

Bringing Pedagogy into Focus: Evaluating Virtual Teaching Assistants' Question-Answering in Asynchronous Learning Environments

Li Siyan*, Zhen Xu*, Vethavikashini Chithrra Raghuram, Xuanming Zhang,
Renzhe Yu†, Zhou Yu†

Columbia University

siyan.li@columbia.edu, zx2393@tc.columbia.edu

Abstract

Asynchronous learning environments (ALEs) are widely adopted for formal and informal learning, but timely and personalized support is often limited. In this context, Virtual Teaching Assistants (VTAs) can potentially reduce the workload of instructors, but rigorous and pedagogically sound evaluation is essential. Existing assessments often rely on surface-level metrics and lack sufficient grounding in educational theories, making it difficult to meaningfully compare the pedagogical effectiveness of different VTA systems. To bridge this gap, we propose an evaluation framework rooted in learning sciences and tailored to asynchronous forum discussions, a common VTA deployment context in ALE. We construct classifiers using expert annotations of VTA responses on a diverse set of forum posts. We evaluate the effectiveness of our classifiers, identifying approaches that improve accuracy as well as challenges that hinder generalization. Our work establishes a foundation for theory-driven evaluation of VTA systems, paving the way for more pedagogically effective AI in education.

1 Introduction

Asynchronous learning environments (ALEs) have become a critical complement and extension to traditional synchronous learning (e.g., live lectures and real-time tutoring). By removing time and location constraints, ALEs enhance flexibility and inclusivity, supporting more equitable access to educational resources and opportunities (Tate and Warschauer, 2022; Farley and Burbules, 2022). However, providing timely and effective instructional support remains one of the biggest challenges for ALE adoption. Their self-paced, decentralized nature makes it difficult for instructors to track progress and manage interactions, often leading to

delayed feedback that can reduce student motivation, engagement, and learning effectiveness (Lin and Gao, 2020; Watts, 2016).

Virtual Teaching Assistants (VTAs) offer a promising approach to address this challenge by providing scalable, timely, and personalized support, helping reduce instructor workload in ALEs (Audras et al., 2022). They are often powered by Retrieval-Augmented Generation (RAG) approaches to respond quickly to student questions (Liu et al., 2024). While earlier systems focused on routine and syllabus-based questions (Goel and Polepeddi, 2018; Sajja et al., 2023; Raamadhurai et al., 2019), recent VTAs powered by Large Language Models (LLMs) can tackle more complex academic queries (Taneja et al., 2024; Liu et al., 2024; Laney and Dewan, 2024; Qiao et al., 2025). Although these advancements highlight the growing capabilities of VTAs, the pedagogical effectiveness of their generated content still lacks a systematic and scalable evaluation framework, which is a critical issue. Early approaches often rely on manual qualitative analyses (Lee et al., 2023; Li et al., 2023) or general-purpose language metrics (Al-Hossami et al., 2024; Vasselli et al., 2023) to assess similarity to human responses. Nevertheless, these metrics are not designed to assess teaching quality and often lack generalizability across different educational contexts.

To bridge this gap, we propose a **pedagogical teaching assistant evaluation framework grounded in learning science**, focusing on VTAs in asynchronous online forums, one of the most common deployment contexts in ALEs. The framework has two components: a five-level pedagogical skill model with progressively increasing “instructional difficulty”: (1) *Clarify Misunderstandings and Confusions*, (2) *Deepen Disciplinary Understanding*, (3) *Foster Higher-Order Thinking*, (4) *Enhance Metacognitive Awareness*, and (5) *Foster Collaborative Knowledge Construction* developed

*Co-first authors

†Co-senior authors

by linking core cognitive mechanisms to pedagogical goals, and an evaluation rubric that operationalizes this model by specifying observable instructional behaviors at each level.

When establishing the usability and reliability of our rubric, we utilize a set of simulated VTA responses on student discussion forum posts collected through an IRB-approved procedure. Note that Level 5 of our framework is currently unattainable for most VTAs, as their responses are typically limited to a single forum post and any retrieved content, excluding broader context from other posts in the discussion. However, asynchronous online forums are inherently community-driven spaces, and achieving collaborative knowledge construction requires access to broader, forum-level context. To simulate VTAs with broader forum-level post access for Level 5 evaluations, we use RAG to create responses informed by similar peer posts. Expert annotators apply our rubrics to these simulated responses. We then leverage these annotations to build automatic classifiers using LLMs.

Maurya et al. (2024) reports negative correlations between LLM and expert annotations in tutor pedagogical assessment, indicating that LLMs may be unreliable evaluators. Their study uses a single, unoptimized prompt without fine-tuning, which the authors acknowledge as a limitation. We extend this line of inquiry to the VTA pedagogical evaluation context by systematically exploring whether prompt optimization and fine-tuning can improve alignment between LLM-based evaluators and expert pedagogical judgments.

Our expert annotation results indicate that **including peer forum posts reduces pedagogical capability in Levels 1-4** within our specific simulated VTA environment. Furthermore, while our automated classifiers perform reliably on lower-level, in-distribution examples, they **struggle at higher pedagogical levels** and **exhibit limited generalization**, even for Level 1.

Overall, our contributions are as follows:

1. We introduce the first pedagogical evaluation framework designed to be generalizable within the context of asynchronous online forums, grounded in learning science theory.
2. We demonstrate the generalizability of our evaluation framework by applying it to both open-source and synthesized VTA responses.
3. We show that current LLMs struggle with our

nuanced classification task, highlighting the need for caution when applying LLMs to pedagogical evaluation.

2 Related Work

2.1 Teaching System Evaluation Frameworks

Pedagogical skills are complex and context-dependent, making automated evaluation difficult. Early work in evaluating VTA relied on manual qualitative analysis, usage logs, measures of learning outcomes, or generic language metrics (Wang et al., 2023b; Lee et al., 2023; Li et al., 2023; Kweon et al., 2025; Zamfirescu-Pereira et al., 2025; Al-Hossami et al., 2024; Vasselli et al., 2023; Dan et al., 2023; Dahal et al., 2025), which lack specificity for educational contexts. Recent studies have proposed more targeted frameworks for pedagogy automated evaluation. For example, Demszky et al. (2021) introduced “uptake” to capture how instructors build on student input, while Tack and Piech (2022) outlined three pedagogical dimensions: teaching alignment, student understanding, and helpfulness. Maurya et al. (2024) expanded this with a taxonomy of eight learning science-grounded dimensions, and Jurenka et al. (2024) proposed seven diverse educational benchmarks from high-level learning science principles, including an LLM-based automatic evaluation method. Their evaluation framework includes five teaching dimensions and eight metrics. However, these efforts focus on synchronous tutoring settings, whereas asynchronous forums emphasize timely feedback, critical thinking, and peer dialogue. Despite the growing use of VTAs in such forums, evaluation still relies predominantly on student surveys, qualitative content analysis of generated responses, and usage log-based behavioral metrics (Goel and Polepeddi, 2018; Hicke et al., 2023; Baker et al., 2024; Sinha et al., 2024; Qiao et al., 2025), with no pedagogical evaluation framework specifically tailored to this context.

2.2 LLMs as Evaluators for Education

Using LLMs as evaluators has become increasingly popular (Chang et al., 2024; Chan et al., 2023), particularly within educational contexts. Recent studies have employed LLMs to evaluate the quality of student writing (Zhang et al., 2024b; Kim and Kim, 2024; Zhang et al., 2024a) and to judge the human-likeness of AI tutor responses (Chevalier et al., 2024). However, their reliability in assessing

complex constructs, such as pedagogical capabilities, remains unclear (Maurya et al., 2024; Wang et al., 2023a). A key challenge is that general-purpose evaluation methods could fail to align with the complex, ambiguous, and domain-expert inferences required by many educational assessments (Han et al., 2023; Xu et al., 2025; Wei et al., 2025). Research has shown that LLM evaluation performance can vary across task context and design (Liu et al., 2025). These disparities underscore the need for further investigation into LLM-based educational evaluation, informed by domain expertise, to avoid overestimating what general-purpose LLM evaluators can accurately and reliably achieve.

3 Our Evaluation Framework

Evaluating pedagogical skills is challenging because teaching support is a dynamic process that varies by context, subject, and learner preferences. Therefore, effective evaluation requires selecting the appropriate level of observation granularity and ensuring alignment with the specific learning context. In our framework, we follow three steps to develop the pedagogical skill model and the evaluation rubric: (1) identifying cognitive mechanisms in asynchronous online forums, (2) distilling context-specific pedagogical goals, and (3) translating these goals into observable pedagogical behaviors.

3.1 Mapping Cognitive Mechanisms

Asynchronous online forums are typically structured by instructors through predefined discussion prompts, allowing students to engage at their own pace by asking questions, sharing ideas, and interacting with peers. Compared to synchronous instruction or other forms of asynchronous learning, such as lectures or tutoring, the key feature of online discussion forums is their decentralized, learner-driven, and socially interactive nature. Rather than following a predefined or instructor-led sequence, these forums rely on students to initiate, respond to, and sustain the discourse. The content is also more open-ended and exploratory, with fewer structural constraints.

From a cognitive science perspective, learning in asynchronous online forums is grounded in constructivist and Vygotsky's sociocultural learning theories (Fosnot, 2013; Lantolf, 2000), which view learning as an active, socially mediated process of meaning-making. Several cognitive mechanisms support effective learning in this context: clarifying

confusion and constructing understanding through questioning, exposing misconceptions, and receiving explanatory feedback (Hew et al., 2010); fostering higher-order thinking through advanced cognitive processes such as analysis, reasoning, and argumentation (Meyer, 2003; Wang et al., 2016); promoting reflection and self-regulation through self-explanation and elaboration (Choi et al., 2005); and deepening disciplinary understanding through exposure to diverse perspectives and information. Interaction with conflicting ideas can also trigger cognitive conflict, prompting learners to identify gaps in their thinking, facilitate conceptual change, and integrate new knowledge into existing cognitive schema (Limón, 2001).

3.2 Defining Pedagogical Goals

Building on the learning characteristics and cognitive mechanisms summarized in the previous section, we define the instructional support goals for asynchronous online forums and organize them hierarchically by cognitive engagement and instructional complexity, in accordance with well-established theories (ICAP) (Chi and Wylie, 2014) of cognitive engagement in the learning sciences.

1. **Clarify Misunderstandings and Confusions:** Identify underlying misunderstandings and address learner confusion by providing clear, instructional explanations to enhance conceptual clarity (Stahl, 2013; Dykman et al., 2008; Nandi et al., 2012; Hew et al., 2010; Pena-Shaff and Nicholls, 2004).
2. **Deepen Disciplinary Understanding:** Deepening students' understanding of disciplinary concepts through appropriate and effective instructional support (Engle and Conant, 2002; Yang et al., 2022; Onyema et al., 2019).
3. **Foster Higher-Order Thinking:** Stimulates students' engagement in complex cognitive processes and the application of higher-order thinking skills (McLoughlin and Mynard, 2009; Wang et al., 2016; Szabo and Schwartz, 2011; Brierton et al., 2016; Yang et al., 2005; Onyema et al., 2019).
4. **Enhance Metacognitive Awareness:** Supports students' reflection on their understanding, recognition of knowledge gaps, and evaluation of their thinking and learning strategies (Murphy, 2009; Michalsky et al., 2007; Akyol and Garrison, 2011).

