Cognitive-Uncertainty Guided Knowledge Distillation for Accurate Classification of Student Misconceptions

Qirui Liu[1,2,*] **Hao Chen**[2,*] **Weijie Shi**[3] **Jiajie Xu**[4] **Jia Zhu**[5,†]

[1]South China University of Technology, [2]Tencent Financial Technology
[3]The Hong Kong University of Science and Technology, [4]Soochow University
[5]Zhejiang Key Laboratory of Intelligent Education Technology and Application, Zhejiang Normal University

Abstract

Accurately identifying student misconceptions is crucial for personalized education but faces three challenges: (1) data scarcity with long-tail distribution, where authentic student reasoning is difficult to synthesize; (2) fuzzy boundaries between error categories with high annotation noise; (3) deployment paradoxlarge models overlook unconventional approaches due to pretraining bias and cannot be deployed on edge, while small models overfit to noise. Unlike traditional methods that increase diversity through large-scale data synthesis, we propose a two-stage knowledge distillation framework that mines high-value samples from existing data. The first stage performs standard distillation to transfer task capabilities. The second stage introduces a dual-layer marginal selection mechanism based on cognitive uncertainty, identifying four types of critical samples based on teacher model uncertainty and confidence differences. For different data subsets, we design difficulty-adaptive mechanism to balance hard/soft label contributions, enabling student models to inherit inter-class relationships from teacher soft labels while distinguishing ambiguous error types. Experiments show that with augmented training on only 10.30% of filtered samples, we achieve MAP@3 of 0.9585 (+17.8%) on the MAP-Charting dataset, and using only a 4B parameter model, we attain 84.38% accuracy on cross-topic tests of middle school algebra misconception benchmarks, significantly outperforming sota LLM (67.73%) and standard fine-tuned 72B models (81.25%). Our code is available at https://github.com/RoschildRui/acl2026_map.

1 Introduction

Understanding how students think and solve problems remains a core challenge in educational research (Carly D. Robinson, 2021; Dyer and Sherin, 2016; Parwati and Suharta, 2020). Traditional assessment methods focusing solely on answer correctness overlook students' reasoning processes, failing to reveal cognitive obstacles or recognize partially valid thinking within incorrect answers. This limits teachers' ability to provide personalized feedback and prevents platforms from making adaptive adjustments based on actual thinking trajectories (Graesser et al., 2004; Wang et al., 2020).

This limitation is particularly acute in mathematics education, where identical incorrect answers may stem from completely different reasoning paths revealing distinct conceptual misunderstandings (Ansari et al., 2025; Sadler et al., 2013). For example, solving $2x + 3 = 11$, one student might correctly subtract 3 then divide by 2 but make a calculation error, while another might directly divide 11 by 2 due to conceptual confusion. Both produce wrong answers requiring fundamentally different interventions (Otero et al., 2025). This necessitates shifting from evaluating "answer correctness" to understanding "problem-solving approaches," which recent NLP advances now make feasible through automatic classification of student reasoning processes (Hsu et al., 2025).

However, accurate classification faces three core challenges:

Challenge 1: Data scarcity and distribution gap. Real student process data is severely limited with long-tail distribution (Hsu et al., 2025). While existing work uses LLMs to generate synthetic data (Tan et al., 2024), real student reasoning features colloquial language, reasoning jumps, and logical errors that models cannot accurately replicate. LLM-generated text's superior logic and fluency creates significant distribution gaps, limiting generalization to real scenarios.

Challenge 2: Label complexity and annotation noise. Student misconception labels are numerous with fuzzy boundaries between error

[*] Equal contribution.
[†] Corresponding Author.

Findings of the Association for Computational Linguistics: ACL 2026, pages 29964–29980

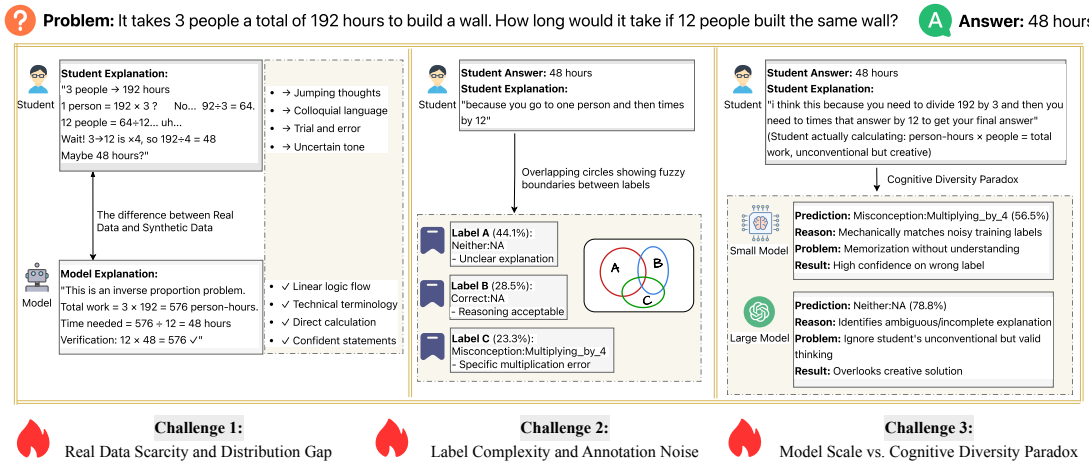


Figure 1: The overview of the key points and corresponding challenges.

types and even correct/incorrect categories (Otero et al., 2025), causing substantial annotation noise. Traditional hard-label models cannot learn subtle inter-category differences or handle inherent uncertainty, performing poorly on complex ambiguous errors.

Challenge 3: Model scale and cognitive diversity paradox. Students' diverse thinking often produces unconventional but reasonable solutions within their cognitive framework (Ansari et al., 2025). Small models meet deployment needs but overfit to noisy labels; large models possess rich knowledge but systematically overlook non-standard approaches due to pretraining bias (Shi et al., 2023), force-fitting innovative solutions into existing frameworks and misjudging error types. Educational privacy requirements and edge device limitations further prevent direct large model deployment.

Addressing these challenges, unlike traditional methods that rely on large-scale data synthesis (Tan et al., 2024), we propose a two-stage knowledge distillation framework that strategically mines high-value samples from existing data. The first stage performs standard distillation for basic task capability transfer; the second stage introduces a dual-layer marginal selection mechanism based on cognitive uncertainty, identifying Near-miss (correct but uncertain) and Hard-hard (severely incorrect) critical samples through teacher uncertainty and confidence differences, with difficulty-adaptive loss functions dynamically balancing hard/soft labels to help students distinguish complex error types while inheriting inter-class relationships. This approach produces lightweight models that accurately identify diverse student approaches while meeting practical requirements for privacy protection and edge deployment.

Our main contributions include:

- A dual-layer marginal selection mechanism based on cognitive uncertainty that precisely filters high-value real samples for incremental training, improving limited-data performance without synthetic data dependency.
- Difficulty-adaptive loss functions with dynamic soft/hard label weighting based on sample difficulty, addressing label noise and boundary ambiguity while improving discrimination of confusable categories.
- Addressing large models' diversity oversight from pretraining bias through two-stage distillation, balancing knowledge transfer with cognitive openness to maintain inclusiveness toward non-standard solutions.

2 Related Work

In this section, we first introduce the evolution in the education domain from answer scoring to process understanding, and then discuss related techniques for handling data scarcity and label noise.

