GateKD: Confidence-Gated Closed-Loop Distillation for Robust Reasoning

Anonymous ACL submission

Abstract

Distilling multi-step reasoning abilities from large language models (LLMs) into compact student models remains challenging due to noisy rationales, hallucinated supervision, and static teacher–student interactions. Existing reasoning distillation methods, including mentor-based approaches, predominantly operate in an open-loop manner, implicitly assuming uniform teacher reliability and consequently propagating erroneous intermediate reasoning. We propose **GateKD**, a confidence-gated closed-loop distillation framework that enables robust reasoning transfer by treating the teacher as a dynamic gatekeeper rather than a static oracle. GateKD introduces three complementary mechanisms: (i) confidence-gated soft supervision that selectively distills reliable predictive signals, (ii) gated hidden-state evolution that aligns intermediate representations only when teacher confidence is high, and (iii) reliability-filtered attention distillation that preserves stable reasoning structures while suppressing noisy patterns. These components jointly form a closed feedback loop in which teacher confidence continuously modulates the distillation process, reducing hallucination transfer and stabilizing student reasoning. Extensive experiments across commonsense, logical, and symbolic reasoning benchmarks, using T5 and Flan-T5 backbones of varying sizes, demonstrate that GateKD consistently outperforms strong open-loop distillation baselines. Notably, GateKD yields substantial gains in logical and symbolic reasoning, remains robust under low-resource distillation settings, and shows clear performance degradation when any gating component is removed. Our results highlight that confidence-gated closed-loop supervision is critical for building reliable and scalable small reasoning models.

1 Introduction

Large language models (LLMs) have demonstrated remarkable multi-step reasoning abilities when guided to explicitly articulate intermediate reasoning steps, a capability formalized through Chain-of-Thought (CoT) prompting (Wei et al., 2022). Subsequent studies show that reasoning reliability can be further enhanced by aggregating multiple reasoning trajectories, as in self-consistency decoding, which mitigates brittle or spurious inference paths (Wang et al., 2022). Complementary to sampling-based approaches, structured prompting strategies such as least-to-most prompting decompose complex problems into simpler subproblems, highlighting the central role of intermediate reasoning structure in effective inference (Zhou et al., 2022). Collectively, these advances suggest that **how** reasoning is constructed is as important as **what** prediction is produced.

Motivated by the high computational cost of LLMs, recent work has focused on transferring these reasoning capabilities to smaller and more efficient models via reasoning distillation. Early approaches distill soft labels or explicit reasoning traces from large teachers, demonstrating that students can acquire non-trivial reasoning skills despite limited capacity. Building on this line of research, Mentor-KD (Lee et al., 2024) introduces a task-specific mentor model that augments both rationales and soft supervision, substantially improving multi-step reasoning distillation, particularly under low-resource settings. This work reveals a key insight: **intermediate supervision tailored to the task can be more effective than relying solely on large, generic LLM teachers**.

Despite these advances, existing reasoning distillation methods—including Mentor-KD—predominantly operate in an *open-loop* fashion. Teacher or mentor signals are treated as uniformly reliable and transferred wholesale to the student. In practice, however, even strong teachers exhibit fluctuating confidence across inputs, producing hallucinated reasoning steps, unstable intermediate representations, or misleading attention patterns. Blindly distilling such signals risks amplifying noise, especially for small models that lack the capacity to recover from erroneous supervision. This observation raises a fundamental question: *should all teacher reasoning be trusted equally during distillation?*

In this work, we answer this question negatively and propose **GateKD**, a confidence-gated closed-loop distillation framework for reasoning transfer. The key idea behind GateKD is simple yet powerful: instead of passively absorbing teacher signals, the student selectively learns from them based on the teacher's predictive confidence. Specifically, GateKD estimates teacher reliability via predictive entropy and uses this signal to dynamically gate three complementary distillation pathways: **(i)** soft-label supervision, **(ii)** hidden-state alignment, and **(iii)** attention distillation. As a result, reliable reasoning patterns are reinforced, while uncertain or unstable signals are suppressed.

Our design is inspired by Mentor-KD (Lee et al., 2024), but departs from it in a crucial way. While Mentor-KD focuses on *where* to source better supervision—by introducing a task-specialized mentor—GateKD focuses on *when and how* supervision should be transferred. By incorporating confidence-aware gating, GateKD transforms reasoning distillation from a static, open-loop procedure into a closed-loop interaction between teacher and student. This perspective aligns with the intuition underlying self-consistency and structured prompting: robust reasoning does not arise from a single trajectory, but from selectively trusting stable and coherent inference paths.

Empirically, we demonstrate that GateKD consistently outperforms strong open-loop distillation baselines across commonsense, logical, and symbolic reasoning benchmarks, and across multiple student model scales. Notably, the gains are most pronounced on logical and symbolic tasks, where erroneous intermediate reasoning is particularly detrimental. These results indicate that **selective reasoning transfer**, rather than indiscriminate distillation, is critical for building reliable small-scale reasoners.

In summary, our contributions are threefold: **(i)** we identify confidence misalignment as a key limitation of existing reasoning distillation methods; **(ii)** we propose GateKD, a unified confidence-gated framework that selectively distills reliable reasoning signals; and **(iii)** we empirically show that closed-loop, confidence-aware distillation yields robust and scalable reasoning improvements for small language models.

