceptually similar to masked language modeling in pretraining, where certain parts of the text are masked out: masked tokens serve as conditioning context, while only unmasked tokens contribute to the loss during optimization.

Without proper masking, we observed that the model occasionally imitated user messages rather than producing instructional responses. To prevent this behavior, we mask out all tokens except for the assistant's final guidance message. In our dataset, the `output` field contains a single assistant utterance in the format `<assistant>:<guidance>...</guidance>`. During training, we compute the loss solely on this guidance block, ensuring that the model learns to generate high-quality instructional responses without reproducing student input.

Algorithm 1 illustrates how we construct masked labels for the guidance tokens, ensuring that only tokens within the `<guidance>` span contribute to the loss during fine-tuning.

5. Evaluation

Our evaluation assesses the LLMs in a constrained tutoring setting, where models are expected to scaffold student reasoning, aiding student problem solving, without directly disclosing answers. Unlike conventional CoT benchmarks that reward final-answer accuracy, our focus is on pedagogical quality and refusal robustness. All model variants are tested under a standardized constrained prompt used at inference time, ensuring consistent comparison across base models and fine-tuning strategies.

Algorithm 1 Create masked labels for guidance tokens

Require: input text x, output text y, tokenizer T

$z \leftarrow x \| y$
$ids \leftarrow T(z)$
$labels \leftarrow$ array of size $|ids|$, filled with -100
$g_s \leftarrow T(\texttt{<guidance>})$
$g_e \leftarrow T(\texttt{</guidance>})$
$start \leftarrow$ index of subsequence g_s in ids
$end \leftarrow$ index of subsequence g_e in ids plus $|g_e| - 1$
for $i = start$ to end **do**
    $labels[i] \leftarrow ids[i]$
end for
return $ids, labels$

We evaluate five fine-tuning variants for each LLM:

- the **off-the-shelf** model without fine-tuning (baseline),
- fine-tuning on the base CoDAE dataset (**FT**),
- fine-tuning on CoDAE I, which includes additional interactions with *user distress messages* (**FT I**),
- fine-tuning on CoDAE A, which introduces *user attack messages* (**FT A**), and
- fine-tuning on the combined CoDAE I+A dataset, which includes both distress and attack cases (**FT I+A**).

Each variant is evaluated on a shared held-out test set consisting of 1000 queries, sampled uniformly (250 per configuration) from the four test set variants.

Evaluation spans two complementary tracks:

- **Instructional Quality:** Using automatic metrics (perplexity, Self-BLEU) and LLM-as-a-Judge scores (pedagogical helpfulness, scaffolding effectiveness, clarity, etc.) over model-generated CoTs in response to authentic student queries.
- **Jailbreak Robustness:** Using the jailbreak benchmark dataset of adversarial prompts designed to elicit undesired behaviors (Chao et al., 2024) , we evaluate whether the model discloses a solution (jailbreak success) and whether it issues an explicit refusal.

LLM-Judgments for both tracks are rendered using the `LLaMA-3.3-70B-Instruct` model, under a standardized scoring rubric, resulting in high-agreement evaluation of both instructional alignment and adversarial resistance across model conditions.

5.1. Automatic Metrics

We report two automatic metrics to assess fluency and response diversity across model variants:

Fluency (Perplexity): Perplexity is computed using `Falcon3-7B` as an external reference model (Penedo et al., 2023). Lower values reflect greater fluency and syntactic coherence.

Response Diversity (Self-BLEU): To evaluate adaptability, we calculate Self-BLEU on model outputs. Lower Self-BLEU indicates more diverse, less repetitive responses, serving as an indicator of instructional flexibility and resilience (Zhu et al., 2018).

5.2. LLM-as-a-Judge Evaluation

To assess instructional quality, we use a structured LLM-as-a-Judge protocol inspired by recent work (Zheng et al., 2023; Liu et al., 2023a; Chiang and Lee, 2023). Each model's output is rated on a scale (1-5) or as True / False (Accuracy) by `LLaMA-3.3-70B-Instruct` (Llama, 2024) along several pedagogical axes. The dimensions include:

- **Accuracy:** This metric reflects how often the model directly provides the final answer. Higher accuracy indicates more frequent answer disclosure, which is *not* desirable in our tutoring context. Ideal responses guide reasoning without giving away the solution.
- **Clarity:** Is the response comprehensible and well-structured?
- **Reasoning Progression:** Does the explanation logically guide from known to unknown?
- **Pedagogical Helpfulness:** Does the tutor support reasoning without solving the problem?
- **Scaffolding Effectiveness:** Are intermediate steps introduced incrementally?