2.1 Student Reasoning Assessment and Misconception Diagnosis

Educational assessment research has undergone a paradigm shift from result-oriented evaluation to process understanding. Early studies focused primarily on the automatic determination of answer

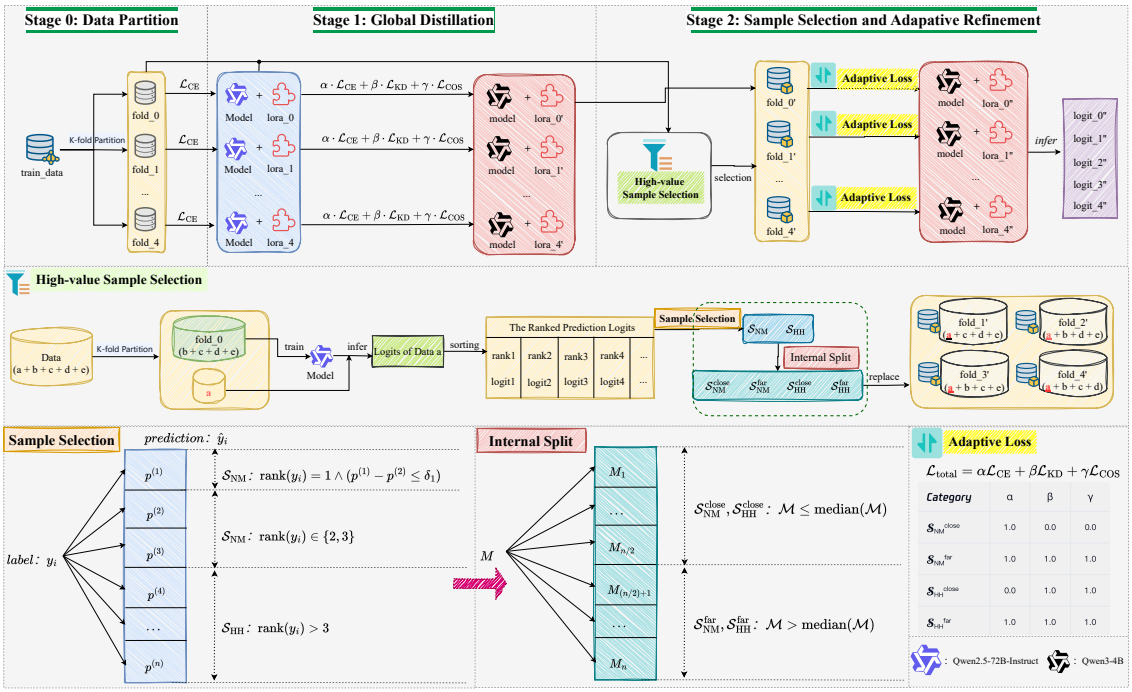


Figure 2: Overview of our two-stage distillation framework, consisting of data partitioning, two-stage knowledge distillation, and sample selection.

correctness (Dyer and Sherin, 2016; Parwati and Suharta, 2020), achieving scoring accuracy close to human level, but lacking the ability to diagnose the underlying causes of errors. Recent work has begun to explore misconception identification. For example, Ansari et al. (2025); Hsu et al. (2025) proposed fine-grained error classification frameworks in the algebra domain, and Otero et al. (2025) summarized 55 types of algebraic misconceptions and built diagnostic benchmarks. However, these approaches still rely mainly on the final answer rather than the complete reasoning process.

Directly analyzing students reasoning processes is key to understanding cognitive barriers. Shi et al. (2023) found that even advanced language models can fail in reasoning when confronted with explanations containing irrelevant information or logical leaps, a phenomenon that mirrors characteristics commonly observed in real student responses. However, obtaining high-quality student process data is extremely challenging. Although large language models (LLMs) can generate synthetic data (Tan et al., 2024), such data tends to exhibit overly standardized expressions that differ markedly from the colloquial, non-standard reasoning of real students, resulting in poor generalization for models trained solely on synthetic data.

To address the distribution bias in synthetic data, we propose a sample selection strategy based on cognitive uncertainty. This approach accurately identifies the most critical *near-miss* and *hard-hard* samples from limited real data, which are most influential to the decision boundary, thereby avoiding reliance on synthetic data.

2.2 Sample-Efficient Learning under Label Noise

In the presence of data scarcity and label noise, the machine learning community has developed various strategies. Curriculum learning (Bengio et al., 2009; Chen et al., 2025) improves learning efficiency by arranging training in an easy-to-hard sequence; subsequent work further introduced uncertainty-based active sample selection (Houlsby et al., 2011; Gal et al., 2017; Kirsch et al., 2019), prioritizing the learning of samples with the highest information gain. The core idea behind these methods is to identify and focus on key samples that can significantly improve the decision boundary (Yuan et al., 2020; Fang et al., 2021; Yin et al., 2020).

Knowledge distillation (Hinton et al., 2015; Mansourian et al., 2025) offers another path for handling label noise. The soft labels from teacher models embed rich inter-class relational informa-

tion, which can mitigate the impact of noisy hard labels. Recent adaptive distillation methods (Chennupati et al., 2021; Song et al., 2022; Yu et al., 2025) further propose dynamically adjusting the weights of soft and hard labels based on sample difficulty, enabling more fine-grained knowledge transfer (Hao et al., 2023; Cazenavette et al., 2022; Liu et al., 2022). However, existing approaches face unique challenges in educational scenarios: biases from large model pretraining make it difficult to accept non-standard student reasoning, and direct distillation can cause small models to excessively inherit these biases (Luan et al., 2019; Hossain et al., 2025; Chen et al., 2022; Guo et al., 2020).

To address the cognitive bias issue in large models, we propose a difficulty-adaptive loss function that dynamically adjusts the weighting of soft and hard labels, allowing small models to inherit knowledge from large models while maintaining tolerance for non-standard student expressions.

3 Methodology

To address the dual challenges of data and labels in identifying students' problem-solving approaches, we propose an Adaptive Knowledge Distillation Framework for High-Value Samples. Instead of relying on synthetic data with distribution gaps, our framework accurately identifies and exploits the most valuable samples from limited authentic data through a two-stage training process of sample selection and adaptive learning, guiding models to tackle real-world ambiguity and complexity.

3.1 Problem Formulation

Our goal is to train a classification model f_s that accurately assigns student math problem-solving processes x_i (including problem description, answer, correctness, and metadata) to predefined reasoning labels y_i. The label set $\mathcal{Y}$ contains "Correct", multiple misconception categories ($\text{Misconception}_1, ..., \text{Misconception}_K$), and "Neither" for unclassifiable cases.

Given training set $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ with N real student samples, we optimize student model f_s parameters using pretrained teacher models f_t to address label noise and data sparsity. The model outputs probability distribution $p(y|x) = \text{softmax}(z/\tau)$, where $z \in R^{|\mathcal{Y}|}$ represents logits and τ is the temperature coefficient for distribution smoothing in knowledge distillation.

3.2 Stage One: Global Knowledge Distillation and Preliminary Learning

In student thinking discrimination, we face a fundamental contradiction: large models possess abundant knowledge but ignore thinking diversity due to overconfidence in prior distributions; small models are flexible but lack necessary knowledge foundations. Stage one thus uses knowledge distillation to equip small models with basic knowledge structures while maintaining openness to non-standard ideas.

We employ n-fold cross-validation to generate soft labels and prevent overfitting. Using `StratifiedKFold` (Raschka, 2020) based on label distribution, we split data into n subsets. For each fold $k \in \{1, ..., n\}$, train teacher model $f_t^{(k)}$ on $\mathcal{D} \setminus \mathcal{D}_k$ and generate soft labels $\mathbf{y}_{\text{soft}}^{(i)}$ for samples in $\mathcal{D}_k$.

We then train n student models with a loss function combining three objectives:

$$\mathcal{L} = \alpha \cdot \mathcal{L}_{\text{CE}}(f_s(x_i), y_i) + \beta \cdot \mathcal{L}_{\text{KD}}(f_s(x_i), \mathbf{y}_{\text{soft}}) + \gamma \cdot \mathcal{L}_{\text{COS}}(f_s(x_i), \mathbf{y}_{\text{soft}}) \quad (1)$$

where $\mathcal{L}_{\text{CE}}$ ensures basic classification accuracy, $\mathcal{L}_{\text{KD}}$ transfers inter-class relationships, and $\mathcal{L}_{\text{COS}}$ constrains student-teacher consistency in representation space.

To ensure the stability and cross-model generalization of hyperparameters, we conducted comprehensive grid search and ablation experiments on the loss weights (α, β, γ) for Stage 1 (see Appendix D).

3.3 Stage Two: High-Value Sample Selection and Adaptive Refinement

After initial training, we employ a two-tier margin selection mechanism to identify the most valuable samples for model enhancement, then perform targeted training on this subset using difficulty-adaptive loss functions.

3.3.1 Two-Tier Margin-Based Sample Selection

In our framework, the teacher's cognitive uncertainty serves only as a guiding signal for high-value sample selection rather than a hard constraint (see Appendix E).

This mechanism extracts samples that most effectively reveal model weaknesses and provide maximum informational value, directly addressing data scarcity and distribution discrepancies

through deep exploration of existing data rather than blind expansion.

Uncertainty-Based Difficulty Partition The most beneficial samples are not ones already mastered, but those near decision boundaries or beyond current understanding. Based on teacher model f_t predictions from stage one, we identify two high-value types:

- **Near-miss samples ($\mathcal{S}_{\text{NM}}$):** Samples with correct but low-confidence predictions, or incorrect predictions where the correct answer is almost reached. These boundary-adjacent samples are crucial for fine-grained discrimination. Formally:

$$\mathcal{S}_{\text{NM}} = \{(x_i, y_i) : [(\hat{y}_i = y_i) \wedge (p^{(1)} - p^{(2)}) \leq \delta)] \vee \text{rank}(y_i) \in \{2, 3\}\} \tag{2}$$

  Here, $p^{(k)}$ represents the k-th highest predicted probability from the model, $\hat{y}_i$ denotes the predicted class, $\text{rank}(y_i)$ is the position of the true label y_i in the sorted prediction probabilities, and δ is a small confidence margin threshold (e.g., 0.05).