2 Related Work

2.1 Reasoning in Large Language Models

Large language models (LLMs) have demonstrated strong multi-step reasoning capabilities when prompted to explicitly generate intermediate reasoning steps, a phenomenon formalized as Chain-of-Thought (CoT) prompting (Wei et al., 2022). Subsequent work shows that aggregating multiple reasoning trajectories via self-consistency decoding improves robustness by mitigating brittle or spurious inference paths (Wang et al., 2022). Complementary to sampling-based approaches, structured prompting strategies such as least-to-most prompting decompose complex reasoning problems into simpler subproblems, highlighting the importance of intermediate reasoning structure in effective inference (Zhou et al., 2022). These techniques collectively suggest that reasoning quality depends not only on model capacity, but also on the stability and coherence of intermediate reasoning processes. The strong reasoning abilities exhibited by large proprietary models such as GPT-4 further reinforce this observation (OpenAI, 2023).

2.2 Reasoning Distillation for Small Language Models

Motivated by the computational cost of LLMs, a growing body of work explores distilling reasoning abilities into smaller student models. Early studies demonstrate that distilling soft labels or explicit reasoning traces enables small models to acquire non-trivial reasoning skills (Ho et al., 2023; Deng et al., 2023). Program-aided distillation (PaD) further shows that structured program supervision can outperform naive CoT fine-tuning for reasoning transfer (Zhu et al., 2023). Mixed Distillation combines heterogeneous teacher signals, including logits and rationales, to improve student generalization (Li et al., 2024a). Other works explore architectural or parameter-efficient strategies, such as mixture-of-experts distillation (Li et al., 2024b) and multilingual reasoning distillation (Payoungkhamdee et al., 2024).

Several recent studies investigate how reasoning supervision should be structured. Cascading decomposed CoT distillation improves generalization by progressively transferring simpler reasoning steps (Dai et al., 2024), while StepER performs step-wise distillation in retrieval-augmented reasoning settings (Lee et al., 2025). CODI compresses explicit CoT into continuous representations via self-distillation, reducing inference cost while preserving reasoning behavior (Shen et al., 2025). Reinforcement-learning-based distillation methods further attempt to uncover implicit multi-branch reasoning structures within teachers (Xu et al., 2025). Comprehensive analyses of CoT distillation identify key factors affecting reasoning transfer, including rationale quality, supervision granularity, and task difficulty (Chen et al., 2025b,a).

2.3 Mentor-Based and Multi-Path Distillation

Mentor-based distillation has emerged as a promising direction for improving reasoning transfer. Mentor-KD introduces a task-specific mentor model to provide refined rationales and soft supervision, significantly improving multi-step reasoning distillation under limited data regimes (Lee et al., 2024). Related work explores learning from diverse reasoning paths via routing and collaboration mechanisms, emphasizing that not all reasoning trajectories are equally informative (Lei et al., 2025). These approaches highlight the importance of intermediate supervision quality, but still largely assume that provided teacher or mentor signals are uniformly reliable.

2.4 Limitations of Open-Loop Distillation

Despite their success, most existing reasoning distillation methods operate in an *open-loop* fashion, treating teacher supervision as static and equally trustworthy across inputs. Empirical analyses show that even strong teachers frequently generate hallucinated or unstable reasoning steps, particularly for challenging logical and symbolic tasks (Hsieh et al., 2023; Song et al., 2025). Blindly distilling such signals risks amplifying noise and degrading student robustness, especially for small models with limited capacity to recover from erroneous supervision.

2.5 Our Position

In contrast to prior work, **GateKD** introduces a confidence-gated closed-loop distillation paradigm that dynamically regulates *when* and *how* teacher reasoning signals are transferred. While Mentor-KD focuses on improving the *source* of supervision via mentor models (Lee et al., 2024), GateKD focuses on selectively trusting supervision based on teacher confidence. By explicitly modeling supervision reliability and gating soft labels, hidden states, and attention patterns, GateKD complements existing reasoning distillation approaches and addresses a fundamental limitation of open-loop reasoning transfer.

3 Approach

We introduce **GateKD**, a *confidence-gated closed-loop distillation framework* for robust multi-step reasoning transfer from large teacher models to compact student models. Unlike prior reasoning distillation approaches that treat teacher supervision as uniformly reliable and static, GateKD dynamically modulates the influence of teacher signals based on their estimated reliability, forming an implicit closed feedback loop during training.

Figure 1 illustrates the overall architecture of GateKD. Given an input instance, both teacher and student models process the input in parallel. The teacher additionally produces a confidence signal that selectively gates which outputs, hidden representations, and attention structures are distilled to the student. This design prevents hallucinated or unstable reasoning trajectories from being propagated, while preserving reliable reasoning knowledge.

3.1 Overview of GateKD

GateKD operates on three complementary distillation pathways:

1. **Confidence-Gated Soft Supervision**, which adaptively reweights teacher predictive distributions;
2. **Gated Hidden-State Evolution**, which selectively aligns intermediate reasoning representations; and
3. **Reliability-Filtered Attention Distillation**, which transfers structural reasoning patterns only when reliable.

All three pathways are jointly optimized alongside the standard task loss, allowing the student to learn accurate predictions while gradually acquiring stable reasoning behaviors. Crucially, the gating mechanism is driven by teacher uncertainty estimated at training time, enabling dynamic modulation rather than fixed heuristics.