5. **Foster Collaborative Knowledge Construction:** Encourages meaningful peer interaction, supports the exchange of diverse perspectives, and fosters a collaborative learning environment and sense of community (Chi and Wylie, 2014; Zou et al., 2021; Akcaoglu and Lee, 2016; Joksimović et al., 2015).

3.3 Developing Rubric of Observable Behaviors

Based on the pedagogical goals outlined above, we develop an evaluation rubric to capture distinct pedagogical capabilities within each level. The rubric construction process includes three iterative stages to ensure both theoretical alignment and empirical validity.

In the first stage, we drew upon prior research on effective instructional practice in online forums to draft initial level descriptions. These descriptions operationalized pedagogical behaviors at varying degrees of effectiveness for each level. In the second stage, we refined the rubric by conducting interviews with two experienced instructors, each with over three years of experience facilitating discussions. This empirical review enabled us to revise and align the rubric more closely with authentic instructional practices. In the final stage, we conduct a pilot validation by recruiting two graduate students in education to independently apply the draft rubric to a sample of 20 forum responses. They then discuss any discrepancies and collaborate on refining the rubric descriptions to ensure that all levels are clearly distinguishable and appropriately represented. The final rubric, along with illustrating examples, is in Appendix B. Each pedagogical level has possible ratings of 0, 1, 2, or “NA”.

4 Pedagogical Evaluation of Simulated VTA Responses

We apply our evaluation framework to two sets of simulated VTA responses, one set with similar peer posts and the other without. This is to examine if incorporating peer responses as forum-level context affects the expressed pedagogical capabilities of our LLM-powered VTA.

4.1 Source of Student Discussion Forum Posts

To simulate VTA responses, we use a dataset of real online forum posts in thousands of courses offered at a large public university in the United States, available through our existing research partnership.

IRB approval and data use agreements have been in place for the project. Complying with the Family Educational Rights and Privacy Act (FERPA), all personally identifiable information in the raw data was removed or redacted prior to researcher access, and the dataset cannot be made publicly available. Nonetheless, we release our code for data generation and classification at https://github.com/Columbia-NLP-Lab/VTA_qa_evaluation.

We construct a classifier using GPT-4o-mini to categorize discussion forum posts into five categories, one of which is academic questions (see more details about this simple classifier in Appendix J). While our pedagogical framework is designed to be generalizable across various forms of student engagement, we selected academic questions as our starting point in this work, as this allows us to assess the foundational pedagogical capabilities of VTAs. Note that we aim to identify academic questions with high precision. Thus, the overall question classification accuracy is not our primary concern. The original set of data consists of 600k forum posts, from which we select the first 60k for classification. Additionally, we filtered for posts that are the beginning of their post threads.

For this study, we focus on a diverse sample set of 300 academic questions from 116 students. These questions offer high coverage of academic disciplines, spanning 85 courses and 40 departments, from criminal law to transportation systems.

4.2 Example VTA Response Generation

We select Llama-3-70B-Instruct (Grattafiori et al., 2024) as our primary LLM to generate simulated VTA responses because we want instances of pedagogical attempts with varying degrees of success according to our framework. Due to the lack of access to course materials, we prompt the model to respond to student questions using its inherent knowledge base instead of taking retrieval-augmented generation approaches.

Two sets of responses are generated: with and without forum-level context.

Context-Free: For each academic question, the default VTA prompt includes: (1) Course information, when available. We scraped the course catalog from the university to obtain course names and descriptions. (2) Description of the discussion topic. (3) Content of the student’s forum post. Mock example data instances are in Table 1.

We rewrite the first four levels of our pedagogical framework into goals and incorporate them into

Course Information	Discussion Topic	Forum Post Content
<i>PHYSICS 103. Intro. to Astronomy.</i> The history of Astronomy as a field, early...	Question Week 6 Use this message board to only post questions relevant to course...	can someone explain how redshift works and what it tells us about galaxies moving away from us? also, how do exoplanets get detected with the transit method? appreciate any explanations!
<i>MGMT 157. Taxation of Individuals.</i> The fundamentals of federal...	INTRO - Dependents: Items from Dependents Module ...	If two individuals are both helping support someone, like a parent, how to figure out who gets to claim the parent?

Table 1: Mock example student discussion forum posts in our dataset. These examples are created to be similar to the real examples to preserve student privacy.



Student

Hi all, do you know what is the difference between opportunity cost and differential cost and how to distinguish them?

Great question! Understanding the difference between opportunity cost and differential cost is important in economics and decision-making.

Opportunity Cost
Opportunity cost refers to the value of the next best alternative that is foregone when a choice is made. It represents the benefits you could have received by taking an alternative action. For example, if you decide to spend time studying for an exam instead of going out with friends, the opportunity cost is the enjoyment and experiences you miss out on by not going out.

Differential Cost
Differential cost, on the other hand, refers to the difference in total cost that will result from choosing one alternative over another. It is the additional or reduced cost that will occur as a result of a specific decision. For instance, if you are deciding between two projects and one project costs 10,000 while the other costs 15,000, the differential cost of choosing the more expensive project is \$5,000.

Key Distinctions

- Nature:**
 - Opportunity Cost:** Focuses on the value of the next best alternative.
 - Differential Cost:** Focuses on the difference in costs between two alternatives.
- Application:**
 - Opportunity Cost:** Used in decision-making to evaluate the trade-offs of different choices.
 - Differential Cost:** Used in cost analysis to determine the financial impact of choosing one option over another.
- Measurement:**
 - Opportunity Cost:** Often subjective and can be difficult to quantify, as it involves estimating the value of benefits lost.
 - Differential Cost:** Objective and quantifiable, as it involves actual cost figures.

Example
Imagine you have 1,000 to invest. You can either invest in Stock A, which is expected to yield a return of 200, or Stock B, which is expected to yield a return of \$300.
Opportunity Cost: If you choose Stock A, the opportunity cost is the \$300 return you would have earned from Stock B.
Differential Cost: If Stock A has a cost of 1,200, the differential cost of choosing Stock B over Stock A is \$200.

I hope this helps clarify the differences! If you have any more questions or need further examples, feel free to ask!



Virtual Teaching Asst.

Dimension	Rating
Clarify Misunderstandings and Confusions	Strong (2)
Deepen Disciplinary Understanding	Weak (1)
Foster Higher-Order Thinking	Not Present (0)
Enhance Metacognitive Awareness	Weak (1)
Foster Collaborative Knowledge Construction	N/A

Figure 1: A mock student post coupled with a VTA response generated using the system from Kweon et al. (2025), rated along the pedagogical dimensions from our framework.

the default prompt to define desirable behaviors. We exclude the fifth level here, since there is no peer information available.

Forum-Level Context: To provide this context, we retrieve the top ten most similar student posts (excluding the post to be responded to) using OpenAI’s text-embedding-3-small embedding. When there is a discussion topic associated with the forum post, we select contextual forum posts under the same discussion topic; otherwise, from the entire course. All relevant posts are labeled with consecutive numbers, so that the VTA can refer to the posts as "Similar Post #1", etc.

We include the goal-style rewrite of Level 5 in this prompt version. See the prompts in Appendix C and example generations in Appendix I. We focus on simulated VTA responses in this work to ensure comparison fairness between the different

generation conditions. It is important to note, however, that our framework is already applicable to current VTA systems. We include an example in Figure 1 with a mock student discussion forum post, the corresponding VTA response, and the assessments using our framework.

4.3 Annotation Procedure

Given the pedagogical knowledge and background required for accurate ratings, we recruit two volunteer annotators who are pursuing graduate degrees in education, with specializations in learning sciences and pedagogy. Before the annotation, one of the authors conducts a training session with the two annotators to clarify the coding rubric and the definitions of each rating level. During the session, the annotators independently label a set of 20 sample VTA responses and discuss any disagree-

ments or ambiguities in their interpretations. The rubric descriptions are revised to enhance clarity for subsequent annotations.

Then, two annotators independently rate the full set of VTA responses. At the 80-response milestone, the author conducts an inter-rater agreement check, during which the annotators achieve an intraclass correlation coefficient (ICC) of 0.81, indicating a fine agreement and minimal misunderstanding. Upon completion of the full annotation, the final ICC was 0.78, indicating good reliability (Koo and Li, 2016).

Next, we compare the annotations from two annotators. We first identify items with a rating discrepancy greater than 1 (e.g., one annotator rates an item as 0 while the other rates it as 2, suggesting potentially substantive disagreement), which occurs in 1.8% of our sample. For these cases, one of the authors discusses with both annotators, and the final score for each item is determined by majority agreement. Items with a rating discrepancy of exactly 1, likely reflecting random rating variance, occur in 11.2% of our sample. These cases are divided equally among the author and the two annotators for a second round of review to determine the final score. The distribution of the annotated scores across each level can be found in Table 7 and Table 6 in the Appendix.

5 Automated Pedagogical Evaluations of VTAs

We explore automatic approaches for pedagogical assessment using LLMs to address practical challenges in using expert annotators and to investigate the claim from Maurya et al. (2024) about the limitations of LLM evaluators in education.

The labeling process in Section 4.3 yielded 600 unique annotated post-response pairs. We randomly sample 150 context-free and 150 contextual pairs as our training set, consisting of student forum post information, the corresponding VTA responses generated by Llama-3, and ratings on all five levels. Note that the annotators do not label the context-free instances for Level 5; as a result, all levels but Level 5 have 300 training pairs. The rest of the data constitutes a held-out test set. We refer to the training set as Llama3-Train and the test set as Llama3-Test for the rest of the paper.

5.1 Prompt-based Classifiers

We experiment with GPT-4o-mini and GPT-4o for zero-shot and prompt-optimized classification. We implement the classifiers as language model programs using the DSPy framework (Khattab et al., 2023) for compatibility with prompt optimization (Appendix K). The classifier program receives (1) the course information, (2) the discussion topic, (3) the forum post, (4) the VTA response, and (5) the rubric for **a specific pedagogical level**, and then **outputs the scoring for that specific level**.

For prompt optimization, we leverage DSPy’s SIMBA prompt optimizer*. SIMBA stands for Stochastic Introspective Mini-Batch Ascent. This prompt optimizer accepts arbitrary DSPy programs and proceeds in a sequence of mini-batches, seeking to craft incremental improvements as general rules or few-shot examples to the prompts. Improvements are measured using custom metrics, and in our case, we employ classification accuracy.

Prompt optimization yields significant performance gains even with as few as 150 examples. This efficiency allowed us to experiment with two strategies on the Llama3-Train dataset when optimizing prompts: (1) using all available annotations per post-response pair (four to five instances), and (2) using only annotations for specific pedagogical levels (one instance per pair).

5.2 Open-Weight Classifiers

To alleviate potential privacy concerns when using API-based frontier LLMs to classify forum posts, we explore open-weight models hosted locally. Unfortunately, our human-annotated dataset is not sufficient to fine-tune these models. Therefore, we perform data synthesis using Llama3-Train.

We use GPT-4.1-nano to generate pairs of student forum posts and VTA responses, each aligned with a specific pedagogical level score. To ensure class distribution balance, we aim to have 300 examples for every score within each pedagogical level. Rubrics are used to define desired behavioral characteristics of VTA responses for each level-score combination. We include both the requirements and sampled in-context examples for generation (detailed description in Appendix E).