- **Hard-hard samples ($\mathcal{S}_{\text{HH}}$):** These are samples for which the models predictions are grossly incorrect, meaning the predicted probability ranking of the true label is very low. Such samples expose fundamental knowledge gaps or severe misunderstandings of certain complex concepts.

$$\mathcal{S}_{\text{HH}} = \{(x_i, y_i) : \text{rank}(y_i) > 3\} \tag{3}$$

Through this tiered partitioning, we narrow the training focus from the entire dataset to $\mathcal{D}_{\text{selected}} = \mathcal{S}_{\text{NM}} \cup \mathcal{S}_{\text{HH}}$, achieving the first stage of concentration on high-value samples.

Fine-Grained Differentiation Based on Probability Margin Even within $\mathcal{S}_{\text{NM}}$ and $\mathcal{S}_{\text{HH}}$, there exists significant variation in sample difficulty. To enable more fine-grained adaptive learning, we introduce a composite difficulty measure that combines the prediction probability margin with distributional uncertainty. The composite difficulty metric $M(x_i, y_i)$ is defined as follows:

1. **Probability Margin:**

$$d(x_i, y_i) = |p_s(y_i|x_i) - \max_{j \in \mathcal{Y}} p_s(j|x_i)| \tag{4}$$

This reflects the models direct fitting degree to the ground truth label. A larger d indicates that the model's understanding deviates more from the true label; a smaller d suggests that the model is closer to correctly understanding it.

2. **Prediction Entropy:**

$$H(x_i) = -\sum_{j \in \mathcal{Y}} p_s(j|x_i) \log p_s(j|x_i) \tag{5}$$

This reflects the overall dispersion of the models predictions.

3. **Composite Difficulty Metric:**

$$\mathcal{M}(x_i, y_i) = d(x_i, y_i) \cdot e^{-H(x_i)} \tag{6}$$

Here, the margin $d(x_i, y_i)$ represents the prediction bias with respect to the ground truth; the entropy $H(x_i)$ adjusts the weight of the bias, with $e^{-H(x_i)}$ amplifying difficulty when entropy is low, and attenuating it when entropy is high. This metric is designed to capture two complementary dimensions of difficulty.

Based on M, we further divide $\mathcal{S}_{\text{NM}}$ and $\mathcal{S}_{\text{HH}}$ into close and far subsets according to the median:

$$\mathcal{S}_t^{\text{close}} = \{(x_i, y_i) \in \mathcal{S}_t : \mathcal{M}(x_i, y_i) \leq \text{median}(\mathcal{M}_t)\} \tag{7}$$

$$\mathcal{S}_t^{\text{far}} = \{(x_i, y_i) \in \mathcal{S}_t : \mathcal{M}(x_i, y_i) > \text{median}(\mathcal{M}_t)\} \tag{8}$$

where $t \in \{\text{NM}, \text{HH}\}$.

This dual-dimensional characterization constructed from the probability margin and prediction entropy precisely depicts different levels of sample difficulty, providing a reliable basis for subsequent adaptive loss design. Hyperparameter search confirms $\delta = 0.05$ and $K = 5$ as optimal settings (see Appendix D).

3.3.2 Difficulty-Adaptive Loss Function

To address the issues of label complexity and annotation noise, we design an adaptive loss function that dynamically fuses information from hard labels and soft labels, with its weights adjusted according to the sample categories identified in the previous section. The complete definition of the total loss is:

$$\mathcal{L}_{\text{total}} = \alpha \mathcal{L}_{\text{CE}} + \beta \mathcal{L}_{\text{KD}} + \gamma \mathcal{L}_{\text{COS}} \tag{9}$$

where $\mathcal{L}_{\text{CE}}$ is the standard cross-entropy loss based on the ground-truth (hard) label y_i; $\mathcal{L}_{\text{KD}}$ is the knowledge distillation loss that guides the student

model f_s to mimic the soft probability distribution output by the teacher model f_t; and $\mathcal{L}_{\text{COS}}$ is the cosine embedding loss, which constrains the directional consistency between the student and teacher models in the representation space. The specific formulations are defined as:

$$\mathcal{L}_{\text{CE}} = -\log p_s(y_i|x_i) \tag{10}$$

$$\mathcal{L}_{\text{KD}} = \tau^2 \cdot \text{KL}(p_t(\cdot|x_i)\|p_s(\cdot|x_i)) \tag{11}$$

$$\mathcal{L}_{\text{COS}} = 1 - \cos\left(p_s(\cdot \mid x_i),\, p_t(\cdot \mid x_i)\right) \tag{12}$$

The key lies in the adaptive allocation strategy of the coefficients (α, β, γ) based on different sample categories. For samples with high uncertainty near the decision boundary, strong constraints from hard labels are necessary to avoid the soft-label smoothing effect. Samples near but slightly away from the boundary benefit from both the precision of hard labels and the inter-class relationships of soft labels. When predictions are close to ground-truth but exhibit significant deviations, soft labels are favored to mitigate noise impact. For extremely difficult samples, both hard- and soft-label guidance need to be strengthened. The specific weight allocation for each sample category is detailed in Appendix A. We also visualize sample characteristics and prediction differences to verify selection effectiveness (see Appendix F).

4 Experiments and Analysis

4.1 Experimental Setup

4.1.1 Datasets

We evaluate the proposed method on two complementary benchmarks, which represent different levels of granularity in student misconception detection:

MAP-Charting dataset (King et al., 2025): This dataset contains real reasoning traces of students in multiple-choice mathematics questions. Each sample includes the problem statement, the student's answer, the correctness label, andcriticallythe student's written explanation of their reasoning process. Labels are divided into three categories: Correct, Misconception (with specific misconception types), or Neither (indicating vague or irrelevant thinking). This fine-grained dataset consists of 36,695 samples in total.

Algebra Misconceptions Benchmark (Nancy et al., 2024): This benchmark covers 55 types of algebraic misconceptions validated by 145 peer-reviewed studies. Unlike MAP-Charting, this dataset requires only the students final answer (without reasoning traces) to identify the misconception type, making it coarser-grained but easier to collect. We randomly select questions from all available topics for evaluation. This dataset contains a total of 220 samples.

4.1.2 Implementation Details

For student models, we use three lightweight architectures: Qwen-3-4B (Team, 2025c), Gemma-2-9B (Team, 2024a), and Llama-3.1-8B (Grattafiori et al., 2024), with Qwen-2.5-72B (Team, 2024b) as the teacher. All models are fine-tuned via AdamW (batch size=16, 4 gradient accumulation steps): student learning rate 2×10^{-4}, teacher learning rate 1×10^{-4}, distillation temperature $\tau = 1.0$, and confidence threshold $\delta = 0.05$ (from preliminary experiments). For second-stage incremental training, student models use learning rate 1×10^{-6} and `max_grad_norm`=4, with other configurations unchanged.

4.1.3 Baseline Methods

Prompting-based models: We evaluate various prompting strategies for advanced large language models, including in-context learning and chain-of-thought reasoning. Tested models include GPT-5 (Team, 2025b), Claude-4-Sonnet (Team, 2025a), DeepSeek-V3 (DeepSeek-AI et al., 2025), and Qwen-2.5-72B.

Classification-based models: We attach a classification head and fine-tune different sizes of language models, including student models (Qwen-3-4B, Gemma-2-9B, Llama-3.1-8B) and the teacher model (Qwen-2.5-72B) directly fine-tuned without distillation.

4.1.4 Evaluation Metrics

We adopt four evaluation metrics to comprehensively assess ranking quality from multiple perspectives: MAP@3, MAP@10, Accuracy, and F1@3. MAP@3 mainly measures the model's ability to identify multiple possible correct answers within the top 3 predictions; MAP@10 extends the scope to the top 10 predictions, evaluating ranking performance on a broader scale; Accuracy measures the overall correctness of predictions; F1@3 considers both precision and recall within the top 3 predictions, reflecting the models balanced performance in multi-candidate scenarios. These four metrics allow us to compare model performance in terms of local precision, overall

Table 1: Performance comparison on two benchmark datasets. The best results in each category are shown in **bold**.