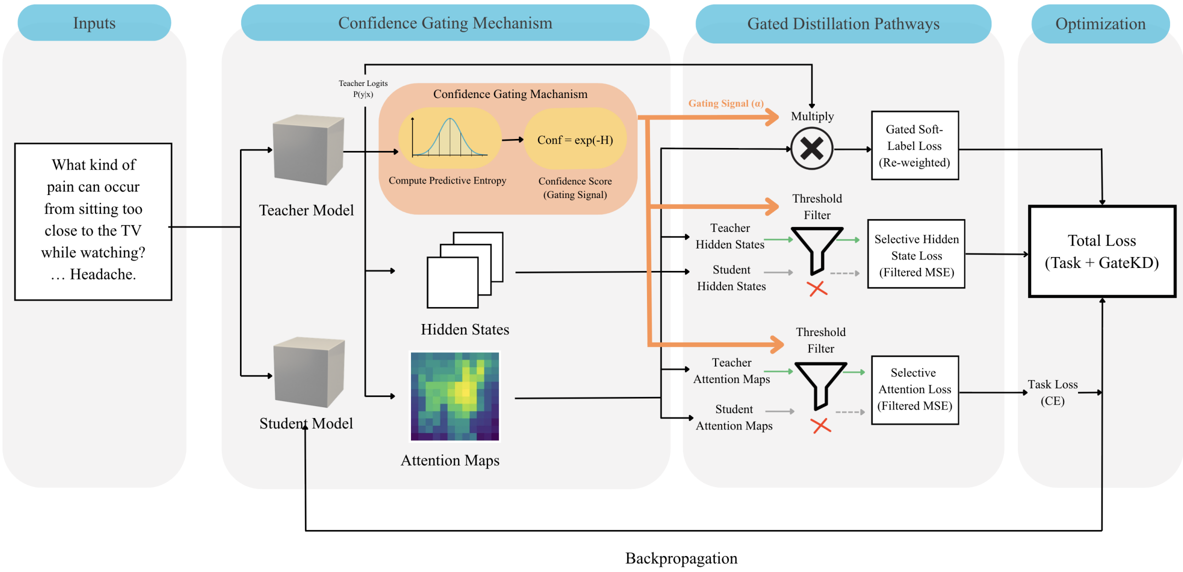

Figure 1: Overview of the proposed **GateKD** framework. Given an input, the teacher and student models process the instance in parallel. The teacher produces predictive distributions, hidden states, and attention maps, along with a confidence score estimated via predictive entropy. This confidence signal acts as a unified gating mechanism that selectively controls three distillation pathways: (i) *confidence-gated soft supervision*, where teacher soft labels are re-weighted by confidence; (ii) *gated hidden-state evolution*, where intermediate representations are aligned only when teacher confidence exceeds a threshold; and (iii) *reliability-filtered attention distillation*, which transfers structural reasoning patterns selectively. All gated distillation losses are jointly optimized with the task loss, forming an implicit closed-loop interaction that suppresses unreliable teacher signals while preserving stable reasoning knowledge.

3.2 Confidence Estimation and Gating Signal

For each input x, the teacher model $\mathcal{T}$ produces a probability distribution $p^{\mathcal{T}} = \text{softmax}(z^{\mathcal{T}})$. We estimate teacher confidence using predictive entropy:

$$C(x) = \exp\left(-\sum_i p_i^{\mathcal{T}} \log p_i^{\mathcal{T}}\right), \qquad (1)$$

where lower entropy corresponds to higher confidence.

This scalar confidence score serves as a *unified gating signal* that modulates all distillation pathways. Intuitively, when the teacher is uncertain, its supervision is downweighted or filtered out, preventing unreliable reasoning signals from influencing the student.

3.3 Confidence-Gated Soft Supervision

Standard knowledge distillation assumes that teacher predictions are equally informative for all training instances. However, in reasoning-intensive tasks, teacher outputs may be unreliable due to spurious reasoning paths or hallucinated intermediate steps.

GateKD addresses this issue by weighting the soft-label distillation loss using the confidence signal:

$$\mathcal{L}_{\text{gate-soft}} = \mathbb{E}_x\left[C(x) \cdot \text{CE}\left(p^{\mathcal{T}}, \log p^{\mathcal{S}}\right)\right], \qquad (2)$$

where $p^{\mathcal{S}}$ denotes the student's predictive distribution.

This formulation softly suppresses low-confidence supervision rather than discarding it entirely, leading to smoother optimization and improved robustness compared to hard filtering strategies.

3.4 Gated Hidden-State Evolution

Beyond output-level supervision, intermediate hidden representations encode step-by-step reasoning trajectories. Blindly distilling these representations can amplify teacher errors and destabilize training.

To mitigate this issue, GateKD selectively aligns hidden states only when the teacher is sufficiently confident. Let $\phi(\cdot)$ denote a projection layer that maps teacher representations to the student's hidden space. The gated hidden-state loss is defined as:

$$\mathcal{L}_{\text{gate-hid}} = \sum_{k=1}^{L_S} \mathbb{I}\left[C_k > \bar{C}\right] \left\| h_k^{\mathcal{S}} - \phi\left(h_{\alpha(k)}^{\mathcal{T}}\right) \right\|_2^2, \tag{3}$$

where $\alpha(k)$ aligns student and teacher layers, and $\bar{C}$ is the batch-level mean confidence.

This selective alignment ensures that the student internalizes only stable reasoning representations, while allowing it to independently develop alternative representations when teacher uncertainty is high.

3.5 Reliability-Filtered Attention Distillation

Attention maps encode structural reasoning patterns such as dependency tracking and variable binding. However, attention behavior is particularly sensitive to noisy or uncertain predictions.

GateKD applies the same confidence-based gating to attention alignment:

$$\mathcal{L}_{\text{gate-att}} = \sum_{k} \mathbb{I}\left[C_k > \bar{C}\right] \left\| A_k^{\mathcal{S}} - A_{\alpha(k)}^{\mathcal{T}} \right\|_2^2. \tag{4}$$

By distilling attention maps only when the teacher's reasoning is reliable, GateKD preserves meaningful structural patterns while suppressing spurious attention behaviors.