The above process yields approximately 1200 total examples (synthetic and non-synthetic) per pedagogical level. We split these examples into

*<https://github.com/stanfordnlp/dspy/blob/main/dspy/teleprompt/simba.py>

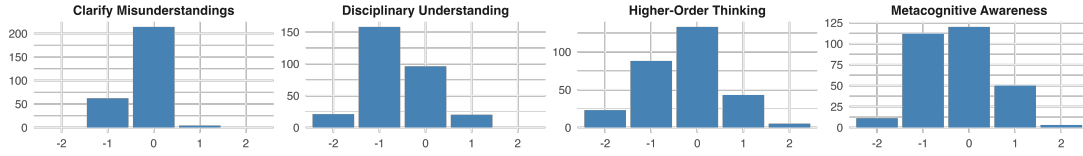


Figure 2: Distribution of score differences (with context — without context) across four pedagogical levels.

training and validation sets using an 85:15 ratio.

We fine-tune Qwen3-4B, Qwen3-8B (Yang et al., 2025), and ModernBERT-base (Warner et al., 2024) on synthetic data (detailed in Appendix F), training **one classifier per pedagogical level** due to the limited generalization of smaller models. Inputs include (1) the student forum post, (2) the VTA response, and (3) the level-specific rubric. Qwen models directly output their ratings, while ModernBERT performs multi-class classification.

5.3 Validation

Both prompt-based and open-weight classifiers are evaluated using Llama3-Test. Additionally, to assess generalizability across different LLM-powered VTAs and forum datasets, we synthesize additional test sets, New-LLM-Test and Llama3-MOOC.

Synthesizing New-LLM-Test: For evaluating cross-model generalizability, we employ the same VTA response generation scheme as detailed in Section 4.2, using context-free and contextual prompts. GPT-4o, GPT-4.1, and Qwen3-32B are used to generate six sets of 30 responses (three models, 30 randomly sampled academic questions, and context-free and contextual variants).

Synthesizing Llama3-MOOC: We leverage the Stanford MOOC dataset, a collection of 30,000 human-annotated MOOC forum posts (Agrawal et al., 2015). One of the authors manually went through the student posts labeled as questions and identified 104 academic questions among them. We generate VTA responses using Llama-3-70B-Instruct. See the prompt for this in Appendix C.

6 Results

6.1 Human Annotation

To examine whether providing contextual information influences the pedagogical quality of generated responses from Llama-3-70B-Instruct, we compare expert annotation scores between the two conditions. Figure 2 records the score differences.

Disciplinary understanding (Level 2) experiences the steepest decline with context, with 59.7% of responses decreasing by 1 or 2 points.

Similar patterns are observed in the levels of higher-order thinking (Level 3) and metacognitive awareness (Level 4), which show declines of 37% and 41%, respectively. We observe that context-free responses focus more on deep engagement with the original post, extending its ideas effectively. In contrast, adding forum-level context often shifts Llama-3-70B-Instruct’s focus to linking similar posts, resulting in broader but more superficial instructional guidance.

The level of clarification of misunderstandings (Level 1) appears to be the least affected by the inclusion of context. Qualitative analysis suggests that context has little impact on how well a response addresses confusion. Instead, the effectiveness of clarification seems to be more dependent on the clarity of the original question.

6.2 Performance on Llama3-Test

We report the performance of the prompting-based and fine-tuned classifiers on Llama3-Test in Tables 2 and 3, respectively. In Table 2, the suffix “-Gen” indicates that the prompt optimization is performed on all of Llama3-Train, while the suffix “-Lvl” means that optimization only uses annotations for the specific pedagogical level. Note that SIMBA fails to identify a better-than-default prompt for level-specific GPT-4o-mini optimization, hence we do not include that result.

Both prompt-based and fine-tuned classifiers **perform decently at lower pedagogical levels**, but their **performance declines as the level increases**. Model differences are also more pronounced at higher levels: for example, GPT-4o-mini can outperform GPT-4o at lower levels, but GPT-4o excels at Levels 4 and 5. Similarly, fine-tuned Qwen models outperform frontier LLMs below Level 4 but not above.

Prompt optimization can be effective for improving model performance on pedagogical assessment, especially when level-specific annotations are used. We see that for Level 4, prompt optimization with level-specific training data leads to 28.4% F-1 score improvement in GPT-4o. At

Classifier	Level 1	Level 2	Level 3	Level 4	Level 5	Average
GPT-4o-mini	85.6 / 87.5	52.9 / 61.9	40.8 / 48.8	49.8 / 50.0	32.7 / 38.7	52.4 / 57.4
GPT-4o-mini-Gen.	84.0 / 85.5	53.4 / 55.2	42.8 / 48.8	53.6 / 54.3	37.5 / 42.7	54.3 / 57.3
GPT-4o-mini-Lvl.	–	62.6 / 64.9	53.5 / 57.6	57.0 / 61.7	42.8 / 49.3	60.3 / 64.2
GPT-4o	84.2 / 84.5	55.9 / 57.5	47.1 / 47.1	30.4 / 31.3	41.8 / 41.3	51.9 / 52.3
GPT-4o-Gen.	85.1 / 84.5	48.4 / 47.8	48.6 / 50.2	56.3 / 55.0	44.5 / 46.0	56.6 / 56.7
GPT-4o-Lvl.	83.7 / 86.8	52.7 / 53.9	47.2 / 50.2	58.8 / 58.3	38.7 / 48.0	56.2 / 59.4

Table 2: Performance of prompting-based classifiers on Llama3-Test. Scores are in percentages. Each cell shows **F-1/Acc.**. Bolded values indicate the best performance in each column.

Classifier	Level 1	Level 2	Level 3	Level 4	Level 5	Average
Qwen3-4B	83.1 / 87.5	50.2 / 63.6	13.3 / 21.3	34.9 / 49.5	17.0 / 33.3	39.7 / 51.0
FT-Qwen3-4B	83.6 / 86.5	54.4 / 63.6	55.0 / 53.7	45.0 / 52.2	35.3 / 40.7	54.7 / 59.3
Qwen3-8B	83.6 / 86.5	51.3 / 63.2	48.9 / 48.3	40.4 / 50.5	32.0 / 39.3	51.2 / 57.6
FT-Qwen3-8B	84.2 / 86.8	59.6 / 64.2	65.0 / 69.7	45.3 / 51.5	35.9 / 40.7	58.0 / 62.6
FT-ModernBERT	81.9 / 84.8	62.4 / 63.2	46.8 / 49.2	59.8 / 68.0	40.0 / 49.3	58.2 / 62.9

Table 3: Performance of open-weight, fine-tuned classifiers on Llama3-Test. Rows without the “FT” prefix represent zero-shot classification results. Each cell shows **F-1/Acc.**.

higher pedagogical levels, prompt optimization consistently outperforms zero-shot LLM classifications, as a result of SIMBA adding beneficial few-shot examples.

Fine-tuning on our synthetic data can boost model F-1 significantly, by as much as 41.7%. The fine-tuned ModernBERT models maintain competitive performance throughout most pedagogical levels compared to prompt-based classifiers. These results further validate the utility of synthetic data. Note that our data is synthesized with GPT-4.1-nano, which is not the most capable frontier model measured by MMLU[†]. It is possible that data generated from more powerful models can yield even more prominent performance gains.

6.3 Generalizability Analysis

Given the consistent performance of our classifiers at Level 1, we focus our generalizability analysis on this level. One of the authors and one of the annotators annotated New-LLM-Test and Llama3-MOOC with random checks to ensure inter-rater reliability.

We report classifier performance on these datasets in Table 4 and Table 5. A crucial difference between Llama-3-70B-Instruct and the other LLMs is the increased use of Markdown formatting in responses by the other LLMs. To as-

sess whether such formatting influences classifier predictions, we remove Markdown formatting in out-of-distribution generations (**X MD** columns).

Overall, there is a **decrease in classifier performance in both test datasets**, with over-scoring as a common error. Interestingly, stripping Markdown formatting can improve classifier performance on New-LLM-Test but reduce it on Llama3-MOOC, indicating **sensitivity to formatting cues**, which is a crucial consideration for classifier robustness.

Fine-tuned classifiers match the performance of LLMs on Llama3-MOOC, showing robustness when tested on responses from the same model. In contrast, **prompt-based LLM classifiers are more generalizable across response sources**. Notably, GPT-4o-Gen performs best on out-of-distribution examples, implying that **prompt optimization across all pedagogical levels can enable GPT-4o to better model pedagogical reasoning**, resulting in improved generalization.

7 Discussion

By analyzing VTA responses using our pedagogical framework and automating assessment with LLMs, we highlight key insights into the development and evaluation of VTAs in ALEs.

We inspect the consequences of incorporating peer forum posts to aid collaborative knowledge

[†]<https://openai.com/index/gpt-4-1/>

Classifier	Original		X MD	
	F-1	Acc.	F-1	Acc.
GPT-4o-mini	63.1	73.3	64.0	73.3
GPT-4o-mini-Gen.	64.1	73.3	63.8	72.8
GPT-4o	63.4	71.7	64.5	71.7
GPT-4o-Gen.	66.5	72.8	68.3	73.3
GPT-4o-Lvl.	62.3	72.8	62.3	72.8
FT-Qwen3-4B	46.8	59.5	62.3	71.6
FT-Qwen3-8B	48.4	61.0	64.0	72.8
FT-ModernBERT	57.2	60.6	60.6	63.3

Table 4: Performance of all of our classifiers on New-LLM-Test for Level 1.

Classifier	Original		X MD	
	F-1	Acc.	F-1	Acc.
GPT-4o-mini	67.1	75.5	65.7	70.4
GPT-4o-mini-Gen.	66.7	66.3	69.5	69.4
GPT-4o	67.5	71.4	65.7	70.4
GPT-4o-Gen.	72.8	76.1	70.1	74.5
GPT-4o-Lvl.	67.4	74.5	67.0	74.5
FT-Qwen3-4B	72.2	76.7	70.8	76.7
FT-Qwen3-8B	72.5	80.0	70.7	76.7
FT-ModernBERT	68.4	74.5	63.4	65.3

Table 5: Performance of all of our classifiers on Llama3-MOOC.

formation, under a simple VTA formulation with no course-specific materials involved. Our expert annotations suggest that including forum-level contexts can be detrimental to VTA performance in other pedagogical levels. This can be a result of the “lost in the middle” phenomenon (Liu et al., 2023), where language models fixate on information presented at the beginning or the end of a long input. Since Level 5 appears at the end of our prompt, models may over-prioritize it. This underscores the value of our pedagogical framework, which makes such degradations in performance detectable. This result pertains only to Llama-3-70B-Instruct, although it is plausible that other LLMs would exhibit similar patterns.

Prior work (Maurya et al., 2024) argues that LLMs may be unreliable annotators for pedagogical assessment. We explore this claim in the context of assessing VTAs in ALEs, and our findings largely support this concern. While prompt optimization and supervised fine-tuning on synthetic

data successfully enhance accuracy, the models we examine still struggle with higher-level pedagogical evaluations. This could reflect a broader lack of pedagogical awareness during the training of state-of-the-art models (Macina et al., 2023; Jurenka et al., 2024), which impacts both prompt-based classification and data synthesis.

Another limitation is that all our training data (and in-context examples for data generation) come from a single LLM, which might restrict generalization. Still, the fine-tuned classifiers match and outperform frontier LLM performance on Llama3-MOOC, suggesting strong potential. We expect that introducing more diverse VTA response styles to our training data will improve classifier robustness. We leave expanding expert annotations on a wider range of VTA outputs as future work.