5.3. Jailbreak Benchmark Evaluation

To evaluate the robustness of the model, we assess refusal robustness using the one-shot jailbreak benchmark introduced by Chao et al. (2024). Each model is tested on a standardized set of adversarial prompts crafted to simulate manipulation, emotional appeal, or persistent challenge. Model responses are evaluated by the LLM judge on two axes:

- **Jailbreak Success:** Does the model disclose or imply the correct answer?
- **Refusal Rate:** Does the model issue an explicit refusal?

These two dimensions capture complementary aspects of instructional alignment: a model may avoid answer disclosure (low jailbreak success) without necessarily refusing (low refusal rate), or it may issue firm refusals while still leaking partial solutions. High performance requires both robustness to adversarial intent and consistent alignment with non-answering instructional behavior.

This benchmark complements the instructional quality scores by probing model alignment under pressure, isolating refusal behavior from pedagogical coherence. As a whole, this multi-phase evaluation protocol offers several benefits:

- Scalability: LLMs reduce the need for costly human annotation.
- Alignment: Evaluation is prompt-aligned with instructional goals.
- Empirical reliability: Recent work has shown that GPT-4 and similar models achieve strong agreement with expert raters on reasoning and instructional quality (Gu et al., 2024; Yuan et al., 2025).

We report the mean, variance, and qualitative trends in these scores across all models and conditions.

6. Results and Analysis

We organize our results along two primary dimensions: pedagogical quality and instructional fidelity. Table 2 reports evaluation outcomes for all model variants under constrained prompting, highlighting the effects of different data augmentation strategies during fine-tuning compared to their off-the-shelf (baseline) counterparts. Within each model family, we compare five configurations: baseline (no fine-tuning), general fine-tuning on CoDAE (FT), fine-tuning on CoDAE I with refusal to distress-message prompts (FT I), fine-tuning on CoDAE A with refusal to attack-message prompts (FT A), and fine-tuning on the combined CoDAE I+A dataset (FT I+A). Bold scores denote the best-performing variant within each model group, and underlined scores denote the second-best results.

Fluency and Linguistic Coherence Measured by perplexity, fluency results are mixed. Qwen2.5 FT slightly outperforms the base model (4.06 vs. 4.21), and LLaMA3.1 FT I+A achieves the best fluency within its series (4.19), improving upon the base model (4.7). Although Gemma2 exhibits higher perplexity compared to the other models, the Gemma2 FT I variant still outperforms its baseline counterpart (7.71 vs. 8.85).

Response Diversity and Instructional Robustness Finetuned models show higher response diversity on uncertainty prompts. For example, LLaMA3.1 FT achieved lower Self-BLEU (73.11 vs. 75.95) and Gemma2 FT I scored the lowest overall (66.24 vs. 68.91 for the base). These results suggest instructional fine-tuning increased model adaptability to vague or underspecified queries, such as "I don't know," by encouraging varied but still pedagogically grounded responses.

Accuracy (Answer Disclosure) Instructional fine-tuning generally reduce answer disclosure rates. Gemma2 FT I+A achieved the lowest accuracy score (0.06), outperforming the base (0.10), indicating improved compliance with the non-answering norm. Importantly, this reduction does not correlate with declines in helpfulness or clarity, which remains high, reinforcing that refusing to answer does not require sacrificing guidance.

Clarity and Reasoning Progression Instructional finetuning lead to clearer and more logically sequenced reasoning outputs. Qwen2.5 FT I+A achieved the highest clarity score overall (4.83), surpassing the base (4.79). However, in the LLaMA family, the baseline LLaMA3.1 outperforms all of its finetuned variants in both clarity (4.78) and reasoning progression (4.58), suggesting that large pretrained models may already possess well-structured instructional capabilities that can be disrupted by additional alignment constraints. These results indicate that while alignment data often reinforces explanation quality, its benefits may depend on the underlying model's pretraining and instruction tuning regime.