3.6 Overall Training Objective and Closed-Loop Interpretation

The final training objective combines task supervision with gated distillation losses:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda_1 \mathcal{L}_{\text{gate-soft}} + \lambda_2 \mathcal{L}_{\text{gate-hid}} + \lambda_3 \mathcal{L}_{\text{gate-att}}. \tag{5}$$

Although GateKD does not explicitly retrain the teacher, the confidence-gated mechanisms induce an implicit *closed-loop* interaction. As the student improves over training, reliable teacher signals increasingly dominate the distillation process, while unreliable supervision is progressively suppressed. This dynamic stands in contrast to open-loop distillation methods, where teacher guidance remains static throughout training.

Overall, GateKD enables robust and selective reasoning transfer, allowing compact student models to acquire stable multi-step reasoning capabilities without inheriting teacher hallucinations.

4 Experiments

We evaluate **GateKD** on a diverse set of reasoning benchmarks covering *commonsense*, *logical*, and *symbolic* reasoning. Following prior work on reasoning distillation (Lee et al., 2024), we adopt T5 and Flan-T5 model families as student backbones and employ `large` variants as task-specific teachers. All student models are trained using identical data splits and optimization settings for fair comparison across distillation strategies.

Tasks. Commonsense reasoning is evaluated on CSQA and StrategyQA (SQA). Logical reasoning is measured using the Shuffled Objects task, and symbolic reasoning is assessed via the Last Letter concatenation task. All metrics are reported in accuracy.

Baselines. We compare GateKD with several representative open-loop distillation methods, including Vanilla-KD, MCC-KD (Lee et al., 2024), and Mentor-KD (Lee et al., 2024). In addition, we report zero-shot CoT performance of GPT-4o-mini as an upper-bound reference. All student results are averaged over 5 random seeds.

4.1 Main Results

Tables 1 and 2 summarize the main experimental results across T5 and Flan-T5 backbones, respectively. We observe that **GateKD consistently outperforms all open-loop distillation baselines** across model scales and reasoning categories.

On **T5-based students** (Table 1), GateKD yields substantial improvements over Mentor-KD, with gains of up to **+4.9** points on logical reasoning and **+4.7** points on symbolic reasoning for the `T5-small` model. Notably, these improvements become more pronounced as the student capacity decreases, highlighting GateKD's effectiveness under severe capacity gaps.

Similar trends are observed for **Flan-T5 backbones** (Table 2). GateKD achieves the strongest overall performance across all tasks, with particularly large gains on logical and symbolic benchmarks. For example, GateKD improves over Mentor-KD by **+3.8** and **+4.4** points on the Shuffled and Last Letter tasks, respectively, when using FlanT5-small.

Overall, these results demonstrate that **confidence-gated distillation enables more reliable knowledge transfer** than existing open-loop approaches, especially for reasoning tasks that are sensitive to noisy or inconsistent teacher signals.

Model	#Params	Method	Commonsense		Logical	Symbolic
			CSQA	SQA	Shuffled	Last Letter
GPT-4o-mini	–	ZS-CoT (teacher)	76.8	61.4	83.1	71.2
T5-large	780M	Vanilla-KD	69.3 ± 0.4	58.6 ± 0.6	88.1 ± 0.3	69.0 ± 0.5
T5-base	250M	Vanilla-KD	61.9 ± 0.5	55.2 ± 0.7	78.4 ± 0.6	56.1 ± 0.8
		MCC-KD	63.0 ± 0.6	56.4 ± 0.5	81.0 ± 0.7	58.2 ± 0.6
		Mentor-KD	64.2 ± 0.4	57.6 ± 0.6	84.9 ± 0.5	61.0 ± 0.7
		GateKD (ours)	**66.8 ± 0.3**	**59.9 ± 0.4**	**90.6 ± 0.4**	**65.7 ± 0.5**
T5-small	80M	Vanilla-KD	55.4 ± 0.6	48.9 ± 0.8	63.7 ± 0.7	49.6 ± 0.9
		MCC-KD	56.8 ± 0.7	49.5 ± 0.6	66.1 ± 0.6	51.3 ± 0.7
		Mentor-KD	58.6 ± 0.5	51.8 ± 0.7	72.9 ± 0.5	55.2 ± 0.6
		GateKD (ours)	**61.3 ± 0.4**	**54.6 ± 0.5**	**80.8 ± 0.4**	**60.1 ± 0.5**

Table 1: Performance comparison on commonsense, logical, and symbolic reasoning benchmarks using T5 backbones. All student results are averaged over 5 runs with different random seeds (mean ± std). GateKD consistently outperforms open-loop distillation baselines, demonstrating robust reasoning transfer under confidence-gated supervision.

Model	#Params	Method	Commonsense		Logical	Symbolic
			CSQA	SQA	Shuffled	Last Letter
GPT-4o-mini	–	ZS-CoT (teacher)	77.6	62.1	84.3	72.4
FlanT5-large	780M	Vanilla-KD	71.2 ± 0.3	60.1 ± 0.4	89.7 ± 0.4	70.8 ± 0.5
FlanT5-base	250M	Vanilla-KD	63.8 ± 0.5	57.2 ± 0.6	82.6 ± 0.6	58.9 ± 0.7
		MCC-KD	65.4 ± 0.4	58.8 ± 0.5	85.1 ± 0.5	60.3 ± 0.6
		Mentor-KD	66.9 ± 0.4	60.0 ± 0.5	88.3 ± 0.4	63.8 ± 0.6
		GateKD (ours)	**69.5 ± 0.3**	**62.4 ± 0.4**	**92.1 ± 0.3**	**67.9 ± 0.4**
FlanT5-small	80M	Vanilla-KD	57.0 ± 0.6	50.1 ± 0.7	68.2 ± 0.7	52.4 ± 0.8
		MCC-KD	58.6 ± 0.5	51.3 ± 0.6	70.0 ± 0.6	54.0 ± 0.7
		Mentor-KD	60.4 ± 0.4	53.7 ± 0.5	76.4 ± 0.5	58.1 ± 0.6
		GateKD (ours)	**63.2 ± 0.3**	**56.1 ± 0.4**	**83.7 ± 0.4**	**62.5 ± 0.5**

Table 2: Results on Flan-T5 backbones. GateKD yields consistent gains across model scales, with particularly strong improvements on logical and symbolic reasoning tasks.