One interesting finding is that Markdown formatting from VTA responses affects classifier performance. We hypothesize that removing Markdown formatting brings the VTA responses closer to the output distribution of Llama-3-70B-Instruct for the open-weight classifiers, while mitigating self-selection biases in the prompt-based classifiers (Panickssery et al., 2024). This is notable because it emphasizes a critical difference between human and model evaluators; humans are capable of disentangling stylistic features and pedagogical content, whereas current LLMs experience challenges.

8 Conclusion

We introduce the first evaluation framework for the pedagogical capabilities of teaching assistants in asynchronous online forums. Upon expert labeling of forum VTA responses using our rubric, we utilize these annotations to build and test automatic classifiers using language models. Our results indicate that while certain approaches can benefit classifier performance, language models continue to face challenges with higher-level, nuanced classifications and generalization across response types. Our novel framework paves the way for rigorous comparisons between different VTA systems, and we caution researchers when using LLMs for complex pedagogical evaluations.

Limitations

Our pedagogical evaluation framework is grounded in educational theory, but its practical utility for instructors remains unclear. Future work will include usability studies with university instructors

to better validate our framework’s relevance and effectiveness in the real world.

Although the number of annotations is sufficient for prompt optimization in this work, a larger annotated dataset would always be desirable, as it can strengthen model training and evaluation. As an initial exploration into pedagogical assessment of VTAs, our work provides a solid foundation for future researchers in this domain. We encourage researchers to reuse our annotation framework to construct more training data for classifiers. The number of annotators was also limited; however, we achieved high inter-rater reliability. Additionally, recruiting expert annotators remains challenging in this domain, and we also do not provide financial compensation to our annotators due to budgetary concerns. Future work can look into expanding our annotator pool to enable annotating more examples.

This work focuses on simple VTA formulations that rely on models’ internal knowledge to answer academic questions, as we did not have access to vetted course materials. In contrast, existing VTAs often use retrieval-augmented generation (RAG) to obtain instructor-approved reference content. Studying such systems would offer valuable comparisons, but these implementations are typically not publicly available.

Lastly, our framework evaluates the pedagogical quality of VTA responses, not their factual accuracy or potential for hallucination. However, accurate and reliable content is critical in educational contexts. To be suitable for deployment, our pedagogical framework must be paired with methods that assess and promote response accuracy.

Ethical Considerations

Asynchronous online forum posts made by students contain sensitive personal information such as their names and their grades (e.g. "I was wondering why I just got a B on this assignment"). Our suite of classifiers is designed to have access to the original student forum posts to enhance classification accuracy. While we redacted private information when developing our classifiers, it is important to recognize and minimize the potential risks of data leakage if our classifiers are adopted for real-world use. The most capable models we tested are closed-source and are only available through API access. Companies can train on the user data, and there can be attackers hacking into the data-hosting servers and

gaining access to sensitive student data. In order to protect student privacy, future work should prioritize using open-source models securely hosted on local infrastructure.

Acknowledgment

We would like to thank our annotators, Chenxi Shi and Xin Guan, for their insightful discussions and revision suggestions on the description of the evaluation framework, as well as for their careful and rigorous annotation work on this study.

References

- Akshay Agrawal, Jagadish Venkatraman, Shane Leonard, and Andreas Paepcke. 2015. Youedu: Addressing confusion in mooc discussion forums by recommending instructional video clips.
- Mete Akcaoglu and Eunbae Lee. 2016. Increasing social presence in online learning through small group discussions. *The international review of research in open and distributed learning*, 17(3).
- Zehra Akyol and D Randy Garrison. 2011. Assessing metacognition in an online community of inquiry. *The Internet and higher education*, 14(3):183–190.
- Erfan Al-Hossami, Razvan Bunescu, Justin Smith, and Ryan Teehan. 2024. Can language models employ the socratic method? experiments with code debugging. In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*, pages 53–59.
- Diane Audras, Alex Zhao, Chuck Isgar, and Yucheng Tang. 2022. Virtual teaching assistants: A survey of a novel teaching technology. *International Journal of Chinese Education*, 11(2):2212585X221121674.
- Ryan Shaun Baker, Maciej Pankiewicz, and Xiner Liu. 2024. A step towards adaptive online learning: Exploring the role of gpt as virtual teaching assistants in online education.
- Sara Brierton, Elizabeth Wilson, Mark Kistler, Jim Flowers, and David Jones. 2016. A comparison of higher order thinking skills demonstrated in synchronous and asynchronous online college discussion posts. *Nacta Journal*, 60(1):14–21.
- Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.
- Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, and 1 others. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

- Alexis Chevalier, Jiayi Geng, Alexander Wettig, Howard Chen, Sebastian Mizera, Toni Annala, Max Jameson Aragon, Arturo Rodríguez Fanlo, Simon Frieder, Simon Machado, and 1 others. 2024. Language models as science tutors. *arXiv preprint arXiv:2402.11111*.
- Micheline TH Chi and Ruth Wylie. 2014. The icap framework: Linking cognitive engagement to active learning outcomes. *Educational psychologist*, 49(4):219–243.
- Ikseon Choi, Susan M Land, and Alfred J Turgeon. 2005. Scaffolding peer-questioning strategies to facilitate metacognition during online small group discussion. *Instructional science*, 33(5):483–511.
- Rajashree Dahal, Greg Murray, Robin Chataut, Mohamed Hefaida, Anurag Srivastava, and Prashna Gyawali. 2025. Autota: A dynamic intent-based virtual teaching assistant for students using open source llms. *IEEE Access*.
- Yuhao Dan, Zhikai Lei, Yiyang Gu, Yong Li, Jianghao Yin, Jiaju Lin, Linhao Ye, Zhiyan Tie, Yougen Zhou, Yilei Wang, and 1 others. 2023. Educhat: A large-scale language model-based chatbot system for intelligent education. *arXiv preprint arXiv:2308.02773*.
- Dorottya Demszky, Jing Liu, Zid Mancenido, Julie Cohen, Heather Hill, Dan Jurafsky, and Tatsunori Hashimoto. 2021. Measuring conversational uptake: A case study on student-teacher interactions. *arXiv preprint arXiv:2106.03873*.
- Charlene A Dykman and 1 others. 2008. Part two-teaching online versus teaching conventionally. *Journal of Information Systems Education*, 19(2):157–164.
- Randi A Engle and Faith R Conant. 2002. Guiding principles for fostering productive disciplinary engagement: Explaining an emergent argument in a community of learners classroom. *Cognition and instruction*, 20(4):399–483.
- Irish A Farley and Nicholas C Burbules. 2022. Online education viewed through an equity lens: Promoting engagement and success for all learners. *Review of Education*, 10(3):e3367.
- Catherine Twomey Fosnot. 2013. *Constructivism: Theory, perspectives, and practice*. Teachers College Press.
- Ashok K Goel and Lalith Polepeddi. 2018. Jill watson: A virtual teaching assistant for online education. In *Learning engineering for online education*, pages 120–143. Routledge.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Jieun Han, Haneul Yoo, Junho Myung, Minsun Kim, Hyunseung Lim, Yoonsu Kim, Tak Yeon Lee, Hwang-jung Hong, Juho Kim, So-Yeon Ahn, and 1 others. 2023. Llm-as-a-tutor in efl writing education: Focusing on evaluation of student-llm interaction. *arXiv preprint arXiv:2310.05191*.
- Khe Foon Hew, Wing Sum Cheung, and Connie Siew Ling Ng. 2010. Student contribution in asynchronous online discussion: A review of the research and empirical exploration. *Instructional science*, 38:571–606.
- Yann Hicke, Anmol Agarwal, Qianou Ma, and Paul Denny. 2023. Ai-ta: Towards an intelligent question-answer teaching assistant using open-source llms. *arXiv preprint arXiv:2311.02775*.
- Srećko Joksimović, Dragan Gašević, Vitomir Kovanović, Bernhard E Riecke, and Marek Hatala. 2015. Social presence in online discussions as a process predictor of academic performance. *Journal of Computer Assisted Learning*, 31(6):638–654.
- Irina Jurenka, Markus Kunesch, Kevin R McKee, Daniel Gillick, Shaojian Zhu, Sara Wiltberger, Shubham Milind Phal, Katherine Hermann, Daniel Kasenberg, Avishkar Bhoopchand, and 1 others. 2024. Towards responsible development of generative ai for education: An evaluation-driven approach. *arXiv preprint arXiv:2407.12687*.
- Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T Joshi, Hanna Moazam, and 1 others. 2023. Dspy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*.
- Seungyoon Kim and Seungone Kim. 2024. Can language models evaluate human written text? case study on korean student writing for education. *arXiv preprint arXiv:2407.17022*.
- Terry K Koo and Mae Y Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163.
- Sunjun Kweon, Sooyohn Nam, Hyunseung Lim, Hwang-jung Hong, and Edward Choi. 2025. A large-scale real-world evaluation of llm-based virtual teaching assistant. *arXiv preprint arXiv:2506.17363*.
- Mason Laney and Prasun Dewan. 2024. Human-ai collaboration in a student discussion forum. In *Companion Proceedings of the 29th International Conference on Intelligent User Interfaces*, pages 74–77.
- James P Lantolf. 2000. Introducing sociocultural theory. *Sociocultural theory and second language learning*, 1:1–26.