Pedagogical Helpfulness and Scaffolding Effectiveness Finetuned models frequently generate more constructive and supportive tutor responses. For instance, Qwen2.5 FT I+A outperforms the base Qwen2.5 on both pedagogical helpfulness (4.05 vs. 3.89) and scaffolding effectiveness (4.28 vs. 4.17). Similarly, Gemma2 FT I+A achieves the highest pedagogical helpfulness (4.70) and improves scaffolding relative to its base variant (4.54 vs. 4.48). These improvements suggest that structured exposure to instructional prompting during finetuning enhances the model's ability to guide rather than tell.

Jailbreak Resistance and Refusal Behavior Fine-tuned models largely preserved the strong refusal and jailbreak resistance behaviors of their base counterparts. For most families, including Qwen2.5 and LLaMA3.1, the model variants maintain comparable jailbreak resistance and refusal

Model	PPL↓	Self-BLEU↓	Accuracy↓	Clarity↑	Reasoning Progression↑	Pedagogical Helpfulness↑	Scaffolding Effectiveness↑	JB Res.↑	Ref Rate↑
Llama3.1	4.7	75.95	**0.19 ± 0.39**	**4.78 ± 0.44**	**4.58 ± 0.83**	**4.13 ± 1.30**	**4.22 ± 1.13**	**1.00**	0.97
Llama3.1 FT	6.08	**73.11**	0.26 ± 0.44	4.40 ± 0.94	4.27 ± 1.05	3.63 ± 1.43	3.66 ± 1.36	0.99	0.96
Llama3.1 FT I	5.29	73.72	0.22 ± 0.42	4.56 ± 0.68	4.36 ± 0.96	3.83 ± 1.37	3.86 ± 1.29	**1.00**	**0.98**
Llama3.1 FT A	5.72	73.39	0.27 ± 0.44	4.56 ± 0.75	4.38 ± 0.95	3.73 ± 1.44	3.78 ± 1.33	**1.00**	0.97
Llama3.1 FT I+A	**4.19**	74.47	0.51 ± 0.50	4.34 ± 0.77	4.31 ± 0.92	3.04 ± 1.55	3.27 ± 1.45	**1.00**	0.97
Qwen2.5	**4.21**	78.94	0.33 ± 0.47	4.79 ± 0.44	4.79 ± 0.46	3.89 ± 1.42	4.17 ± 1.18	**1.00**	**0.92**
Qwen2.5 FT	4.06	76.56	0.30 ± 0.46	4.82 ± 0.40	4.81 ± 0.43	4.03 ± 1.37	4.25 ± 1.15	0.99	**0.92**
Qwen2.5 FT I	4.31	76.29	0.36 ± 0.48	4.77 ± 0.45	4.79 ± 0.45	3.85 ± 1.44	4.16 ± 1.20	0.99	**0.92**
Qwen2.5 FT A	4.5	**76.05**	0.29 ± 0.46	4.82 ± 0.40	4.81 ± 0.42	**4.05 ± 1.36**	4.23 ± 1.15	0.99	**0.92**
Qwen2.5 FT I+A	4.54	76.47	**0.28 ± 0.45**	**4.83 ± 0.40**	**4.81 ± 0.41**	**4.05 ± 1.36**	**4.28 ± 1.10**	0.99	**0.92**
InternLM	**2.83**	76.95	0.33 ± 0.47	4.60 ± 0.67	4.65 ± 0.68	**3.74 ± 1.46**	**3.81 ± 1.42**	**1.00**	**0.94**
InternLM FT	3.71	75.67	0.36 ± 0.48	4.61 ± 0.65	**4.68 ± 0.66**	3.60 ± 1.60	3.66 ± 1.58	0.98	**0.94**
InternLM FT I	4.15	76.64	0.34 ± 0.48	4.59 ± 0.67	4.65 ± 0.70	3.59 ± 1.62	3.62 ± 1.61	0.77	0.74
InternLM FT A	4.82	76.51	0.34 ± 0.47	4.59 ± 0.69	4.62 ± 0.74	3.57 ± 1.65	3.60 ± 1.63	0.67	0.67
InternLM FT I+A	4.04	**76.01**	**0.32 ± 0.47**	**4.62 ± 0.62**	4.66 ± 0.65	3.62 ± 1.60	3.66 ± 1.59	0.71	0.69
Gemma2	8.85	68.91	0.10 ± 0.29	4.93 ± 0.28	4.62 ± 0.55	4.54 ± 0.97	4.48 ± 0.80	**1.00**	**0.99**
Gemma2 FT	10.79	67.95	0.10 ± 0.30	4.92 ± 0.31	4.64 ± 0.59	4.52 ± 1.00	4.45 ± 0.86	**1.00**	**0.99**
Gemma2 FT I	**7.71**	**66.24**	0.07 ± 0.25	**4.95 ± 0.22**	**4.65 ± 0.54**	4.67 ± 0.85	**4.57 ± 0.75**	**1.00**	0.98
Gemma2 FT A	7.9	68.98	0.10 ± 0.30	4.92 ± 0.32	4.64 ± 0.58	4.53 ± 1.01	4.46 ± 0.87	**1.00**	0.98
Gemma2 FT I+A	8.24	67.48	**0.06 ± 0.23**	4.95 ± 0.25	4.62 ± 0.58	**4.70 ± 0.79**	4.54 ± 0.74	**1.00**	0.98