5 Analysis

To better understand the behavior of GateKD, we conduct a series of targeted analyses that examine its generalization and internal mechanisms. Specifically, we investigate whether GateKD consistently improves reasoning performance across different student architectures and model scales, and analyze how each confidence-gated component contributes to the overall effectiveness of the framework. These analyses provide deeper insight into when and why GateKD succeeds beyond aggregate benchmark results.

5.1 Generalization Across Student Models (RQ1)

GateKD is evaluated on multiple student backbones with varying capacities, including T5 and Flan-T5 in {`base`, `small`} configurations. As shown in Tables 1 and 2, GateKD consistently improves performance across all evaluated settings.

Importantly, the gains are not confined to a specific architecture or task type. GateKD improves commonsense reasoning accuracy while simultaneously delivering larger relative gains on logical and symbolic tasks, which are known to be more vulnerable to unreliable teacher supervision. These findings indicate that GateKD generalizes well across both model families and reasoning paradigms.

5.2 Ablation Study (RQ2)

To analyze the contribution of each gating mechanism, we perform ablation experiments on the `T5-small` model. The results are reported in Table 3.

Removing any gating component leads to a clear

Model	Method	Shuffled	Last Letter
T5-small	**GateKD (ours)**	**80.8 $\pm$ 0.4**	**60.1 $\pm$ 0.5**
	w/o confidence gating	72.3 $\pm$ 0.6	55.4 $\pm$ 0.7
	w/o hidden-state gate	75.6 $\pm$ 0.5	56.8 $\pm$ 0.6
	w/o attention gate	77.1 $\pm$ 0.4	58.2 $\pm$ 0.6

Table 3: Ablation study on logical and symbolic reasoning. Removing any gating component degrades performance, highlighting the complementary role of confidence-gated soft labels, hidden states, and attention alignment.

performance degradation. In particular, disabling confidence gating causes the largest drop, confirming the importance of suppressing low-confidence teacher signals. Ablating hidden-state or attention gating also degrades performance, indicating that intermediate representation alignment and structural reasoning transfer play complementary roles.

These results validate that GateKD's improvements arise from the *synergistic interaction* of confidence-gated soft supervision, hidden-state alignment, and attention distillation, rather than from any single component in isolation.

5.3 Expertise-Oriented Qualitative Analysis

Beyond quantitative gains, GateKD yields more *expertise-aligned* reasoning behavior by selectively distilling reliable intermediate signals. Figure 2 presents a representative example from StrategyQA, where the correct answer is a binary factual judgment.

In this example, the teacher model generates a verbose reasoning trace that explores hypothetical construction scenarios and advanced technologies. Although the reasoning appears fluent and detailed, it ultimately concludes that building a house on an asteroid is *theoretically possible*, leading to an incorrect prediction. This illustrates a common failure mode of large models: speculative reasoning that prioritizes plausibility over physical and practical constraints.

In contrast, the mentor model produces a more grounded reasoning trajectory that emphasizes feasibility, material limitations, gravity, and sustainability. While acknowledging theoretical possibilities, the mentor explicitly distinguishes between *technical feasibility* and *practical realizability*, ultimately arriving at the correct answer. This form of reasoning reflects expert judgment, where constraints and real-world viability dominate speculative extrapolation.

GateKD enables the student to internalize such expert-aligned reasoning by confidence-gated supervision. In this case, the teacher exhibits low confidence—as evidenced by high predictive uncertainty—despite producing a fluent explanation. GateKD therefore suppresses unreliable teacher signals and instead prioritizes stable mentor supervision. As a result, the distilled model learns to avoid overconfident speculative reasoning and favors constraint-aware inference patterns.

This qualitative example highlights a key advantage of GateKD: it does not merely improve final answer accuracy, but also shapes *how* reasoning is performed. By selectively trusting reliable reasoning trajectories, GateKD promotes cautious, physically grounded, and expertise-consistent inference, which is particularly critical for binary and commonsense reasoning tasks where hallucinated explanations can be misleading.

6 Conclusion

We introduced **GateKD**, a confidence-gated closed-loop distillation framework for transferring multi-step reasoning abilities from large language models to compact student models. Unlike prior open-loop reasoning distillation methods that treat teacher supervision as uniformly reliable, GateKD dynamically regulates the flow of supervision based on teacher confidence, selectively distilling soft labels, hidden representations, and attention structures. This design mitigates hallucination transfer and stabilizes reasoning acquisition in small models.

Extensive experiments across commonsense, logical, and symbolic reasoning benchmarks demonstrate that GateKD consistently outperforms strong open-loop baselines, particularly in challenging settings where erroneous intermediate reasoning is most harmful. Ablation studies further confirm that confidence-aware gating is essential to these gains, highlighting the importance of selectively trusting teacher reasoning rather than indiscriminately distilling it.