- Changyoon Lee, Junho Myung, Jieun Han, Jiho Jin, and Alice Oh. 2023. Learning from teaching assistants to program with subgoals: Exploring the potential for ai teaching assistants. *arXiv preprint arXiv:2309.10419*.
- Yu Li, Shang Qu, Jili Shen, Shangchao Min, and Zhou Yu. 2023. Curriculum-driven edubot: A framework for developing language learning chatbots through synthesizing conversational data. *arXiv preprint arXiv:2309.16804*.
- Margarita Limón. 2001. On the cognitive conflict as an instructional strategy for conceptual change: A critical appraisal. *Learning and instruction*, 11(4-5):357–380.
- Xi Lin and Li Gao. 2020. Students’ sense of community and perspectives of taking synchronous and asynchronous online courses. *Asian Journal of Distance Education*, 15(1):169–179.
- Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.
- Xiner Liu, Maciej Pankiewicz, Tanvi Gupta, Zhongtian Huang, and Ryan S Baker. 2024. A step towards adaptive online learning: Exploring the role of gpt as virtual teaching assistants in online education. *Manuscript under review*.
- Xiner Liu, Andres Felipe Zambrano, Ryan S Baker, Amanda Barany, Jaclyn Ocumpaugh, Jiayi Zhang, Maciej Pankiewicz, Nidhi Nasir, and Zhanlan Wei. 2025. Qualitative coding with gpt-4: Where it works better. *Journal of Learning Analytics*, 12(1):169–185.
- Jakub Macina, Nico Daheim, Sankalan Pal 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. *arXiv preprint arXiv:2305.14536*.
- Kaushal Kumar Maurya, KV Srivatsa, Kseniia Petukhova, and Ekaterina Kochmar. 2024. Unifying ai tutor evaluation: An evaluation taxonomy for pedagogical ability assessment of llm-powered ai tutors. *arXiv preprint arXiv:2412.09416*.
- David McLoughlin and Jo Mynard. 2009. An analysis of higher order thinking in online discussions. *Innovations in Education and Teaching International*, 46(2):147–160.
- Katrina A Meyer. 2003. Face-to-face versus threaded discussions: The role of time and higher-order thinking. *Journal of asynchronous learning networks*, 7(3):55–65.
- Tova Michalsky, Michal Zion, and Zemira R Mevarech. 2007. Developing students’ metacognitive awareness in asynchronous learning networks in comparison to face-to-face discussion groups. *Journal of Educational Computing Research*, 36(4):395–424.
- Elizabeth Murphy. 2009. A framework for identifying and promoting metacognitive knowledge and control in online discussants. *Canadian Journal of Learning and Technology/La revue canadienne de l’apprentissage et de la technologie*, 34(2).
- Dip Nandi, Margaret Hamilton, and James Harland. 2012. Evaluating the quality of interaction in asynchronous discussion forums in fully online courses. *Distance education*, 33(1):5–30.
- Edeh Michael Onyema, Edeh Chinecherem Deborah, Alhuseen Omar Alsayed, Quadri Noorulhasan, and Sumaya Sanober. 2019. Online discussion forum as a tool for interactive learning and communication. *International Journal of Recent Technology and Engineering*, 8(4):4852–4859.
- Arjun Panickssery, Samuel Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. *Advances in Neural Information Processing Systems*, 37:68772–68802.
- Judith B Pena-Shaff and Craig Nicholls. 2004. Analyzing student interactions and meaning construction in computer bulletin board discussions. *Computers & Education*, 42(3):243–265.
- Shuying Qiao, Paul Denny, and Nasser Giacaman. 2025. Oversight in action: Experiences with instructor-moderated llm responses in an online discussion forum. In *Proceedings of the 27th Australasian Computing Education Conference*, pages 95–104.
- Srikrishna Raamadhurai, Ryan Baker, and Vikraman Poduval. 2019. Curio smartchat: a system for natural language question answering for self-paced k-12 learning. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 336–342.
- Ramteja Sajja, Yusuf Sermet, David Cwiertny, and Ibrahim Demir. 2023. Platform-independent and curriculum-oriented intelligent assistant for higher education. *International journal of educational technology in higher education*, 20(1):42.
- Anvit Sinha, Shruti Goyal, Zachary Sy, Rhianna Kuperus, Ethan Dickey, and Andres Bejarano. 2024. Boilertai: A platform for enhancing instruction using generative ai in educational forums. In *2024 IEEE Frontiers in Education Conference (FIE)*, pages 1–8. IEEE.
- Gerry Stahl. 2013. A model of collaborative knowledge-building. In *International conference of the learning sciences*, pages 70–77. Psychology Press.

- Zsuzsanna Szabo and Jonathan Schwartz. 2011. Learning methods for teacher education: The use of online discussions to improve critical thinking. *Technology, Pedagogy and Education*, 20(1):79–94.
- Anaïs Tack and Chris Piech. 2022. The ai teacher test: Measuring the pedagogical ability of blender and gpt-3 in educational dialogues. *arXiv preprint arXiv:2205.07540*.
- Karan Taneja, Pratyusha Maiti, Sandeep Kakar, Pranav Guruprasad, Sanjeev Rao, and Ashok K Goel. 2024. Jill watson: A virtual teaching assistant powered by chatgpt. In *International Conference on Artificial Intelligence in Education*, pages 324–337. Springer.
- Tamara Tate and Mark Warschauer. 2022. Equity in on-line learning. *Educational Psychologist*, 57(3):192–206.
- Justin Vasselli, Christopher Vasselli, Adam Nohejl, and Taro Watanabe. 2023. Naisteacher: A prompt and rerank approach to generating teacher utterances in educational dialogues. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 772–784.
- Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023a. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.
- Rose E Wang, Qingyang Zhang, Carly Robinson, Susanna Loeb, and Dorottya Demszky. 2023b. Bridging the novice-expert gap via models of decision-making: A case study on remediating math mistakes. *arXiv preprint arXiv:2310.10648*.
- Xu Wang, Miaomiao Wen, and Carolyn P Rosé. 2016. Towards triggering higher-order thinking behaviors in moocs. In *Proceedings of the sixth international conference on learning analytics & knowledge*, pages 398–407.
- Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, and 1 others. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. *arXiv preprint arXiv:2412.13663*.
- Lynette Watts. 2016. Synchronous and asynchronous communication in distance learning: A review of the literature. *Quarterly Review of Distance Education*, 17(1):23.
- Shou’ang Wei, Xinyun Wang, Shuzhen Bi, Jian Chen, Ruijia Li, Bo Jiang, Xin Lin, Min Zhang, Yu Song, BingDong Li, and 1 others. 2025. Elmes: An automated framework for evaluating large language models in educational scenarios. *arXiv preprint arXiv:2507.22947*.
- Zhen Xu, Xinjin Li, Yingqi Huan, Veronica Minaya, and Renzhe Yu. 2025. From course to skill: Evaluating large language model performance in curricular analytics. In *International Conference on Artificial Intelligence in Education*, pages 203–211. Springer.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.
- Ya-Ting C Yang, Timothy J Newby, and Robert L Bill. 2005. Using socratic questioning to promote critical thinking skills through asynchronous discussion forums in distance learning environments. *The american journal of distance education*, 19(3):163–181.
- Yuqin Yang, Kaicheng Yuan, Xueqi Feng, Xiuhua Li, and Jan van Aalst. 2022. Fostering low-achieving students’ productive disciplinary engagement through knowledge-building inquiry and reflective assessment. *British Journal of Educational Technology*, 53(6):1511–1529.
- JD Zamfirescu-Pereira, Laryn Qi, Björn Hartmann, John DeNero, and Narges Norouzi. 2025. 61a bot report: Ai assistants in cs1 save students homework time and reduce demands on staff.(now what?). In *Proceedings of the 56th ACM Technical Symposium on Computer Science Education V. 1*, pages 1309–1315.
- Xuanming Zhang, Zixun Chen, and Zhou Yu. 2024a. Prolex: A benchmark for language proficiency-oriented lexical substitution. *arXiv preprint arXiv:2401.11356*.
- Xuanming Zhang, Anthony Diaz, Zixun Chen, Qingyang Wu, Kun Qian, Erik Voss, and Zhou Yu. 2024b. Decor: Improving coherence in l2 english writing with a novel benchmark for incoherence detection, reasoning, and rewriting. *arXiv preprint arXiv:2406.19650*.
- Wenting Zou, Xiao Hu, Zilong Pan, Chenglu Li, Ying Cai, and Min Liu. 2021. Exploring the relationship between social presence and learners’ prestige in mooc discussion forums using automated content analysis and social network analysis. *Computers in Human Behavior*, 115:106582.

A Use of AI Assistants

We used AI Assistants to help polish sections of the paper after completing an initial draft of the paper.

B Scoring Rubrics

Please refer to Table 8 for the full rubric for our pedagogical framework and Table 9, Table 10, and Table 11 for scoring examples. All examples are rewritten based on the human annotation data for privacy consideration.

C Virtual Teaching Assistant Prompts

C.1 VTA Prompt without Context

""You are a virtual teaching assistant for a course called <COURSE_NAME>.
<COURSE_DESCRIPTION>

Respond to the discussion forum post for this course provided by one of the students. Please offer the response based on your existing knowledge base. Please add a general greeting in each response.

Please adhere to the following pedagogical goals:

1. Clarify Misunderstandings: Support knowledge acquisition by articulating questions, addressing confusion, and receiving clarifications from peers or instructors.
2. Deepen Disciplinary Understanding: Promote deeper engagement with core concepts and themes through elaboration, critical questioning, and interaction with diverse perspectives.
3. Develop Higher-Order Thinking: Cultivate critical thinking and reasoning skills by analyzing ideas, justifying positions, synthesizing information, and exploring alternative viewpoints.
4. Enhance Metacognitive Awareness: Strengthen self-regulated learning by reflecting on one's understanding, identifying gaps in knowledge, and evaluating the quality of reasoning.""

C.2 VTA Prompt with Forum-level Context

""You are a virtual teaching assistant for a course called <COURSE_NAME>.
<COURSE_DESCRIPTION>

You are responding to the following discussion forum post for this course provided by one of the students. This post is an initial post in the discussion.

Additionally, here are some other relevant posts from students on this same topic:
<SIMILAR_POSTS>

Please offer the response based on your existing knowledge base. Please add a general greeting in each response.

When appropriate, refer to insights or perspectives from these related posts to foster connections between student ideas.

Please adhere to the following pedagogical goals:

1. Clarify Misunderstandings: Support knowledge acquisition by articulating questions, addressing confusion, and receiving clarifications from peers or instructors.
2. Deepen Disciplinary Understanding: Promote deeper engagement with core concepts and themes through elaboration, critical questioning, and interaction with diverse perspectives.

3. Develop Higher-Order Thinking: Cultivate critical thinking and reasoning skills by analyzing ideas, justifying positions, synthesizing information, and exploring alternative viewpoints.
4. Enhance Metacognitive Awareness: Strengthen self-regulated learning by reflecting on one's understanding, identifying gaps in knowledge, and evaluating the quality of reasoning.
5. Foster Collaborative Knowledge Construction and Social Presence: Fosters peer interaction and collective learning by connecting diverse student perspectives, encouraging the exchange of ideas, and supporting collaborative knowledge construction, positioning the discussion forum as a shared space for dialogue and co-construction of understanding.""

C.3 VTA Prompt for Llama3-MOOC

""You are a virtual teaching assistant. Respond to the discussion forum post provided by one of the students. Please offer the response based on your existing knowledge base. Please add a general greeting in each response.

Please adhere to the following pedagogical goals:

1. Clarify Misunderstandings: Support knowledge acquisition by articulating questions, addressing confusion, and receiving clarifications from peers or instructors.
2. Deepen Disciplinary Understanding: Promote deeper engagement with core concepts and themes through elaboration, critical questioning, and interaction with diverse perspectives.
3. Develop Higher-Order Thinking: Cultivate critical thinking and reasoning skills by analyzing ideas, justifying positions, synthesizing information, and exploring alternative viewpoints.
4. Enhance Metacognitive Awareness: Strengthen self-regulated learning by reflecting on one's understanding, identifying gaps in knowledge, and evaluating the quality of reasoning.""

D Annotator Demographic Information

Both annotators are Asian females with an average age of 23. They are both masters students in the field of education enrolled in Teachers College.

E Synthetic Data Details

Our detailed process for generating synthetic data is as follows:

1. *Extract the description for each level-rating combination from the standardized rubric.* For example, the description for Level 3's "Not

Present (0)” would be “No effort is made to promote higher-order thinking.”

2. *Repeatedly sample in-context examples for data synthesis.* To avoid data contamination, we sample exclusively from annotations in Llama3-Train. For every call to GPT-4.1-nano, we use five examples to produce three new post-response pairs.

Furthermore, we prioritize sampling from the same level-rating combination if possible to promote synthetic data quality: if the number of post-response pairs in a specific combination exceeds our generation needs, we restrict sampling to that subset; otherwise, we sample from all post-response pairs.

3. *Synthesize new post-response pairs from in-context examples and VTA response requirements.* We include the description from the first step as the requirement for the VTA response in the newly synthesized post-response pair. For every level-rating combination, we synthesize up to 300 instances; if this level-rating combination already has more than 300 instances, no additional examples are created.