Method	MAP-Charting				Algebra Misconception Benchmark			
	MAP@3	MAP@10	Accuracy	F1@3	MAP@3	MAP@10	Accuracy	F1@3
Prompting-based Methods								
GPT-5	**0.8137**	**0.8145**	**0.7225**	**0.4626**	**0.7409**	**0.7418**	**0.6773**	**0.4091**
Claude-4-Sonnet	0.7833	0.7841	0.6914	0.4579	0.6636	0.6645	0.5636	0.3932
DeepSeek-V3	0.7665	0.7673	0.6601	0.4505	0.6485	0.6494	0.5545	0.3815
GPT-OSS-120B	0.7661	0.7669	0.6794	0.4375	0.6550	0.6559	0.5680	0.3725
Qwen-2.5-72B (prompting)	0.7285	0.7293	0.6222	0.4328	0.6280	0.6289	0.5320	0.3670
Fine-tuned-based Methods								
Qwen-2.5-72B (fine-tuned)	**0.9497**	**0.9501**	**0.9014**	**0.4993**	**0.8438**	**0.8612**	**0.8125**	**0.4375**
Qwen-3-4B (fine-tuned)	0.9472	0.9475	0.8987	0.4992	0.7552	0.7669	0.7188	0.4062
Gemma-2-9B (fine-tuned)	0.9439	0.9442	0.8919	0.4992	0.7708	0.7862	0.7188	0.4219
Llama-3.1-8B (fine-tuned)	0.9453	0.9456	0.8954	0.4990	0.7760	0.7917	0.7500	0.4062
Our Method (Two-Stage Distillation)								
Qwen-3-4B + Ours	**0.9585**	**0.9587**	**0.9198**	**0.4996**	**0.8750**	**0.8915**	**0.8438**	**0.4531**
Gemma-2-9B + Ours	0.9560	0.9562	0.9148	0.4995	0.8015	0.8155	0.7656	0.4375
Llama-3.1-8B + Ours	0.9553	0.9555	0.9134	0.4995	0.7865	0.7995	0.7564	0.4281

accuracy, and ranking coverage from multiple angles.

4.2 Main Results

Table 1 shows that two-stage distillation markedly boosts MAP@3 and MAP@10 on both MAP-Charting and Algebra Misconception benchmarks, outperforming prompt-based reasoning and direct fine-tuning. In MAP-Charting, the best prompt-based GPT5 score (0.8137/0.8145) rises to 0.9497/0.9501 via direct fine-tuning of Qwen-2.5-72B, while the distilled Qwen-3-4B attains 0.9585/0.9587, about 0.9% above the teacher model and surpassing the 72B model; Gemma-2-9B and Llama-3.1-8B show similar gains. On Algebra Misconception, Qwen-3-4B improves from 0.7552/0.7669 to 0.8750/0.8915 (15.8%/16.2%), exceeding GPT5 by 18.1%/20.2%. Consistent gains across models and metrics indicate stable advantages. Overall, two-stage distillation enables lightweight students to close and sometimes surpass the gap with large teacher models, combining stability, generality, and efficiency. The 4B-parameter student model outperforms the 72B-parameter teacher model, with key reasons detailed in Appendix G.

4.3 Ablation Study

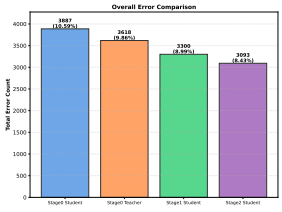
Our ablation study consists of two parts. The first part removes (ablates) the main components of the method one by one to verify the impact of different modules on model performance. The second part takes into account the multi-stage distillation training nature of our method and designs a cross-stage performance comparison experiment.

4.3.1 Ablation of Main Components

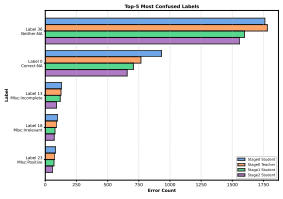
Table 2 reports ablation results for MAP@10, MAP@3, and Accuracy using the two-stage distillation scheme based on Qwen-3-4B. Removing any core component consistently lowers both MAP@10 and MAP@3, indicating stable gains across different evaluation dimensions. On the MAP-Charting dataset, the full method achieves 0.9587/0.9585 (MAP@10/MAP@3) and 0.9198 (Accuracy); removing adaptive loss or high-value sample selection reduces performance by 0.4%–0.7%, while omitting second-stage distillation causes the largest drop, with Accuracy falling by over 1.7%. On the Algebra Misconception benchmark, the full method obtains 0.8915/0.8750 and 0.8438; removing second-stage distillation reduces MAP by more than 10% and Accuracy to 0.7577. Overall, each stage of the distillation process provides significant, consistent improvements across different evaluation dimensions.

4.3.2 Performance Across Different Stages

Figures 3a and 3b show error count changes across key stages and top-5 confused categories, with later-stage models consistently reducing total and per-category errors. This confirms multi-stage distillation lowers error rates and enhances fine-grained classification, especially for difficult categories. Detailed top-10 and 37-category analyses are in Appendix B.



(a) Error count changes across different stage models



(b) Error count changes in top-5 confused categories

Figure 3: Visualization of error counts in multi-stage distillation training.

Table 2: Ablation study results. The best result for each benchmark dataset is in **bold**.

Method Variant	MAP-Charting		
	MAP@10	MAP@3	Accuracy
Full Method	**0.9587**	**0.9585**	**0.9198**
w/o Adaptive Loss	0.9542	0.9540	0.9123
w/o Sample Selection	0.9521	0.9519	0.9085
w/o Stage-1 Distillation	0.9548	0.9546	0.9132
w/o Stage-2 Distillation	0.9495	0.9493	0.9024
Method Variant	**Algebra Misconception**		
	MAP@10	MAP@3	Accuracy
Full Method	**0.8915**	**0.8750**	**0.8438**
w/o Adaptive Loss	0.8802	0.8657	0.8321
w/o Sample Selection	0.8741	0.8603	0.8269
w/o Stage-1 Distillation	0.8823	0.8679	0.8342
w/o Stage-2 Distillation	0.8001	0.7893	0.7577

4.4 Efficiency Analysis

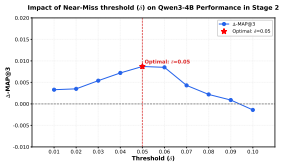
To evaluate practical value, we tested inference efficiency on 7,339 samples (Table 3). Qwen-3-4B outperforms GPT-5 by 18.0 pp in MAP@3 (0.9599 vs 0.8137) with 187.5× speedup (0.008 h vs 1.50 h), slightly outperforms teacher Qwen-2.5-72B (0.9599 vs 0.9497) with 23.25× speedup, and surpasses GPT-OSS-120B by 25.3 pp with 137.5× speedup (runnable on PC). These results confirm our methods gains in producing accurate, lightweight models for misconception classification.

4.5 Parameter Analysis

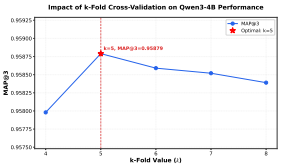
In the second training stage, the class probability threshold δ is used to filter high-uncertainty samples. Based on Qwen-3-4B (Team, 2025c), experiments in the range $[0.01, 0.10]$ evaluated changes in validation MAP@3, with gain defined as $\text{MAP@3}(\delta) - \text{MAP@3}(\delta_{\text{baseline}})$. Results (Figure 4a) show $\delta = 0.05$ yields the largest improvement ($+0.012$), while $\delta = 0.10$ causes a drop (-0.004), indicating that moderate thresholds effectively select useful samples, whereas overly high thresholds risk overfitting. K-fold validation further shows performance peaks at $K = 5$, achieving **0.95879** MAP@3, suggesting this split optimally balances training coverage and validation stability.

Table 3: Inference efficiency comparison over 7,339 samples

Model	MAP@3	Time (h)	Hardware
API Models			
GPT-5	0.8137	1.50	Cloud API
Claude-4-Sonnet	0.7665	1.80	Cloud API
Self-deployment Models			
GPT-OSS-120B	0.7661	1.10	32 × H20
Qwen-2.5-72B	0.7285	1.30	32 × H20
Our Models			
Qwen-2.5-72B (teacher)†	0.9497	0.186	8 × H20
Qwen-3-4B (student)*	**0.9599**	**0.008**	**8 × H20**



(a) Impact of δ threshold



(b) Impact of different K-fold

Figure 4: Results of two hyperparameter search experiments in the second stage.