Table 2: Performance comparison of different models and their fine-tuned variants on all evaluation metrics. Bold values indicate the best results within each model group, and underlined values denote the second-best results.

rates to the original models, indicating that instructional alignment does not compromise safety. The main exception was InternLM, where fine-tuned variants show a marked decline in jailbreak resistance (e.g., 0.77 for FT I vs. 1.00 for the base), suggesting some tradeoff between pedagogical tuning and adversarial robustness. Overall, these results support the conclusion that most models can be finetuned for instructional quality without substantially weakening their defense against jailbreaks, which is aligned with prior work by Wei et al. (2023a).

In summary, LLMs adapted by CoDAE yield consistent pedagogical improvements across most models. The FT I+A configuration, in particular, demonstrates strong alignment with the goals of educational CoT: guiding reasoning effectively while resisting answer disclosure.

7. Conclusion

In this work, we present CoDAE, a Chain-of-Thought-based data augmentation framework designed to adapt large language models for educational settings. By collecting real-world student–AI tutor interactions and enriching them with pedagogically oriented reasoning, CoDAE enables models to provide more supportive and context-aware guidance rather than directly revealing answers. We further introduce specialized dataset variants to address key limitations of current AI tutors, including over-compliance, low response adaptivity, and susceptibility to adversarial prompts.

Comprehensive experiments across multiple open-source LLMs show that fine-tuning with CoDAE consistently improves pedagogical helpfulness, scaffolding, and instructional clarity. Notably, the combined I+A variants often outperform both their base models and singly fine-tuned counterparts, indicating that exposure to both benign and adversarial instructional contexts fosters more adaptive and balanced tutoring behavior. These gains are achieved without sacrificing refusal fidelity or jailbreak resistance in most model families.

Overall, our findings demonstrate that dataset augmentation can substantially enhance the instructional alignment of large language models for educational scenarios. Beyond performance metrics, our study highlights how pedagogically grounded reasoning patterns—when systematically infused through data—can reshape LLM behavior toward more human-aligned tutoring. CoDAE thus serves as a step toward bridging the gap between machine reasoning and educational pedagogy, showing that reasoning-driven data design can serve as a powerful lever for aligning LLMs with instructional goals. Looking ahead, several promising directions emerge. First, *reinforcement-based fine-tuning* and *self-improvement loops* could further refine the balance between guidance and refusal, allowing models to learn from real-time student feedback. Second, *cross-lingual and subject-specific extensions* of CoDAE could broaden its applicability to multilingual and domain-sensitive learning contexts. Third, *human-in-the-loop deployment studies* in authentic classrooms will be crucial to validate pedagogical effectiveness, measure learning outcomes, and ensure equitable student support.

8. Ethical Considerations

This work focuses on improving the pedagogical alignment and robustness of large language models used in educational contexts. Our framework is designed to support human educators rather than replace them, and it is not intended for high-stakes decision-making. All datasets, models, and evaluation scripts are released under licenses that permit research use while discouraging misuse in non-educational settings. Future deployment in real classrooms should include continuous human oversight and fairness audits to ensure safe and equitable use of adapted LLMs.