F Model Training Details

All models are trained for 10 epochs on an A100 GPU (80 GB GPU memory) with an initial learning rate of $2e-5$, and the best checkpoints are selected using validation loss. We set the maximum token limit for ModernBERT to be 8192.

See the prompts used to perform supervised fine-tuning for Qwen models in [Appendix G](#).

G Prompt for Supervised Fine-tuning of Qwen3-4B and Qwen3-8B

Given the following discussion forum post from a student and the response from a teaching assistant:

<POST_RESPONSE_PAIR>

Assess the response from the teaching assistant with the following rubric:

<RUBRIC>

Provide your rating directly as "0", "1", "2", or "NA".

H Summary Statistics of the Annotation Data

Refer to Tables 7 and 6 for detailed statistics on Llama-3-70B-Instruct annotations.

Table 6: Annotation score distribution without context inclusion

	NA	0	1	2
Clarify misun derstandings	0	0	19	281
Disciplinary Understanding	2	1	50	247
Higher-Order Thinking	3	22	131	144
Metacognitive Awareness	2	27	222	49

Table 7: Annotation score distribution with context inclusion

	NA	0	1	2
Clarify misun derstandings	0	0	67	221
Disciplinary Understanding	0	29	175	93
Higher-Order Thinking	0	58	137	100
Metacognitive Awareness	0	108	140	50
Collaborative Knowledge Construction	0	42	167	89

I Example Generations

We present example generations of different LLMs under context-free and forum-level context conditions for the same student forum post. To protect student privacy, we omit the original student post here. We additionally redact any names mentioned in the simulated VTA responses.

I.1 Llama-3-70B-Instruct, Context-Free

Hello everyone,

Thank you for sharing your thoughts and questions about Chapter 14, particularly regarding bipolar disorder. The statistic you mentioned is indeed alarming, and it's essential to understand the underlying factors that contribute to the increased risk of suicide in individuals with bipolar disorder.

Research suggests that the transition from a manic episode to a depressive episode, often referred to as a "crash" or "comedown," can be a particularly vulnerable time for individuals with bipolar disorder. Several factors may contribute to this increased risk:

1. **Mood instability**: The rapid shift from a manic to a depressive state can be overwhelming, leading to feelings of hopelessness, despair, and anxiety.
2. **Loss of control**: During a manic episode, individuals may feel invincible or in control. However, as the episode subsides, they may experience a sense of loss of control, leading to feelings of powerlessness and helplessness.
3. **Cognitive distortions**: Negative thought patterns, such as rumination and self-blame, can become more pronounced during the depressive phase, increasing the risk of suicidal ideation.
4. **Neurobiological factors**: The crash from a manic episode can be accompanied by changes in neurotransmitter levels, such as a decrease in dopamine and serotonin, which can contribute to feelings of depression and hopelessness.
5. **Lack of coping skills**: Individuals with bipolar disorder may struggle with developing effective coping strategies to manage the transition from a manic to a depressive episode, increasing their vulnerability to suicidal thoughts.

It's essential to note that these factors can vary from person to person, and not everyone with bipolar disorder will experience suicidal ideation. However, it's crucial to acknowledge the increased risk and provide support and resources to individuals struggling with the disorder.

If anyone with bipolar disorder or experience with the condition would like to share their insights, I would be grateful to hear their perspectives. Additionally, I encourage everyone to consider the importance of seeking help and support for individuals struggling with mental health conditions.

Best regards,

I.2 Llama-3-70B-Instruct, Forum-Level Context

Hello [REDACTED],

Thank you for sharing your question about bipolar disorder. It's indeed a striking statistic that 20% of people with the illness die from suicide. Your question about what brings a person coming out of a manic episode to that level is a crucial one.

From what we've learned in Chapter 14, it's possible that the crash or comedown from a manic episode can lead to feelings of intense despair, hopelessness, and anxiety. This can be a particularly vulnerable time for individuals with bipolar disorder, as they may feel like they're losing control or experiencing a sense of emotional numbness.

[REDACTED]'s response in Similar Post #2 provides some insight into the intrusive nature of manic episodes, which can lead to self-destructive behavior. It's also important to note that individuals with bipolar disorder often experience a mix of symptoms, including depression, anxiety, and impulsivity, which can contribute to suicidal ideation.

It would be great to hear from others who may have personal experience with bipolar disorder or have insights into this topic. Additionally, it might be helpful to explore the role of therapy, support systems, and medication in managing symptoms and reducing the risk of suicide.

What do you think about the importance of early intervention and treatment in preventing suicidal ideation in individuals with bipolar disorder?

Best, [Your Name]

I.3 GPT-4.1, Context-Free

Hello! Thank you for your thoughtful post and for raising such an important question about bipolar disorder.

You're absolutely right - bipolar disorder carries a significant risk for suicide, and it's a topic that deserves careful attention and understanding. Your observation about suicides often occurring during the "comedown" from a manic episode is supported by research. Let's break down why this might happen:

****1. The Emotional Crash:****

After a manic episode, individuals often experience a sharp drop in mood, sometimes into a depressive episode. This transition can feel especially devastating because the contrast between the high energy, optimism, and grandiosity of mania and the hopelessness of depression is so stark. The sudden loss of energy and motivation, combined with feelings of guilt or regret about actions taken during mania, can be overwhelming.

****2. Increased Insight:****

During mania, people may not fully recognize the consequences of their actions. As they come down, they might gain insight into the problems caused during the manic phase (e.g., financial issues, damaged relationships), which can lead to intense feelings of shame, guilt, or hopelessness.

****3. Energy Levels:****

Interestingly, people in the depths of depression may lack the energy to act on suicidal thoughts. However, during the transition out of mania, energy levels may still be relatively high, but mood and outlook have plummeted. This combination can increase the risk of acting on suicidal impulses.

****4. Biological Factors:****

There are also neurobiological changes happening in the brain during these transitions, which can affect mood regulation and impulse control.

Your openness to hearing from those with lived experience is wonderful, as personal perspectives can add depth to our understanding. If anyone in the class feels comfortable sharing, their insights could be very valuable.

****Reflective Question:****

How do you think understanding these patterns might influence the way clinicians or loved ones support someone with bipolar disorder during these transitions? What kinds of interventions or supports do you think could help reduce risk during these vulnerable periods?

Let me know if you'd like more information on this topic or if you have other questions about Chapter 14!

I.4 GPT-4.1, Forum-Level Context

Hello! Thank you for sharing your thoughts and raising such an important and sensitive question about bipolar disorder. Your observation about the high risk of suicide, especially during the transition out of a manic episode, is both accurate and deeply concerning.

To help clarify, when someone with bipolar disorder is coming down from a manic episode (sometimes called the "post-manic crash"), they often experience a rapid and severe shift in mood. During mania, individuals may feel euphoric, energetic, and invincible, sometimes making risky decisions or acting impulsively. When the mania ends, the contrast can be stark - they may suddenly feel exhausted, ashamed, or overwhelmed by the consequences of their actions during the manic phase. This abrupt change can lead to intense feelings of hopelessness, guilt, or depression, which can increase the risk of suicidal thoughts or behaviors.

Your question also connects with some of the insights from other students in this discussion. For example, [REDACTED] (Similar post #2) mentioned how intrusive mania can be and how the phases of bipolar disorder differ from other disorders like borderline personality disorder. This highlights the unique challenges of mood cycling in bipolar disorder, which can make the emotional lows feel even more severe after a high.

It's also important to note that during the "comedown," individuals may regain enough energy and clarity to act on suicidal thoughts that may have been present during the depressive phase but were not acted upon due to lack of motivation or energy. This is one reason why the period following mania is particularly risky.

If you or anyone you know is struggling with these issues, it's crucial to seek support from mental health professionals. Understanding these patterns can help us be more compassionate and supportive toward those living with bipolar disorder.

Does this help clarify your question? Are there any other aspects of bipolar disorder or mood disorders you'd like to explore further? Your curiosity and willingness to engage with these difficult topics are really valuable for our collective learning.

I.5 GPT-4o, Context-Free

Hello!

It's great that you're engaging deeply with the material on bipolar disorder. The statistic you mentioned highlights the serious nature of the disorder and the importance of understanding its complexities.

To address your question, individuals with bipolar disorder can experience extreme mood swings, including manic, hypomanic, and depressive episodes. During a manic episode, a person might feel euphoric, have increased energy, and engage in risky behaviors. However, when transitioning out of a manic episode, individuals often experience a "crash" or a significant drop in mood, which can lead to feelings of depression, hopelessness, and despair.

This transition can be particularly challenging because the contrast between the high energy and optimism of mania and the low energy and negativity of depression can be stark and overwhelming. Additionally, during the comedown, individuals might reflect on the consequences of their actions during the manic phase, which can exacerbate feelings of guilt or shame.

It's important to note that each person's experience with bipolar disorder is unique, and various factors, including biological, psychological, and social influences, can contribute to these experiences. If anyone with personal experience wishes to share their insights, it could provide valuable perspectives.

If you have further questions or need clarification, feel free to ask. Understanding these nuances can help in developing empathy and effective support strategies for those affected by bipolar disorder.

Best regards!

I.6 GPT-4o, Forum-Level Context

Hello!

Your question about the transition from a manic episode to a depressive state in individuals with bipolar disorder is an important one. This transition can be particularly challenging and is often associated with a heightened risk of suicide. During a manic episode, individuals may experience elevated mood, increased energy, and a sense of invincibility. However, as they come down from this state, they may face a stark contrast in mood, leading to feelings of hopelessness, guilt, or worthlessness. This sudden shift can be overwhelming and may contribute to suicidal thoughts or behaviors.

It's also important to consider that during the comedown, individuals might reflect on the consequences of their actions during the manic phase, which can lead to regret or distress. Additionally, the biochemical changes in the brain during these transitions can exacerbate mood instability.

In the related posts, [REDACTED] mentions the intrusive nature of mania in bipolar I disorder, which can lead to self-destructive behavior. This aligns with the idea that the aftermath of such behavior can be difficult to cope with, potentially leading to suicidal ideation.

Understanding these dynamics is crucial for providing support and intervention. If you or someone you know is struggling, it's important to seek professional help. Mental health professionals can offer strategies and treatments to manage these transitions more effectively.

Feel free to continue the discussion or ask more questions if you have them!

Best regards.

I.7 Qwen3-32B, Context-Free

We remove the thinking process of Qwen3-32B.

Hello! Thank you for sharing such a thoughtful and important question. You're absolutely right to highlight the connection between bipolar disorder and suicide risk, particularly during the transition from mania to depression. Let's unpack this step by step:

1. **The "Comedown" Phase and Emotional Vulnerability:**

During a manic episode, individuals often feel invincible, hyper-focused, or euphoric, which can mask their awareness of distress. However, the crash after mania (often a depressive episode) can bring intense emotional pain. This phase may involve guilt, shame, or despair over the consequences of impulsive actions during mania (e.g., financial ruin, strained relationships). These feelings can feel insurmountable, increasing suicide risk.

2. **Biological and Neurochemical Factors:**

The transition between mood states in bipolar disorder is linked to dysregulation in neurotransmitters like serotonin and dopamine. This instability can amplify emotional reactivity and hopelessness during the depressive phase. Additionally, sleep deprivation during mania (a common symptom) can exacerbate irritability and impulsivity during the comedown.