5 Conclusion

This study proposes a hierarchical-difficulty-based sample selection and probability-constrained multi-scale knowledge distillation framework, achieving the high accuracy required for real-world deployment of educational AI under extremely data-scarce conditions. Leveraging a dual-level margin mechanism to precisely identify *Near-miss* (correct predictions with uncertainty) and *Hard-hard* (severely incorrect predictions) samples, and designing adaptive loss functions for different types of samples, we achieve a MAP@3 of 0.9585 (relative improvement of 17.8%) on real student data, significantly enhancing the reliability of erroneous concept diagnosis. This framework can be extended in the future to progressive difficulty scheduling and multi-task learning, improving the personalization and scalability of automated tutoring systems.

Limitation

Although this study demonstrates promising results in multi-stage distillation and high-value sample selection, there are still several limitations:

First, the overhead associated with the K-fold partition. Although stratified 5-fold cross-validation helps mitigate overfitting (see Appendix H), we adopt a K-fold cross-partition approach to ensure reliability and prevent data leakage. However, searching for the optimal K is complex, as each evaluation requires a complete global training cycle. This high overhead may limit the exploration of optimal configurations in resource-constrained scenarios, thereby constraining the method's deployment potential.

Second, the limited improvement for low-quality data. While our multi-stage distillation framework enhances performance by selecting valuable samples, its efficacy is limited when incoming data is inherently of poor quality. In the presence of large quality gaps, the model cannot achieve significant gains solely through sample selection. Therefore, it is crucial to design effective high-quality data synthesis strategies to actively generate and repair data, rather than relying only on filtering existing samples.

Acknowledgments

This work was supported by the "Pioneer" and "Leading Goose" R&D Program of Zhejiang (No. 2026C02A1236), the National Natural Science Foundation of China (No. 62577050), and the Jinhua Major Science and Technology Project (No. 2024-1-005).

References

Shahina Mohd Azam Ansari, James Bywater, Sarah Lilly, Donald Brown, and Jennifer Chiu. 2025. Misstepmath: A diverse student mistake dataset forăai inămathematics teacher training. In *Artificial Intelligence in Education*, pages 381–394, Cham. Springer Nature Switzerland.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 4148, New York, NY, USA. Association for Computing Machinery.

Susanna Loeb Carly D. Robinson. 2021. High-impact tutoring: State of the research and priorities for future learning.

George Cazenavette, Tongzhou Wang, Antonio Torralba, Alexei A. Efros, and Jun-Yan Zhu. 2022. Dataset distillation by matching training trajectories. *Preprint*, arXiv:2203.11932.

Defang Chen, Jian-Ping Mei, Hailin Zhang, Can Wang, Yan Feng, and Chun Chen. 2022. Knowledge distillation with the reused teacher classifier. *Preprint*, arXiv:2203.14001.

Xiaoyin Chen, Jiarui Lu, Minsu Kim, Dinghuai Zhang, Jian Tang, Alexandre Piché, Nicolas Gontier, Yoshua Bengio, and Ehsan Kamalloo. 2025. Self-evolving curriculum for llm reasoning. *Preprint*, arXiv:2505.14970.

Sumanth Chennupati, Mohammad Mahdi Kamani, Zhongwei Cheng, and Lin Chen. 2021. Adaptive distillation: Aggregating knowledge from multiple paths for efficient distillation. *Preprint*, arXiv:2110.09674.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 4 others. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Elizabeth B. Dyer and Miriam Gamoran Sherin. 2016. Instructional reasoning about interpretations of student thinking that supports responsive teaching in secondary mathematics. *ZDM Mathematics Education*, 48(1-2):69–82.

Gongfan Fang, Jie Song, Xinchao Wang, Chengchao Shen, Xingen Wang, and Mingli Song. 2021. Contrastive model inversion for data-free knowledge distillation. *Preprint*, arXiv:2105.08584.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep bayesian active learning with image data. *Preprint*, arXiv:1703.02910.

Arthur C. Graesser, Shulan Lu, George Tanner Jackson, Heather Hite Mitchell, Mathew Ventura, Andrew Olney, and Max M. Louwerse. 2004. Autotutor: A tutor with dialogue in natural language. *Behavior Research Methods, Instruments, & Computers*, 36(2):180–192.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, and Abhinav Pandey. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Qiushan Guo, Xinjiang Wang, Yichao Wu, Zhipeng Yu, Ding Liang, Xiaolin Hu, and Ping Luo. 2020. Online knowledge distillation via collaborative learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhiwei Hao, Jianyuan Guo, Kai Han, Yehui Tang, Han Hu, Yunhe Wang, and Chang Xu. 2023. One-for-all: Bridge the gap between heterogeneous architectures in knowledge distillation. *Preprint*, arXiv:2310.19444.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *Preprint*, arXiv:1503.02531.

Md. Ismail Hossain, M M Lutfe Elahi, Sameera Ramasinghe, Ali Cheraghian, Fuad Rahman, Nabeel Mohammed, and Shafin Rahman. 2025. Luminet: Perception-driven knowledge distillation via statistical logit calibration. *Preprint*, arXiv:2310.03669.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. 2011. Bayesian active learning for classification and preference learning. *Preprint*, arXiv:1112.5745.

Wei-Ling Hsu, Yu-Chien Tang, and An-Zi Yen. 2025. Mathedu: Towards adaptive feedback for student mathematical problem-solving. *Preprint*, arXiv:2505.18056.

Jules King, Kennedy Smith, L Burleigh, Scott Crossley, Maggie Demkin, and Walter Reade. 2025. Map - charting student math misunderstandings. `https://kaggle.com/competitions/map-charting-student-math-misunderstandings`. Kaggle.

Andreas Kirsch, Joost van Amersfoort, and Yarin Gal. 2019. Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning. *Preprint*, arXiv:1906.08158.

Songhua Liu, Kai Wang, Xingyi Yang, Jingwen Ye, and Xinchao Wang. 2022. Dataset distillation via factorization. *Preprint*, arXiv:2210.16774.

Yunteng Luan, Hanyu Zhao, Zhi Yang, and Yafei Dai. 2019. Msd: Multi-self-distillation learning via multi-classifiers within deep neural networks. *Preprint*, arXiv:1911.09418.

Amir M. Mansourian, Rozhan Ahmadi, Masoud Ghafouri, Amir Mohammad Babaei, Elaheh Badali Golezani, Zeynab Yasamani Ghamchi, Vida Ramezanian, Alireza Taherian, Kimia Dinashi, Amirali Miri, and Shohreh Kasaei. 2025. A comprehensive survey on knowledge distillation. *Preprint*, arXiv:2503.12067.

Otero Nancy, Druga Stefania, and Lan Andrew. 2024. A benchmark for math misconceptions: Bridging gaps in middle school algebra with ai-supported instruction. *Preprint*, arXiv:2412.03765.

Nancy Otero, Stefania Druga, and Andrew Lan. 2025. A benchmark for math misconceptions: bridging gaps in middle school algebra with ai-supported instruction. *Discover Education*, 4(1):277.

Ni Parwati and I. Suharta. 2020. Effectiveness of the implementation of cognitive conflict strategy assisted by e-service learning to reduce students mathematical misconceptions. *International Journal of Emerging Technologies in Learning (iJET)*, 15(11):102–118.

Sebastian Raschka. 2020. Model evaluation, model selection, and algorithm selection in machine learning. *Preprint*, arXiv:1811.12808.

Philip M. Sadler, Gerhard Sonnert, Heather P. Coyle, Nancy Cook-Smith, and Jaimie L. Miller. 2013. The influence of teachers knowledge on student learning in middle school physical science classrooms. *American Educational Research Journal*, 50(5):1020–1049.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. *arXiv preprint arXiv:2302.00093*.

Jie Song, Ying Chen, Jingwen Ye, and Mingli Song. 2022. Spot-adaptive knowledge distillation. *IEEE Transactions on Image Processing*, 31:33593370.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. *Preprint*, arXiv:2402.13446.

Claude Team. 2025a. Introducing claude 4.

Gemma Team. 2024a. Gemma.

OPENAI Team. 2025b. Gpt-5 system card.

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

Qwen Team. 2025c. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Zichao Wang, Sebastian Tschiatschek, Simon Woodhead, José Miguel Hernández-Lobato, Simon Peyton Jones, and Cheng Zhang. 2020. Large-scale educational question analysis with partial variational auto-encoders. *CoRR*, abs/2003.05980.

Hongxu Yin, Pavlo Molchanov, Zhizhong Li, Jose M. Alvarez, Arun Mallya, Derek Hoiem, Niraj K. Jha, and Jan Kautz. 2020. Dreaming to distill: Data-free knowledge transfer via deepinversion. *Preprint*, arXiv:1912.08795.