More broadly, our results suggest that effective reasoning transfer requires moving beyond static teacher–student pipelines toward adaptive, reliability-aware supervision. We hope this work encourages future research on closed-loop and confidence-driven learning paradigms for building robust, efficient, and trustworthy reasoning models.

StrategyQA		
Question: Can I build a house on an asteroid?		
Answer: No		
Role	Content	Prediction
Teacher	First, let's consider the feasibility of building a house on an asteroid. While there are currently no known bodies with the capability to build structures, there are potential alternatives such as using existing structures or creating artificial structures using advanced technology. Next, we would need to consider the size and composition of the asteroid. Asteroids are small, irregular objects with very low gravity and may have very different structural and environmental characteristics from terrestrial bodies. Additionally, the composition of an asteroid can also play a role in the design of the house. Some asteroids have very thin atmospheres and lack the necessary structural strength to support human habitation. **Based on these considerations, it is theoretically possible to build a house on an asteroid with the right technology and resources. However, it would require significant resources, such as materials, technology, and space to construct and sustain a house.**	✗ Yes
Mentor	Building a house on an asteroid would be a complex and challenging task. Here are some steps to consider: 1. Selection of an appropriate asteroid: It is important to consider factors such as the size of the asteroid, gravity, and potential for habitats. The asteroid itself may not be suitable for building a house due to its small size and lack of suitable resources. 2. Location of the asteroid: Asteroids are small in size and have limited access to basic infrastructure compared to larger bodies. Finding a suitable asteroid with suitable orbits and habitats would be challenging. 3. Construction materials: Asteroids are typically made of lightweight materials like sulfide and titanium. These materials are not strong enough to support the weight and structure of a house. 4. Sustainability: Sustainability is an important aspect of building a house on an asteroid. Solar power and other renewable energy sources would be necessary to sustain the asteroid. 5. Energy source: Energy sources on asteroid bodies may be generated by burning fossil fuels or by using solar panels. **In conclusion, while it may be technically possible to build a house on an asteroid, it is not a practical or sustainable project. It would require careful planning, technology, and resources to ensure sustainability and sustainability of the space.**	✓ No

Figure 2: **Qualitative comparison of reasoning behaviors on StrategyQA.** The teacher model produces a fluent but speculative reasoning trace and incorrectly predicts "Yes," conflating theoretical possibility with practical feasibility. In contrast, the mentor model emphasizes physical constraints, sustainability, and real-world viability, leading to the correct prediction "No." GateKD selectively suppresses low-confidence teacher supervision and prioritizes reliable mentor reasoning, enabling the student to acquire more expertise-aligned, constraint-aware inference patterns rather than overconfident speculation.

7 Limitations

While GateKD demonstrates consistent improvements across a range of reasoning benchmarks, several limitations remain. First, GateKD relies on teacher confidence estimation, which we approximate using predictive entropy. Although this signal is effective in practice, confidence estimation is not perfect and may still be miscalibrated, particularly for out-of-distribution inputs or highly ambiguous problems. Exploring more robust or learned confidence estimators is a promising direction for future work.

Second, our framework assumes access to intermediate teacher representations, such as hidden states and attention maps. This limits direct applicability to black-box or API-only models, where such internal signals are unavailable. Extending confidence-gated distillation to settings with restricted teacher access remains an open challenge.

Third, GateKD introduces additional computational overhead during training due to confidence estimation and gated alignment across multiple representation levels. While this cost is incurred only during distillation and not at inference time, it may pose challenges for extremely large-scale training scenarios.

Finally, our experiments focus on established reasoning benchmarks in English. We do not explicitly evaluate multilingual, multimodal, or real-world interactive reasoning settings, where confidence dynamics and reasoning structures may differ. We leave these extensions to future work.

References

General Chen et al. 2025a. Effectiveness of chain-of-thought in distilling reasoning capability from large teacher models. In *Proceedings of the 2025 International Natural Language Generation Conference*, pages 49–60.

Xinghao Chen, Zhijing Sun, Wenjin Guo, Miaoran Zhang, et al. 2025b. Unveiling the key factors for distilling chain-of-thought reasoning. pages 15094–15119.

Chengwei Dai, Kun Li, Wei Zhou, and Songlin Hu. 2024. Improve student's reasoning generalizability through cascading decomposed cots distillation. *arXiv preprint*, arXiv:2405.19842.

Yuntian Deng, Kiran Prasad, Roland Fernandez, et al. 2023. Implicit chain of thought reasoning via knowledge distillation. In *ArXiv preprint*.

Nathan Ho, Madhu Shridhar, et al. 2023. Teaching small language models to reason. In *ArXiv preprint*.

Alex Hsieh et al. 2023. Rationale-based knowledge distillation for neural reasoning. *Artificial Intelligence Review*. Review of rationale distillation techniques for LLMs.

Hojae Lee, Junho Kim, and SangKeun Lee. 2024. Mentor-kd: Making small language models better multi-step reasoners. *arXiv preprint arXiv:2410.09037*.

Kyumin Lee, Minjin Jeon, Sanghwan Jang, and Hwanjo Yu. 2025. Steper: Step-wise knowledge distillation for enhancing reasoning ability in multi-step retrieval-augmented language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29501–29523. Association for Computational Linguistics.

Zhenyu Lei, Zhen Tan, Song Wang, et al. 2025. Learning from diverse reasoning paths with routing and collaboration. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2832–2845. Association for Computational Linguistics.

Chenglin Li, Qianglong Chen, Liangyue Li, Caiyu Wang, Yicheng Li, et al. 2024a. Mixed distillation helps smaller language models reason better. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 375–392. Association for Computational Linguistics.