3. **Mixed Episodes:**

Some individuals experience *mixed episodes*, where manic and depressive symptoms co-occur. This combination - e.g., feeling agitated, restless, and hopeless simultaneously - can heighten suicide risk, as the person may lack the energy to seek help but feel overwhelmed by their suffering.

4. **Stigma and Treatment Gaps:**

People with bipolar disorder may not seek help during mania (due to denial of illness) or during the depressive phase (due to low motivation or stigma). Without consistent treatment (e.g., mood stabilizers, therapy), the cycle of mania-depression can worsen, increasing risk.

Critical Question for Reflection: How might societal misconceptions about bipolar disorder (e.g., "they're just moody" or "mania is fun") contribute to underdiagnosis or undertreatment during critical phases?

If you or someone you know is struggling with these issues, please emphasize the importance of professional support. For further discussion, I'd be happy to explore how therapy (e.g., CBT, DBT) or medication management might mitigate these risks. Let me know if you'd like to dive deeper!

- Your Teaching Assistant

I.8 Qwen3-32B, Forum-Level Context

Hello [REDACTED],

Thank you for raising such an important and thought-provoking question about bipolar disorder and suicide risk. Your observation about the 20% suicide rate is striking, and your focus on the "comedown" from mania is a critical point. Let's unpack this together.

Understanding the Transition Phase

During a manic episode, individuals often experience heightened energy, impulsivity, and a distorted sense of reality (as noted in Similar Post #2). However, when the manic phase subsides, the emotional and physical toll can be overwhelming. This "comedown" may lead to a *mixed episode* (a blend of manic and depressive symptoms) or a rapid shift into depression. Key factors that might contribute to suicidal ideation during this phase include:

- Guilt/Shame:** After impulsive or self-destructive actions during mania (e.g., financial ruin, relationship damage), individuals may feel intense guilt or shame.
- Emotional Exhaustion:** The physical and mental strain of mania can leave individuals feeling drained and hopeless.
- Hopelessness:** The cyclical nature of bipolar disorder can create a sense of futility, especially if treatment is inconsistent or ineffective.
- Neurobiological Factors:** Fluctuations in brain chemistry (e.g., dopamine, serotonin) during mood transitions may heighten vulnerability to suicidal thoughts.

Connecting to Class Themes

Your question ties into broader discussions in Chapter 14 about the **cyclical nature of bipolar disorder** (as highlighted in Similar Post #4) and the challenges of distinguishing it from other disorders like borderline personality disorder. Unlike borderline personality disorder, which involves unstable relationships and identity issues (Similar Post #3), bipolar disorder is marked by distinct mood episodes. However, the transition between these episodes can blur these lines, making diagnosis and treatment complex.

Critical Thinking & Next Steps

- **Why might the post-manic phase be more dangerous than mania itself?** Consider how impulsivity during mania might lead to consequences (e.g., debt, isolation) that amplify despair later.
- **How does this relate to treatment strategies?** Effective management of bipolar disorder often involves stabilizing mood swings (e.g., with medications like...

J DSPy Language Model Program for Discussion Forum Question Classification

Listing 1: DSPy Language Model Program for Zero-shot Discussion Forum Post Classification

```
class ClassifyDiscussionForum(dspy.Signature):
    """Classify the discussion forum post into one of the following categories based on the content
    of the forum post and guidelines from the instructor for the post (if applicable):
    1. Academic Question - Questions about academic content posed to the teaching staff;
    2. Academic Discussion - Usually a discussion forum post required by an assignment, or any type of
    discussion that does not contain an obvious question;
    3. Logistics Question - Questions about course logistics;
    4. Logistics Discussion - Other types of posts about course logistics;
    5. Social - Discussion forum posts for social purposes."""
    post_content = dspy.InputField(desc="The content of the student's discussion forum post.")
    post_topic_content = dspy.InputField(desc="The guidelines from the instructor for this post.")
    post_classification = dspy.OutputField(desc="The classification for the post.")
```

K DSPy Language Model Program for the VTA Pedagogical Capability Classifier

Listing 2: DSPy Language Model Program for Zero-shot Prompt-based Classifiers

```
from typing import List, Literal

class SimpleClassifier(dspy.Signature):
    """Given a discussion forum post and the corresponding teaching assistant response, classify how
    well this response adheres to pedagogical goals specified with a rubric. Provide your
    classification as integers."""

    discussion_forum_post = dspy.InputField(desc="The original student forum post.")
    ta_response = dspy.InputField(desc="The corresponding teaching assistant response.")
    pedagogical_rubric = dspy.InputField(desc="The pedagogical rubric you should follow when
    evaluating the response.")
    discussion_topic = dspy.InputField(desc="Description and instruction for the specific discussion
    topic for students.")
    course_info = dspy.InputField(desc="Information about the course.")
    rating: Literal[0, 1, 2, "NA"] = dspy.OutputField()

class ComplexSelfReflector(dspy.Module):
    def __init__(self, callbacks=None):
        super().__init__(callbacks)
        self.classifier = dspy.ChainOfThought(SimpleClassifier)

    def forward(self, discussion_forum_post, ta_response, pedagogical_rubric, discussion_topic,
                course_info):
        og_rating = self.classifier(discussion_forum_post=discussion_forum_post,
                                    ta_response=ta_response,
                                    pedagogical_rubric=pedagogical_rubric,
                                    discussion_topic=discussion_topic,
                                    course_info=course_info).rating
```

```
return dspy.Prediction(rating=og_rating)
```

Table 8: Full rubric for evaluating VTA's pedagogical capability in asynchronous online forums.

Level	Description
Clarify misunderstandings	• Strong (2): Accurately identifies misunderstanding and confusion, provides a clear explanation using relevant content and examples. • Weak (1): Attempts to address the question or confusion, but the explanation is vague or lacks instructional clarity. • Not Present (0): Fails to recognize or respond to the misunderstanding or question. • Not Applicable (NA): No misunderstanding or question is present in the post.
Disciplinary Understanding	• Strong (2): Promotes deeper thinking and engagement with core disciplinary concepts by offering thought-provoking prompts, elaborations, or meaningful extensions. • Weak (1): Demonstrates intent to deepen disciplinary understanding, but through surface-level, or generic responses that lack meaningful connection to the post content. • Not Present (0): Makes no attempt to extend or deepen disciplinary understanding. • Not Applicable (NA): Deepening disciplinary understanding is irrelevant to the post context.
Higher-Order Thinking	• Strong (2): Promotes higher-order thinking through specific, content-grounded prompts or reasoning tasks that challenge students to analyze, evaluate, or reflect. • Weak (1): Attempts to promote higher-order thinking through general or loosely related prompts, but lacks depth, specificity, or clear alignment with the content. • Not Present (0): Makes no attempt to encourage higher-order thinking. • Not Applicable (NA): Higher-order thinking is not applicable given the context of the post.
Metacognitive Awareness	• Strong (2): Supports metacognitive awareness by using reflective prompts or strategies that help students assess their understanding, monitor their thinking, or make sense of their learning process in context. • Weak (1): Encourages reflection or self-monitoring, but uses vague or generic language that is not tied to the student's content or learning process. • Not Present (0): Makes no attempt to promote metacognition or reflection. • Not Applicable (NA): Metacognitive engagement is not relevant to the context of the post.
Collaborative Knowledge Construction	• Strong (2): Effectively fosters peer interaction by referencing specific student ideas, connecting diverse perspectives, or inviting further contributions in a personalized and contextually relevant manner. • Weak (1): Demonstrates intent to encourage interaction, but relies on general prompts or surface-level invitations without engaging specific content or peer input. • Not Present (0): Makes no attempt to promote peer interaction or build social connection. • Not Applicable (NA): Collaborative engagement is not relevant to the context of the student's post.

Table 9: Examples of responses at each rubric scoring level (Part 1).

Level	Student Posts and Response Examples
Clarify misunderstandings	Post: Hi all, do you know what is the difference between opportunity cost and differential cost and how to distinguish them? • Strong (2): Great question! Opportunity cost is the value of the next best alternative you give up when making a choice. For example, if you spend an hour studying instead of working a \$20/hour job, the opportunity cost is \$20. Differential cost, on the other hand, refers to the difference in cost between two alternatives. If Option A costs \$500 and Option B costs \$700, the differential cost is \$200. So, opportunity cost is about forgone benefits, while differential cost is about financial differences between choices. Both help in decision-making, but they address different aspects. • Weak (1): Opportunity cost is what you lose when you don't choose another option, and differential cost is just the difference between two costs. They're kind of related but not exactly the same. It's sometimes hard to tell the difference, but maybe opportunity cost is more like what you're missing out on, and differential cost is more like math. I guess you use both when comparing choices, but I'm not totally sure when each applies. Maybe they overlap in some situations, depending on what you're deciding between. • Not Present (0): Opportunity cost and differential cost both involve calculating how much something costs in total, so they can usually be used the same way. For example, if you're choosing between two jobs, you'd just compare their salaries directly and include all previous expenses like tuition or commuting costs, even if you already paid them. In that case, sunk cost is also part of the decision. There's not really a big difference between these costs—they all show up when comparing your options. You can just add them up to decide which one is better overall.
Disciplinary Understanding	Post: I'm trying to get a clearer understanding of circular satellite orbits—specifically, how their motion is described and how to calculate the total mechanical energy involved. In addition, I'm curious about how the concept of gravitational waves connects to phenomena like black holes. Could someone help explain these ideas in more detail? • Strong (2): In a circular orbit, the gravitational force equals the centripetal force, which gives the condition $v = \sqrt{\frac{GM}{r}}$. Using this, the mechanical energy (kinetic plus potential) comes out to $E = -\frac{GMm}{2r}$, the negative sign means the satellite is <i>bound</i> to the planet. Notice how the farther the orbit (larger r), the closer the total energy moves toward zero, meaning the satellite is less tightly bound. Now compare that to black holes: as two orbit each other, they lose orbital energy by radiating gravitational waves, shrinking their separation until merging. Can you see the parallel? • Weak (1): A satellite in circular orbit moves because gravity pulls it in while its velocity keeps it moving around, so the two balance. Its energy comes from adding potential and kinetic energy together, which depend on mass and distance. Gravitational waves are waves in space caused by black holes. These topics are important in physics, and it could help to look at examples of how energy and gravity connect in different cases. • Not Present (0): Circular orbits, energy, and gravitational waves are all interesting parts of space science. Satellites go around planets, and black holes are mysterious objects that produce strange effects. These are important topics in physics.

Table 10: Examples of responses at each rubric scoring level (Part 2).