Qianjin Yu, Keyu Wu, Zihan Chen, Chushu Zhang, Manlin Mei, Lingjun Huang, Fang Tan, Yongsheng Du, Kunlin Liu, and Yurui Zhu. 2025. Rethinking the generation of high-quality cot data from the perspective of llm-adaptive question difficulty grading. *Preprint*, arXiv:2504.11919.

Fei Yuan, Linjun Shou, Jian Pei, Wutao Lin, Ming Gong, Yan Fu, and Daxin Jiang. 2020. Reinforced multi-teacher selection for knowledge distillation. *Preprint*, arXiv:2012.06048.

A Adaptive Loss Weight Allocation

Table 4 presents the specific weighting strategy for the adaptive loss function based on sample categories. This mechanism organically combines sample difficulty characterization with label credibility assessment, and in particular, demonstrates a careful handling of real-world noise for $\mathcal{S}_{\text{HH}}^{\text{close}}$ samples. This is a key design choice that enhances model robustness in complex categories.

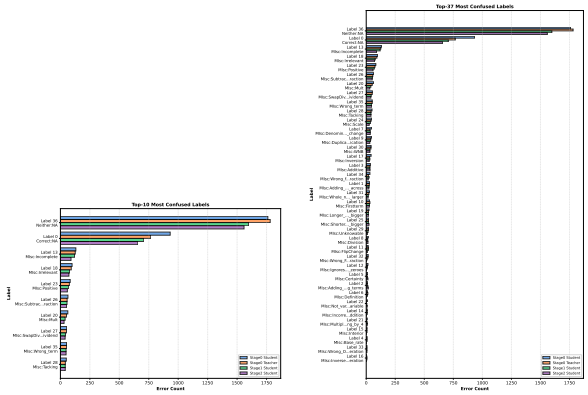
Table 4: Adaptive loss weight allocation strategy based on sample categories

Category	$\alpha(\mathcal{L}_{\text{CE}})$	$\beta(\mathcal{L}_{\text{KD}})$	$\gamma(\mathcal{L}_{\text{COS}})$
$\mathcal{S}_{\text{NM}}^{\text{close}}$	1.0	0.0	0.0
$\mathcal{S}_{\text{NM}}^{\text{far}}$	1.0	1.0	1.0
$\mathcal{S}_{\text{HH}}^{\text{close}}$	0.0	1.0	1.0
$\mathcal{S}_{\text{HH}}^{\text{far}}$	1.0	1.0	1.0

B Detailed Error Analysis Across Categories

This section provides comprehensive error analysis results for different numbers of categories. Figure 5 shows the error count changes for the top-10 most commonly confused categories and all 37 categories across different stage models.

The analysis reveals that the multi-stage distillation approach consistently reduces errors across all category groups. The improvement is particularly notable in the most frequently misclassified categories, where later-stage models demonstrate substantial error reduction compared to earlier stages.



(a) Error count changes in top-10 confused categories

(b) Error count changes in all 37 categories

Figure 5: Extended error analysis for different category groups

C LLM Model Configuration Details

Tables 5, 6 and 7 present the sampling parameters and inference configuration details for all state-of-the-art language models evaluated in our benchmark experiments.

Table 5: Sampling parameters for state-of-the-art language models (Part 1)

Model	Temperature	Max Tokens	Top-p
GPT-5	0.7	8192	0.9
DeepSeek-V3	0.3	8192	0.9
GPT-OSS-120B	0.7	8192	0.9
Claude Sonnet 4	0.3	8192	0.9
Qwen2.5-72B	0.7	8192	0.9

Table 6: Sampling parameters for state-of-the-art language models (Part 2)

Model	Repetition Penalty	Frequency Penalty	Presence Penalty
GPT-5	–	0	0
DeepSeek-V3	1.05	0	0
GPT-OSS-120B	–	0	0
Claude Sonnet 4	1.05	0	0
Qwen2.5-72B	1.05	0	0

Table 7: Inference configuration for state-of-the-art language models

Model	Batch Size	Rate Limit
GPT-5	500	2000
DeepSeek-V3	25	500
GPT-OSS-120B	500	5000
Claude Sonnet 4	500	2000
Qwen2.5-72B	50	1000

D Hyperparameter-Setting and Cross-Setting Generalization

Stage-1 Loss Weights. To facilitate reproducibility, we report our Stage-1 loss as a weighted combination of cross-entropy (CE), knowledge distillation (KD), and cosine similarity (COS): $\mathcal{L} = \alpha\mathcal{L}_{\text{CE}} + \beta\mathcal{L}_{\text{KD}} + \gamma\mathcal{L}_{\text{COS}}$, where $\alpha + \beta + \gamma = 1$. We performed a grid-style ablation across three student backbones. A balanced configuration $(\alpha, \beta, \gamma) = (0.33, 0.33, 0.34)$ (approximately 1:1:1) consistently yields the best (or near-best) performance across architectures and scales, suggesting that our hyperparameter choice is stable rather than overfitted.

α (CE)	β (KD)	γ (COS)	MAP@3	Acc.
0.25	0.25	0.50	0.9476	0.8988
0.25	0.50	0.25	0.9466	0.8969
0.50	0.25	0.25	0.9445	0.8927
0.20	0.20	0.60	0.9476	0.8991
0.20	0.60	0.20	0.9489	0.9013
0.60	0.20	0.20	0.9414	0.8870
0.20	0.40	0.40	0.9493	0.9020
0.40	0.20	0.40	0.9452	0.8943
0.40	0.40	0.20	0.9459	0.8958
0.33	**0.33**	**0.34**	**0.9495**	**0.9024**
0.00	0.50	0.50	0.9478	0.8995
0.50	0.00	0.50	0.9420	0.8886
0.50	0.50	0.00	0.9466	0.8969
0.00	0.00	1.00	0.9360	0.8838
0.00	1.00	0.00	0.9488	0.9010
1.00	0.00	0.00	0.9386	0.8821

Table 8: Stage-1 loss weight ablation on Qwen-3-4B.

α (CE)	β (KD)	γ (COS)	MAP@3	Acc.
0.25	0.25	0.50	0.9452	0.8940
0.25	0.50	0.25	0.9453	0.8942
0.50	0.25	0.25	0.9434	0.8903
0.20	0.20	0.60	0.9451	0.8938
0.20	0.60	0.20	0.9460	0.8957
0.60	0.20	0.20	0.9402	0.8846
0.20	0.40	0.40	0.9470	0.8975
0.40	0.20	0.40	0.9417	0.8873
0.40	0.40	0.20	0.9450	0.8931
0.33	**0.33**	**0.34**	**0.9474**	**0.8981**
0.00	0.50	0.50	0.9461	0.8961
0.50	0.00	0.50	0.9394	0.8836
0.50	0.50	0.00	0.9444	0.8924
0.00	0.00	1.00	0.5130	0.2904
0.00	1.00	0.00	0.9468	0.8969
1.00	0.00	0.00	0.9353	0.8753

Table 9: Stage-1 loss weight ablation on Gemma-2-9B.

α (CE)	β (KD)	γ (COS)	MAP@3	Acc.
0.25	0.25	0.50	0.9447	0.8928
0.25	0.50	0.25	0.9464	0.8962
0.50	0.25	0.25	0.9419	0.8876
0.20	0.20	0.60	0.9446	0.8934
0.20	0.60	0.20	0.9463	0.8958
0.60	0.20	0.20	0.9414	0.8868
0.20	0.40	0.40	0.9458	0.8955
0.40	0.20	0.40	0.9425	0.8891
0.40	0.40	0.20	0.9433	0.8901
0.33	**0.33**	**0.34**	**0.9467**	**0.8971**
0.00	0.50	0.50	0.9457	0.8951
0.50	0.00	0.50	0.9399	0.8846
0.50	0.50	0.00	0.9431	0.8901
0.00	0.00	1.00	0.4201	0.3957
0.00	1.00	0.00	0.9466	0.8968
1.00	0.00	0.00	0.9367	0.8788

Table 10: Stage-1 loss weight ablation on Llama-3.1-8B.

Stage-2 Adaptive Distillation. In Stage-2, we categorize high-value samples (e.g., NM vs. HH) and apply targeted loss compositions. Across three backbones, the complete adaptive strategy consistently outperforms uniform loss designs.

Method (Qwen-3-4B)	MAP@3	Acc.
All selected: CE only	0.9521	0.9085
All selected: CE+KD+COS	0.9536	0.9117
NM: CE; HH: KD+COS	0.9540	0.9123
NM: CE+KD+COS; HH: KD+COS	0.9574	0.9178
Complete method	**0.9585**	**0.9198**

Table 11: Stage-2 ablation on Qwen-3-4B.