Xiang Li, Shizhu He, Jiayu Wu, Zhao Yang, Yao Xu, et al. 2024b. Mode-cotd: Chain-of-thought distillation for complex reasoning tasks with mixture of decoupled lora-experts. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics (LREC-COLING)*, pages 11475–11485. ELRA.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint*, arXiv:2303.08774.

Patomporn Payoungkhamdee, Peerat Limkonchotiwat, Jinheon Baek, et al. 2024. An empirical study of multilingual reasoning distillation for question answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 7739–7751. Association for Computational Linguistics.

Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. 2025. Codi: Compressing chain-of-thought into continuous space via self-distillation. *arXiv preprint*, arXiv:2502.21074.

Author Song et al. 2025. Knowledge distillation and dataset distillation of large language models: emerging trends, challenges, and future directions. *Artificial Intelligence Review*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, et al. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint*, arXiv:2203.11171. Published at ICLR 2023.

Jason Wei, Xuezhi Wang, Dale Schuurmans, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*.

Shicheng Xu, Liang Pang, Yunchang Zhu, et al. 2025. Distilling the implicit multi-branch structure in llms' reasoning via reinforcement learning. *arXiv preprint*, arXiv:2505.16142.

Denny Zhou, Nathanael Schärli, Le Hou, et al. 2022. Least-to-most prompting enables complex reasoning in large language models. In *arXiv preprint*.

Xuekai Zhu, Biqing Qi, Kaiyan Zhang, Xinwei Long, Zhouhan Lin, and Bowen Zhou. 2023. Pad: Program-aided distillation can teach small models reasoning better than chain-of-thought fine-tuning. *arXiv preprint*, arXiv:2305.13888.

A Mathematical Formulation and Analysis of GateKD

This appendix provides a formal description of the confidence-gated distillation mechanism underlying GateKD and offers theoretical intuition for its effectiveness in stabilizing reasoning transfer.

A.1 Preliminaries

Let x denote an input question and $y \in \mathcal{Y}$ the corresponding output space. Let T and S represent the teacher and student models, respectively. For autoregressive decoding, the teacher defines a predictive distribution

$$p_T(y \mid x) = \prod_{t=1}^{L} p_T(y_t \mid y_{<t}, x),$$

where L is the output length. Similarly, the student defines $p_S(y \mid x)$.

We denote the teacher hidden states at layer ℓ and timestep t as $\mathbf{h}_{T,t}^{(\ell)}$, and the corresponding student representations as $\mathbf{h}_{S,t}^{(\ell)}$. Attention matrices are denoted by $\mathbf{A}_{T,t}^{(\ell)}$ and $\mathbf{A}_{S,t}^{(\ell)}$.

A.2 Confidence Estimation via Predictive Entropy

GateKD estimates teacher reliability using predictive entropy. For each decoding step t, the teacher confidence score is defined as

$$c_t = 1 - \frac{H(p_T(\cdot \mid y_{<t}, x))}{\log |\mathcal{V}|},$$

where $H(\cdot)$ denotes Shannon entropy and $\mathcal{V}$ is the vocabulary. This normalization ensures $c_t \in [0, 1]$, with higher values indicating more confident predictions.

We aggregate token-level confidence into a sequence-level score

$$c(x) = \frac{1}{L} \sum_{t=1}^{L} c_t.$$

A.3 Confidence-Gated Soft Label Distillation

Standard knowledge distillation minimizes the KL divergence between teacher and student output distributions. GateKD modulates this objective by confidence-aware gating:

$$\mathcal{L}_{\text{soft}} = c(x) \cdot \mathrm{KL}(p_T(\cdot \mid x) \parallel p_S(\cdot \mid x)).$$

This formulation suppresses gradients induced by low-confidence teacher predictions, preventing the student from overfitting to speculative or unstable reasoning trajectories.

A.4 Gated Hidden-State Alignment

To encourage the transfer of intermediate reasoning representations, GateKD aligns teacher and student hidden states only when the teacher is confident:

$$\mathcal{L}_{\text{hid}} = c(x) \sum_{\ell,t} \left\| \mathbf{h}_{S,t}^{(\ell)} - \mathbf{h}_{T,t}^{(\ell)} \right\|_2^2.$$

This gating prevents the propagation of noisy intermediate states that arise from hallucinated or inconsistent reasoning.

A.5 Reliability-Filtered Attention Distillation

Attention patterns often encode structural reasoning information but can be highly unstable under uncertainty. GateKD therefore applies confidence-filtered attention alignment:

$$\mathcal{L}_{\text{attn}} = c(x) \sum_{\ell,t} \mathrm{KL}\Big(\mathbf{A}_{T,t}^{(\ell)} \parallel \mathbf{A}_{S,t}^{(\ell)} \Big).$$

A.6 Overall Objective

The final GateKD training objective is

$$\mathcal{L}_{\text{GateKD}} = \mathcal{L}_{\text{hard}} + \lambda_1 \mathcal{L}_{\text{soft}} + \lambda_2 \mathcal{L}_{\text{hid}} + \lambda_3 \mathcal{L}_{\text{attn}},$$

where $\mathcal{L}_{\text{hard}}$ denotes standard supervised loss with ground-truth labels, and $\lambda_1, \lambda_2, \lambda_3$ control the contribution of each gated component.

A.7 Why Confidence-Gated Distillation Works

From an optimization perspective, reasoning distillation can be viewed as minimizing the expected discrepancy between student and teacher trajectories under teacher-induced supervision noise. When teacher confidence is low, the variance of gradient estimates increases, leading to unstable student updates. GateKD reduces this variance by down-weighting uncertain supervision, effectively performing reliability-aware risk minimization.