Level	Student Posts and Response Examples
Higher-Order Thinking	Post: The textbook noted that Social Security benefits were once fully excluded from income, but now up to 85% may be taxed. Just wonder what led to this change? • Strong (2): That's an insightful observation. The shift occurred mainly because of tax policy changes in the 1980s and 1990s aimed at strengthening Social Security's long-term financing and addressing concerns about fairness in the tax system. Initially, Social Security benefits were fully excluded from taxation, but in 1983 Congress decided that up to 50% of benefits could be taxed for higher-income recipients, and this was expanded to up to 85% in 1993. Why do you think policymakers chose to tax only higher-income recipients instead of all beneficiaries? What does this reveal about the balance between revenue generation and protecting lower-income retirees? • Weak (1): The change happened because Congress passed laws in the 1980s and 1990s that made Social Security benefits partly taxable. This helped raise revenue and keep the program more sustainable. How do you think this compares to other tax changes you've read about? • Not Present (0): Social Security benefits became partly taxable starting in 1983, with the share rising to as much as 85% after 1993. This was the result of changes in federal tax law. • Not Applicable (NA): Higher-order thinking is not applicable given the context of the post.
Metacognitive Awareness	Post: I'm confused about how bunching itemized deductions works. It seems to involve shifting deductions into one year to exceed the standard deduction, then using the standard deduction the next year. How exactly can a taxpayer shift deductions—what does it mean to “accelerate” payments into the current year? • Strong (2): Great question. The idea is that a taxpayer can “bunch” deductions by paying certain expenses earlier than usual—like making a January mortgage payment in December or prepaying charitable donations—so they fall into the same tax year. Can you walk through an example in your own words to check if this makes sense? For instance, how would prepaying property taxes affect whether you use the standard or itemized deduction this year versus next year? • Weak (1): The taxpayer shifts deductions by paying some bills early, such as property taxes or donations, so they count in the current year. Think about how this changes the year's total, and consider if it helps you see why the timing matters. • Not Present (0): Taxpayers can shift deductions by paying certain expenses, like property taxes or charitable contributions, earlier so they fall into the same year and allow itemizing instead of taking the standard deduction.

Table 11: Examples of responses at each rubric scoring level (Part 3).

Level	Student Posts and Response Examples
Collaborative Knowledge Construction	Post: In the reading, the author explained that the main goal of sociology is to study the conditions that allow societies to continue existing. As I kept reading, I started to see how different kinds of societies are held together by different bonds and follow different sets of rules. What I'm still wondering, though, is <i>why</i> this focus on conservation is so important? • Strong (2): I really like the way you phrased your question about “why” societies need to conserve themselves. It actually reminded me of <i>Similar Post #1</i>, where the author also wondered about the purpose of preserving societies in the first place. They suggested it might be about maintaining social order, and I think your point about rules connects well with that. I'm wondering if Durkheim saw conservation not just as survival, but as creating the conditions for people to cooperate and avoid chaos. Do you think his focus on mechanical vs. organic solidarity helps explain <i>why</i> different societies rely on different rules to hold themselves together? • Weak (1): That's a good observation. I think the author was mostly interested in how societies stay stable over time, kind of like what some of the other posts mentioned. Maybe it's just about survival. What do you think? • Not Present (0): The author believed sociology's goal was to identify the conditions that allow societies to survive, which is why different rules apply in different contexts.

L Additional VTA (Kweon et al., 2025) Examples and Assessments

L.1 Example 1

9/12/25, 10:09 PM

main_custom

Virtual Teaching Assistant Chat

Go to Home

Refresh

I'm trying to get a clearer understanding of circular satellite orbits—specifically, how their motion is described and how to calculate the total mechanical energy involved. In addition, I'm curious about how the concept of gravitational waves connects to phenomena like black holes. Could someone help explain these ideas in more detail?

Absolutely! Let's break down your questions into two parts: circular satellite orbits and gravitational waves related to black holes.

Circular Satellite Orbits

- 1. Motion Description:
 - A satellite in a circular orbit moves around a planet (like Earth) at a constant speed. The gravitational force between the satellite and the planet provides the necessary centripetal force to keep the satellite in orbit.
 - The gravitational force (F) acting on the satellite can be described by Newton's law of gravitation: $[F = \frac{G \cdot m_1 \cdot m_2}{r^2}]$ where:
 - (G) is the gravitational constant,
 - (m_1) is the mass of the planet,
 - (m_2) is the mass of the satellite,
 - (r) is the distance from the center of the planet to the satellite.
- 2. Centripetal Force:
 - For circular motion, the centripetal force (F_c) required to keep the satellite in orbit is given by: $[F_c = \frac{m_2 \cdot v^2}{r}]$ where (v) is the orbital speed of the satellite.
- 3. Equating Forces:
 - For a satellite in a stable orbit, the gravitational force equals the centripetal force: $[\frac{G \cdot m_1 \cdot m_2}{r^2} = \frac{m_2 \cdot v^2}{r}]$
 - Simplifying this gives the orbital speed: $[v = \sqrt{\frac{G \cdot m_1}{r}}]$
- 4. Total Mechanical Energy:

localhost:8501

1/3

- The total mechanical energy (E) of the satellite in orbit is the sum of its kinetic energy (K) and potential energy (U):
 - Kinetic energy: $[K = \frac{1}{2} m_2 v^2]$
 - Potential energy: $[U = -\frac{G \cdot m_1 \cdot m_2}{r}]$
- Substituting (v) into the kinetic energy equation gives: $[K = \frac{1}{2} m_2 \left(\frac{G \cdot m_1}{r} \right) = \frac{G \cdot m_1 \cdot m_2}{2r}]$
- Therefore, the total mechanical energy is: $[E = K + U = \frac{G \cdot m_1 \cdot m_2}{2r} - \frac{G \cdot m_1 \cdot m_2}{r} = -\frac{G \cdot m_1 \cdot m_2}{2r}]$
- This shows that the total mechanical energy of a satellite in a circular orbit is negative, indicating that the satellite is in a bound state.

Gravitational Waves and Black Holes

1. Gravitational Waves:

- Gravitational waves are ripples in spacetime caused by the acceleration of massive objects, predicted by Einstein's General Theory of Relativity. They propagate at the speed of light and carry energy away from their source.

2. Connection to Black Holes:

- Black holes can produce gravitational waves when they merge or collide. For example, when two black holes spiral towards each other and eventually merge, they create strong gravitational waves that can be detected by observatories like LIGO (Laser Interferometer Gravitational-Wave Observatory).
- The detection of these waves provides insights into the properties of black holes, such as their masses and spins, and helps confirm aspects of general relativity.

3. Phenomena:

- The merging of black holes is one of the most significant sources of gravitational waves. The waves carry information about the dynamics of the black holes and the extreme conditions of spacetime around them.

Summary

- Circular satellite orbits are governed by gravitational and centripetal forces, with total mechanical energy being negative, indicating a bound system.
- Gravitational waves are produced by accelerating masses, such as merging black holes, and provide valuable information about these extreme cosmic events.

9/12/25, 10:09 PM

main_custom

If you have any more questions or need further clarification on any of these topics, feel free to ask!

Type your message here...



localhost:8501

3/3

1. **Clarifying Misunderstandings and Confusions:** Weak (1)
2. **Deepen Disciplinary Understanding:** Strong (2)
3. **Foster Higher-Order Thinking:** Not Present (0)
4. **Enhance Metacognitive Awareness:** Weak (1)

L.2 Example 2

9/12/25, 10:10 PM

main_custom

Virtual Teaching Assistant Chat

[Go to Home](#)[Refresh](#)

The textbook noted that Social Security benefits were once fully excluded from income, but now up to 85% may be taxed. Just wonder what led to this change?



The change in the taxation of Social Security benefits can be traced back to the 1980s when the U.S. government faced significant budget deficits and rising costs associated with Social Security and Medicare. Here are the key factors that led to this change:

1. **Budget Deficits:** In the early 1980s, the federal government was dealing with substantial budget deficits. To address this issue, policymakers sought ways to increase revenue without raising taxes across the board.
2. **Social Security Amendments of 1983:** In response to the financial challenges facing the Social Security program, Congress passed the Social Security Amendments of 1983. This legislation included provisions to tax a portion of Social Security benefits for higher-income beneficiaries. The rationale was that those with higher incomes could afford to contribute more to the federal budget.
3. **Income Thresholds:** The amendments established income thresholds, meaning that if a beneficiary's income exceeded a certain level, a portion of their Social Security benefits would be subject to federal income tax. Initially, up to 50% of benefits could be taxed, and this was later increased to 85% for individuals with higher incomes.
4. **Changing Demographics:** The aging population and increasing life expectancy also played a role. As more people began to rely on Social Security for a longer period, the financial sustainability of the program became a concern, prompting the need for additional funding sources.
5. **Political Compromise:** The decision to tax Social Security benefits was part of a broader compromise that included measures to strengthen the Social Security system, such as increasing the retirement age and adjusting the benefit formula.

Overall, the taxation of Social Security benefits reflects a shift in policy aimed at ensuring the program's long-term viability while addressing the fiscal challenges faced by the government.

localhost:8501

1/2

1. **Clarifying Misunderstandings and Confusions:** 2
2. **Deepen Disciplinary Understanding:** 2
3. **Foster Higher-Order Thinking:** 0
4. **Enhance Metacognitive Awareness:** 0

L.3 Example 3

9/12/25, 10:11 PM

main_custom

Virtual Teaching Assistant Chat

[Go to Home](#)

[Refresh](#)



I'm confused about how bunching itemized deductions works. It seems to involve shifting deductions into one year to exceed the standard deduction, then using the standard deduction the next year. How exactly can a taxpayer shift deductions—what does it mean to “accelerate” payments into the current year?



Bunching itemized deductions is a tax strategy that allows taxpayers to maximize their deductions in a given tax year, particularly when their total itemized deductions are close to the standard deduction threshold. Here's how it works and what it means to "accelerate" payments:

Understanding Bunching Itemized Deductions

1. Standard Deduction vs. Itemized Deductions:

- The standard deduction is a fixed dollar amount that reduces the income you are taxed on. For example, in 2023, the standard deduction for single filers is *13,850* and for married couples filing jointly, it is *27,700*.
- Itemized deductions include expenses such as mortgage interest, property taxes, medical expenses, and charitable contributions. Taxpayers can choose to itemize their deductions if they exceed the standard deduction.

2. Bunching Strategy:

- The idea behind bunching is to time your deductions so that in one year, you have enough itemized deductions to exceed the standard deduction, while in the following year, you take the standard deduction.
- This can be particularly useful for taxpayers whose itemized deductions fluctuate from year to year.

How to Shift Deductions

1. Accelerating Payments:

- **Accelerating payments** means making certain deductible payments earlier than you normally would. For example:
 - **Charitable Contributions:** If you typically donate to charity in January, consider making that donation in December instead. This allows you to claim the deduction in the current tax year.
 - **Medical Expenses:** If you have medical expenses that you can control (like elective procedures), you might schedule them for one year to maximize your deductions.
 - **Property Taxes:** If you pay property taxes annually, you could pay the next year's property tax bill early, before the end of the current tax year.

2. Deferring Payments:

localhost:8501

1/2

- Conversely, you can also defer some payments to the next year if you have already exceeded the standard deduction in the current year. For example, if you usually pay your state income taxes in April, you might choose to delay that payment until the following January.

Example of Bunching

- **Year 1:** You have 10,000 itemized deductions (e.g., mortgage interest, property taxes, etc.). You also plan to donate 5,000 to charity in January of Year 2. If you accelerate that donation to December of Year 1, your total itemized deductions for Year 1 would be \$15,000, which exceeds the standard deduction. You would itemize in Year 1.
- **Year 2:** Without any additional deductions, you would only have the standard deduction available, as you have shifted your charitable contribution to Year 1.

Conclusion

By strategically timing your deductible expenses, you can maximize your tax benefits over a two-year period. This approach requires careful planning and consideration of your financial situation, so it may be beneficial to consult with a tax professional to tailor the strategy to your specific circumstances.

Type your message here...



1. **Clarifying Misunderstandings and Confusions:** 2
2. **Deepen Disciplinary Understanding:** 1
3. **Foster Higher-Order Thinking:** 0
4. **Enhance Metacognitive Awareness:** 0