Method (Gemma-2-9B)	MAP@3	Acc.
All selected: CE only	0.9503	0.9024
All selected: CE+KD+COS	0.9516	0.9107
NM: CE; HH: KD+COS	0.9524	0.9167
NM: CE+KD+COS; HH: KD+COS	0.9550	0.9139
Complete method	**0.9560**	**0.9148**

Table 12: Stage-2 ablation on Gemma-2-9B.

Method (Llama-3.1-8B)	MAP@3	Acc.
All selected: CE only	0.9502	0.9003
All selected: CE+KD+COS	0.9514	0.9067
NM: CE; HH: KD+COS	0.9514	0.9109
NM: CE+KD+COS; HH: KD+COS	0.9548	0.9124
Complete method	**0.9553**	**0.9134**

Table 13: Stage-2 ablation on Llama-3.1-8B.

Our adaptive approach that categorizes high-value samples into four types and applies targeted loss combinations consistently outperforms uniform loss strategies across all models, further demonstrating the robustness of our method. These extensive ablation studies demonstrate that: (1) our Stage 1 hyperparameters generalize well across different model architectures, and (2) our Stage 2 adaptive strategy provides consistent improvements. The high degree of cross-model consistency suggests that our hyperparameter choices are principled rather than overfitted to specific datasets.

Cross-validation for generalization. We report full K-fold(K=5) results under the selected balanced setting $(\alpha, \beta, \gamma) = (0.33, 0.33, 0.34)$.

Fold	α	β	γ	MAP@3	Acc.
fold0	0.33	0.33	0.34	0.9495	0.9024
fold1	0.33	0.33	0.34	0.9499	0.9020
fold2	0.33	0.33	0.34	0.9463	0.8962
fold3	0.33	0.33	0.34	0.9464	0.8969
fold4	0.33	0.33	0.34	0.9490	0.9013
Mean±Std	–	–	–	**0.9482±0.0017**	**0.8998±0.0028**

Table 14: Complete 5-fold results for Stage-1 (Qwen-3-4B).

Fold	α	β	γ	MAP@3	Acc.
fold0	0.33	0.33	0.34	0.9474	0.8981
fold1	0.33	0.33	0.34	0.9480	0.8986
fold2	0.33	0.33	0.34	0.9457	0.8947
fold3	0.33	0.33	0.34	0.9465	0.8968
fold4	0.33	0.33	0.34	0.9474	0.8973
Mean±Std	–	–	–	**0.9470±0.0009**	**0.8971±0.0015**

Table 15: Complete 5-fold results for Stage-1 (Gemma-2-9B).

Fold	α	β	γ	MAP@3	Acc.
fold0	0.33	0.33	0.34	0.9467	0.8971
fold1	0.33	0.33	0.34	0.9455	0.8943
fold2	0.33	0.33	0.34	0.9445	0.8907
fold3	0.33	0.33	0.34	0.9462	0.8972
fold4	0.33	0.33	0.34	0.9465	0.8966
Mean±Std	–	–	–	**0.9459±0.0009**	**0.8952±0.0026**

Table 16: Complete 5-fold results for Stage-1 (Llama-3.1-8B).

Method (Qwen-3-4B)	MAP@3	Acc.
fold0_exp5 (Complete method)	0.95845	0.91982
fold1_exp5	0.95911	0.92006
fold2_exp5	0.95692	0.91745
fold3_exp5	0.95877	0.91849
fold4_exp5	0.96031	0.92147
Mean±Std	**0.9567±0.0013**	**0.9195±0.0016**

Table 17: Complete 5-fold results for Stage-2 (Qwen-3-4B), complete method (exp5).

Method (Gemma-2-9B)	MAP@3	Acc.
fold0_exp5 (Complete method)	0.95600	0.91475
fold1_exp5	0.95798	0.91681
fold2_exp5	0.95658	0.91573
fold3_exp5	0.95731	0.91713
fold4_exp5	0.95827	0.91893
Mean±Std	**0.9572±0.0009**	**0.9167±0.0014**

Table 18: Complete 5-fold results for Stage-2 (Gemma-2-9B), complete method (exp5).

Method (Llama-3.1-8B)	MAP@3	Acc.
fold0_exp5 (Complete method)	0.95530	0.91338
fold1_exp5	0.95668	0.91574
fold2_exp5	0.95570	0.91328
fold3_exp5	0.95682	0.91403
fold4_exp5	0.95834	0.91867
Mean±Std	**0.9566±0.0011**	**0.9150±0.0021**

Table 19: Complete 5-fold results for Stage-2 (Llama-3.1-8B), complete method (exp5).

The consistent high performance across all folds and different model architectures validates the generalizabilit of our approach, while the small-dataset experiments confirm its practicality.

E Dependence on Teacher-Model Uncertainty

Our framework utilizes teacher cognitive uncertainty as a *guiding signal* rather than a hard constraint. It serves to identify high-value samples and modulate the student's reliance via difficulty-adaptive weighting. Crucially, when teacher signals are unreliable, the adaptive objective reduces their contribution in favor of ground-truth supervision, effectively preventing the inheritance of teacher flaws.

F Comparisons/Visualizations of Selected High-Value Samples

We provide side-by-side, box-rendered examples to illustrate how our selection distinguishes low-value vs. high-value samples and why different sample types require different supervision.

Case 1: Easy/Normal Sample (Low Learning Value)

QuestionText: What fraction of the shape is not shaded? Give your answer in its simplest form.
MC_Answer: 1/3
StudentExplanation: "one third is equal to tree nineth"
pred_top3: [0, 36, 13]
pred_probability: [0.910085, 0.060027, 0.002885]
label: 0
Interpretation: The teacher predicts correctly with high confidence (91%), indicating limited marginal training value.

Case 2: NM-close (Borderline Confusion)

QuestionText: What fraction of the shape is not shaded? Give your answer in its simplest form.
MC_Answer: 1/3
StudentExplanation: "Because its simplified from 3 ninth"
pred_top3: [36, 0, 13]
pred_probability: [0.532756, 0.428080, 0.004627]
label: 0
Interpretation Probabilities are close (decision boundary). These samples benefit from stronger ground-truth supervision (CE) to sharpen the boundary.

Case 3: NM-far (High-Confidence Misjudgment / Teacher Bias)

QuestionText: What fraction of the shape is not shaded? Give your answer in its simplest form.
MC_Answer: 1/3
StudentExplanation: "Because there are nine 3rds all together so simply that"
pred_top3: [36, 0, 13]
pred_probability: [0.936680, 0.043809, 0.003186]
label: 0
Interpretation. The teacher is confidently wrong. Joint supervision (ground truth + teacher representation) helps the student learn semantic features while correcting the teachers mistaken label preference.

Case 4: HH-near (Complex Semantics / Confusable Error Types)

QuestionText: Calculate 1/2 ÷ 6
MC_Answer: 3
StudentExplanation: "I think this because half of 6 is 3 and 2 divided by 6 is 3."
pred_top3: [20, 11, 27]
pred_probability: [0.877975, 0.029347, 0.017252]
label: 36
Interpretation. The explanation superficially matches multiple misconceptions, making it easy to confuse. For such samples, we prioritize teacher signals (KD+COS) to transfer richer semantic discrimination.

Case 5: HH-far (Diverse Expression / Teacher Blind Spot)

QuestionText: A/10 = 9/15. What is the value of A?
MC_Answer: 1/3
StudentExplanation: "because half is added so we got 9 from 6"
pred_top3: [36, 3, 10]
pred_probability: [0.912709, 0.037672, 0.008021]
label: 0
Interpretation. The teacher misjudges unconventional yet potentially valid reasoning with high confidence. These samples expose teacher blind spots; difficulty-adaptive weighting increases reliance on ground truth, enabling the student to surpass the teacher.

G Why the Student Can Outperform the Teacher

We observed that a smaller student can match or surpass a larger teacher on this task due to:

- **Teacher pretraining bias and cognitive blind spots.** Large general-purpose teachers may over-prefer standardized reasoning, while student explanations are often non-standard but self-consistent.
- **Task specialization via high-value selection.** Our selection focuses learning on uncertainty-revealing regions that are most diagnostic for misconception classification.
- **Adaptive correction of teacher errors.** When the teacher is confidently wrong (e.g., NM-far / HH-far), difficulty-adaptive weighting increases the contribution of ground-truth supervision, enabling the student to inherit general knowledge while correcting teacher-specific mistakes.