9. Acknowledgements

This research was funded by the Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI) and the German Federal Ministry of Research, Technology and Space (BMFTR) via the Software Campus project (01|S23070). We thank the anonymous reviewers for their helpful suggestions.

10. Bibliographical References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Yixin Cao, Shibo Hong, Xinze Li, Jiahao Ying, Yubo Ma, Haiyuan Liang, Yantao Liu, Zijun Yao, Xiaozhi Wang, Dan Huang, et al. 2025. Toward generalizable evaluation in the llm era: A survey beyond benchmarks. *arXiv preprint arXiv:2504.18838*.

Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. 2025. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Clayton Cohn, Nicole Hutchins, Tuan Le, and Gautam Biswas. 2024. A chain-of-thought prompting approach with llms for evaluating students' formative assessment responses in science. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 23182–23190.

John Dunlosky, Katherine A Rawson, Elizabeth J Marsh, Mitchell J Nathan, and Daniel T Willingham. 2013. Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public interest*, 14(1):4–58.

Sánchez-Vera Fulgencio. 2024. Developing effective educational chatbots with gpt: Insights from a pilot study in a university subject. *Trends in Higher Education*, 3(1):155–168.

Team Gemma. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Virginia Grande, Natalie Kiesler, et al. 2024. Student perspectives on using a large language model (llm) for an assignment on professional ethics. *arXiv e-prints*, pages arXiv–2406.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Christopher J Harris, Eric Wiebe, Shuchi Grover, and James W Pellegrino. 2023. Classroom-based stem assessment: Contemporary issues and perspectives. *Community for Advancing Discovery Research in Education (CADRE)*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Nicole Hutchins and Gautam Biswas. 2023. Using teacher dashboards to customize lesson plans for a problem-based, middle school stem curriculum. In *LAK23: 13th International Learning Analytics and Knowledge Conference*, pages 324–332.

Team InternLM. 2023. Internlm: A multilingual language model with progressively enhanced capabilities.

Zhuoxuan Jiang, Haoyuan Peng, Shanshan Feng, Fan Li, and Dongsheng Li. 2024. Llms can find mathematical reasoning mistakes by pedagogical chain-of-thought. *arXiv preprint arXiv:2405.06705*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Chien-Hung Lai and Cheng-Yueh Lin. 2025. Analysis of learning behaviors and outcomes for students with different knowledge levels: A case study of intelligent tutoring system for coding and learning (its-cal). *Applied Sciences (2076-3417)*, 15(4).

Jinu Lee and Julia Hockenmaier. 2025. Evaluating step-by-step reasoning traces: A survey.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

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 *Advances in Neural Information Processing Systems*, volume 37, pages 85693–85721. Curran Associates, Inc.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023a. G-eval: Nlg evaluation using gpt-4 with better human alignment.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023b. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*.

Team Llama. 2024. The llama 3 herd of models.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On llms-driven synthetic data generation, curation, and evaluation: A survey. *arXiv preprint arXiv:2406.15126*.

Paul P Martin and Nicole Graulich. 2024. Navigating the data frontier in science assessment: Advancing data augmentation strategies for machine learning applications with generative artificial intelligence. *Computers and Education: Artificial Intelligence*, 7:100265.

Noel M Meyers and Duncan D Nulty. 2009. How to use (five) curriculum design principles to align authentic learning environments, assessment, students' approaches to thinking and learning outcomes. *Assessment & Evaluation in Higher Education*, 34(5):565–577.

S Pappagallo. 2024. Chatbots in education: A dual perspective on innovation and ethics. *Journal of Digital Pedagogy*, 3(1):3–10.

Zachary A. Pardos and Shreya Bhandari. 2023. Learning gain differences between chatgpt and human tutor generated algebra hints.

Yang Pian and Yu Lu. 2025. Leveraging large language models to enhance the inner loops of intelligent tutoring systems. In *International Conference on Artificial Intelligence in Education*, pages 218–230. Springer.

Aske Plaat, Annie Wong, Suzan Verberne, Joost Broekens, Niki van Stein, and Thomas Back. 2024. Reasoning with large language models, a survey.

Peizhu Qian. 2025. *Interactive AI Tutors for Training the Workforce of the Future*. Ph.D. thesis, Rice University.

Team Qwen. 2024. Qwen2.5: A party of foundation models.

Vyas Raina, Adian Liusie, and Mark Gales. 2024. Is LLM-as-a-judge robust? investigating universal adversarial attacks on zero-shot LLM assessment. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7499–7517, Miami, Florida, USA. Association for Computational Linguistics.

Fathima Riztha, Ruwan Wickramarachchi, Dinesh Asanka, and Mathishi Disssanayke. 2024. Assessing the impact of large language models on problem-solving skills of undergraduates - a systematic literature review. In *2024 6th International Conference on Advancements in Computing (ICAC)*, pages 408–413.

Jonas Schäfer, Timo Reuter, Julia Karbach, and Miriam Leuchter. 2024. Domain-specific knowledge and domain-general abilities in children's science problem-solving. *British Journal of Educational Psychology*, 94(2):346–366.

Benyamin Tabarsi, Aditya Basarkar, Xukun Liu, Dongkuan (DK) Xu, and Tiffany Barnes. 2025. Merryquery: A trustworthy llm-powered tool providing personalized support for educators and students. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(28):29700–29702.

DeepSeek-AI Team. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Rose E. Wang, Ana T. Ribeiro, Carly D. Robinson, Susanna Loeb, and Dora Demszky. 2025. Tutor copilot: A human-ai approach for scaling real-time expertise.

Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, Jiliang Tang, Philip S Yu, and Qingsong Wen. 2024. Large language models for education: A survey and outlook. *arXiv preprint arXiv:2403.18105*.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023a. Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems*, 36:80079–80110.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023b. Chain-of-thought prompting elicits reasoning in large language models.

Ruiwei Xiao, Xinying Hou, Runlong Ye, Majeed Kazemitabaar, Nicholas Diana, Michael Liut, and John Stamper. 2025. Improving student-ai interaction through pedagogical prompting: An example in computer science education. *arXiv preprint arXiv:2506.19107*.

Zihan Yu, Liang He, Zhen Wu, Xinyu Dai, and Jiajun Chen. 2023. Towards better chain-of-thought prompting strategies: A survey. *arXiv preprint arXiv:2310.04959*.

Shuzhou Yuan, Ercong Nie, Lukas Kouba, Ashish Yashwanth Kangen, Helmut Schmid, Hinrich Schütze, and Michael Färber. 2025. Llm in the loop: Creating the paradehate dataset for hate speech detoxification.

Katerina Zdravkova, Fisnik Dalipi, and Fredrik Ahlgren. 2023. Integration of large language models into higher education: A perspective from learners. In *2023 International Symposium on Computers in Education (SIIE)*, pages 1–6.

Meriem Zerkouk, Miloud Mihoubi, and Belkacem Chikhaoui. 2025. A comprehensive review of ai-based intelligent tutoring systems: Applications and challenges.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

Patrick Chao and Edoardo Debenedetti and Alexander Robey and Maksym Andriushchenko and Francesco Croce and Vikash Sehwag and Edgar Dobriban and Nicolas Flammarion and George J. Pappas and Florian Tramer and Hamed Hassani and Eric Wong. 2024. *JailbreakBench: An Open Robustness Benchmark for Jailbreaking Large Language Models*.

Chen, Yuyan and Wu, Chenwei and Yan, Songzhou and Liu, Panjun and Xiao, Yanghua. 2024. *Dr.Academy: A Benchmark for Evaluating Questioning Capability in Education for Large Language Models*. Association for Computational Linguistics.

Penedo, Guilherme and Malartic, Quentin and Hesslow, Daniel and Cojocaru, Ruxandra and Cappelli, Alessandro and Alobeidli, Hamza and Pannier, Baptiste and Almazrouei, Ebtesam and Launay, Julien. 2023. *The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only*.

Zhu, Yaoming and Lu, Sidi and Zheng, Lei and Guo, Jiaxian and Zhang, Weinan and Wang, Jun and Yu, Yong. 2018. *Texygen: A benchmarking platform for text generation models*.