H Limitations: Small Dataset Scale and Data Availability

A key limitation is the **difficulty of collecting authentic student reasoning at scale**. While we mitigate overfitting concerns through stratified 5-fold validation on the 36k dataset and additionally verify practicality on a smaller curated set (220 samples), broader generalization is still constrained by: (i) limited public datasets that contain authentic student explanations with sufficient quality, and (ii) the inherent cost of obtaining large-scale real-world student reasoning. We view expanding dataset coverage (languages, curricula, demographics) as important future work.

I Prompt for Experiments

To validate the effectiveness of our method, we designed three different experimental scenarios, with corresponding prompts shown below. Through these three different types of prompts, we can comprehensively evaluate the models performance under both data synthesis and direct prediction modes.

Prompt: Synthetic Student Explanation Generator

You will be generating new student explanations that match a specific explanation type for a given math question. You need to simulate how different students at the same grade level might explain their reasoning when arriving at the same answer.
Here is the question, student's answer, answer correctness and an example student explanation:
<question> {QUESTIONTEXT} </question>

<answer> {ANSWER} </answer>

<answer_correctness> {ANSWER CORRECTNESS} </answer_correctness>

<student_explanation> {STUDENT EXPLANATION} </student_explanation>

Here is the explanation type you need to match and related student explanations from other students:
<explanation_type>
{STUDENT_EXPLANATION_TYPE}
</explanation_type>

<related_explanations>
{RELATED_STUDENT_EXPLANATION}
</related_explanations>

You need to generate {N} new student explanations.

EXPLANATION TYPE CATEGORIES:

- **Correct:NA** - The student's reasoning process is mathematically sound and leads logically to the correct answer
- **Neither:NA** - The student's explanation is vague, unclear, or unrelated to the mathematical concepts in the question
- **Misconception:[Specific type]** - The student has a specific mathematical misconception. The specific type describes what mathematical concept they misunderstand (e.g., "Incorrect_equivalent_fraction_addition")

INSTRUCTIONS:

1. Analyze the given explanation type to understand what kind of reasoning pattern you need to replicate
2. Study the related student explanations to understand the common patterns for this explanation type
3. Generate new explanations that:
- Lead to the same answer as provided
- Match the specified explanation type category
- Sound like they come from different students at the appropriate grade level
- Show variety in wording and approach while maintaining the same underlying reasoning pattern - Are age-appropriate in language and mathematical sophistication

<scratchpad>
Before generating the explanations, think through:
- What grade level is this question appropriate for?
- What is the specific reasoning pattern shown in the explanation type?
- How do the related explanations demonstrate this pattern?
- What variations in language and approach can I use while maintaining the same reasoning type?
- If it's a misconception, what is the specific mathematical error being made?
</scratchpad>

Generate {N} new student explanations that match the specified explanation type. Each explanation should be distinct but follow the same reasoning pattern. Present each explanation numbered and in separate tags:
<explanation_1> [First new student explanation] </explanation_1>
<explanation_2> [Second new student explanation] </explanation_2>
[Continue for all N explanations...]

Prompt: MAP-Charting dataset

SYSTEM PROMPT:
You are a mathematics education analysis expert. You need to analyze students' explanations for solving math problems, identify correct reasoning, vague statements, or specific misunderstandings, and categorize them into predefined types. Please remain professional and objective, focusing on the students' thought processes.

CLASSIFICATION PROMPT:
You will be analyzing a student's explanation for a math problem and classifying it into one of several predefined explanation types.
Here is the problem data:
<problem_data>
{PROBLEM_DATA}
</problem_data>

Here are the possible student explanation types with their categories and corresponding indices:
<student_explanation_types>
General Categories
- [0] **Correct:NA** - The student's problem-solving approach and process are correct
- [36] **Neither:NA** - The explanation is vague, unclear, or logically incomplete
Fraction Operation Misconceptions
- [1] **Misconception:Adding_across** - Adding numerators and denominators directly (e.g., 1/3 + 2/5 = 3/8)
- [7] **Misconception:Denominator-only_change** - Only denominator is changed (e.g., 1/3 + 2/5 = 3/15)
... (6 more fraction operation types)
Basic Operation Misconceptions
- [2] **Misconception:Adding_terms** - Using addition instead of multiplication (e.g., 2y = 24 interpreted as y = 22)
- [8] **Misconception:Division** - Misunderstanding division concept (e.g., confusing "of" with "œ")
... (5 more basic operation types)
Other Misconception Categories
- [9] **Misconception:Duplication** - Multiplying both numerator and denominator (e.g., 2/3 Œ 5 = 10/15)
- [19] **Misconception:Longer_is_bigger** - Believing more decimal places means larger numbers
- [22] **Misconception:Not_variable** - Not understanding variables (e.g., interpreting 2y as 2 and y)
... (18 more misconception types across various mathematical concepts)
</student_explanation_types>

Your task is to determine which explanation type from STUDENT_EXPLANATION_TYPE best matches the Student Explanation provided in the PROBLEM_DATA.
The category system works as follows:
- **Correct:NA** - The student's explanation process is correct
- **Neither:NA** - The student's explanation is vague, unclear, or irrelevant to the problem
- **Misconception:[specific_type]** - The student's explanation contains a specific misunderstanding or error in reasoning (e.g., "Misconception:Incorrect_equivalent_fraction_addition")
Before providing your final answer, analyze the student's explanation carefully in scratchpad tags. Consider:
1. What mathematical concepts or processes does the student's explanation involve?
2. Is the reasoning correct, incorrect, or unclear?
3. If incorrect, what specific type of misconception does it represent?
4. Which of the available explanation types best matches this analysis?
<scratchpad>
Your analysis here
</scratchpad>

Provide the three most likely explanation type indices in order from highest to lowest probability. Format your answer as [idx1, idx2, idx3, ...] where idx1 is the most likely match, idx2 is the second most likely, and idx3 is the third most likely.

Prompt: Algebra Misconceptions Benchmark

SYSTEM PROMPT:
You are a mathematics education diagnostic expert specializing in misconception analysis. Your task is to analyze students' final answers to math problems, identify whether the answer reflects specific mathematical misconceptions, and categorize them into predefined misconception types. Please remain professional and objective, focusing on the students' submitted answers and the underlying errors they reveal.
CLASSIFICATION PROMPT:
You will be analyzing a student's answer to a math problem and classifying it into one of several predefined misconception types.

Here is the problem data:
<problem_data>
{PROBLEM_DATA}
</problem_data>

Here are the possible misconception types with their categories and corresponding indices:
<misconception_types>
Representative Misconception Categories (55 total)
Proportional Relationships:
- [MaE01] when students don't understand how to represent proportional relationships
- [MaE02] Students misunderstand proportional relationships, not realizing parts must be equal
Fractions:
- [MaE06] when students inaccurately simplify fractions by guessing instead of dividing
- [MaE08] incorrectly add/subtract fractions by summing numerators and denominators separately
Decimals & Negatives:
- [MaE16] mistakenly position decimal point left of sum, assuming units/tenths combine separately
- [MaE18] when students are unsure of correct sign when adding positive and negative numbers
Ratios & Percentages:
- [MaE24] struggle to understand that ratios can compare same or different units
- [MaE29] incorrectly apply single proportion formula: (smaller)/(larger) = (x/100)
Operations:
- [MaE31] incorrectly assume commutative/associative properties apply to subtraction/division
- [MaE34] incorrectly perform operations left to right, neglecting order of operations
Functions & Graphing:
- [MaE38] struggle to grasp that linear function represents consistent rate of change
- [MaE43] struggle with plotting points, reversing x- and y-coordinates
Variables & Algebra:
- [MaE46] mistakenly perceive variables as labels/units, or associate value with alphabetical position
- [MaE51] misunderstand equal sign as "the answer is" rather than relationship between quantities
- [MaE55] struggle to recognize when to combine like terms (e.g., 4x+2x+x=7x)
... (42 more misconception types covering exponents, mixed numbers, division, patterns, slopes, equations, and various algebraic concepts)
</misconception_types>

Your task is to determine which misconception type best matches the Student Answer.
Before providing your final answer, analyze carefully in scratchpad tags. Consider:
1. What is the correct answer to this problem?
2. How does the student's answer differ from the correct answer?
3. What mathematical error or misconception could lead to this specific incorrect answer?
4. Which misconception types best explain the error pattern?
5. Are there alternative misconceptions that could also explain this answer?

<scratchpad>
Your analysis here
</scratchpad>

Provide the three most likely misconception type indices in order from highest to lowest probability.
Format your answer as [MaE01, MaE02, MaE03, ...] where MaE01 is the most likely match.