Intuitively, GateKD biases learning toward regions of the input space where the teacher exhibits consistent and stable reasoning behavior. This mirrors expert learning paradigms, where uncertain demonstrations are discounted rather than blindly imitated. As a result, GateKD enables students to acquire robust reasoning patterns while avoiding overconfident hallucinations, which is particularly important for small models with limited corrective capacity.

B Implementation and Training Details

B.1 Model Architectures

We conduct experiments using encoder–decoder architectures from the T5 and Flan-T5 families, including `small`, `base`, and `large` variants. Teacher models are initialized from publicly available pretrained checkpoints and remain frozen during distillation. Student models are trained from pretrained weights and updated using the proposed GateKD objective.

B.2 Training Configuration

All models are trained using the AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and weight decay of 1×10^{-2}. We use a linear learning rate scheduler with warm-up over the first 5% of training steps. The base learning rate is set to 3×10^{-4} for student models across all experiments.

Training is performed with a batch size of 128, implemented via gradient accumulation when necessary. It typically requires 3–5 epochs for convergence, depending on the dataset and student model size. Early stopping is applied based on validation accuracy.

B.3 GateKD Hyperparameters

GateKD introduces three gated distillation components: soft-label distillation, hidden-state alignment, and attention alignment. The corresponding loss weights are fixed across all experiments:

$$\lambda_1 = 1.0, \quad \lambda_2 = 0.5, \quad \lambda_3 = 0.1.$$

Predictive entropy is used to estimate teacher confidence, and confidence scores are normalized to $[0, 1]$ as described in Appendix **??**. No task-specific tuning of gating thresholds is performed.

B.4 Hardware and Runtime

All experiments are conducted on NVIDIA H200 GPUs with 141 GB of HBM3 memory. Training a `T5-base` student model requires approximately 6 hours on a single H200 GPU, while `T5-small` models converge within 3 hours. Multi-GPU data parallelism is employed for larger models when applicable.

Mixed-precision training with FP16 is used throughout to improve computational efficiency without observable degradation in performance.

B.5 Reproducibility

All reported results are averaged over 5 independent runs with different random seeds. We fix the random seed for model initialization, data shuffling, and dropout layers. Evaluation is performed using exact match accuracy for symbolic and logical reasoning tasks, and standard accuracy metrics for commonsense benchmarks.

Implementation is based on the HuggingFace `transformers` library, and all hyperparameters are kept consistent across baselines to ensure fair comparison.

C Additional Qualitative Analysis

Figures 3–6 provide qualitative evidence for the effectiveness of confidence-gated distillation in GateKD. Across all examples, a consistent pattern emerges: teacher models often produce fluent but brittle reasoning trajectories that contain subtle intermediate errors, which are difficult for small student models to recover from when distilled in an open-loop manner.

In the shuffled object tracking example (Figure 3), the teacher incorrectly updates object ownership during the final swap, despite maintaining locally plausible transitions. This error arises from an unstable intermediate representation rather than from surface-level ambiguity. The mentor, by contrast, maintains a coherent state evolution with lower predictive entropy. GateKD detects this reliability gap and suppresses teacher supervision while amplifying mentor guidance, allowing the student to internalize correct state-tracking dynamics.

A similar phenomenon appears in date understanding (Figure 4), where the teacher makes an early temporal misalignment that propagates through subsequent steps. Although the final reasoning is fluent, it is anchored to an incorrect initial assumption. Because GateKD gates supervision at the intermediate level, high-entropy teacher signals are down-weighted, preventing the student from inheriting this systematic bias. This highlights the importance of closed-loop supervision in temporal reasoning tasks, where early errors are especially destructive.

In arithmetic reasoning on SVAMP (Figure 5), the teacher collapses the problem into a single-step heuristic, yielding an incorrect solution. The mentor instead follows a structured algebraic derivation with consistent intermediate states. GateKD reinforces this low-entropy reasoning trajectory through gated hidden-state and attention alignment, enabling the student to acquire symbolic manipulation skills rather than shallow pattern matching.

Finally, Figure 6 demonstrates that these effects generalize across reasoning domains. Despite substantial differences in task structure, GateKD consistently favors stable, low-entropy supervision and suppresses unreliable reasoning signals. This suggests that the gains of GateKD stem not from task-specific heuristics, but from a general principle: *reasoning distillation should be selective, confidence-aware, and closed-loop*.

Overall, these qualitative results complement our quantitative findings by illustrating how GateKD prevents error amplification, stabilizes intermediate representations, and explainably improves student reasoning fidelity.

D Limitations and Ethical Considerations

D.1 Limitations

While GateKD demonstrates consistent improvements across a range of reasoning tasks and model scales, several limitations remain. First, GateKD relies on the availability of confidence estimates from teacher and mentor models. Although predictive entropy and related uncertainty measures are widely supported in modern language models, their calibration quality may vary across architectures and training regimes, potentially affecting the precision of the gating mechanism.

Second, the closed-loop distillation process introduces additional computational overhead due to iterative teacher evaluation and intermediate-state alignment. While this overhead is modest relative to full-scale pretraining, it may limit applicability in extremely resource-constrained settings or when distilling very large teachers over long reasoning traces.

Third, GateKD assumes that mentor models provide more reliable reasoning signals than the teacher on average. In scenarios where both teacher and mentor exhibit correlated failure modes or systematic biases, confidence gating alone may not fully prevent error propagation. Extending GateKD to incorporate multiple heterogeneous mentors or external verification signals is a promising direction for future work.

Finally, our qualitative analyses focus on representative reasoning benchmarks. Although these benchmarks capture diverse reasoning patterns, they do not exhaustively represent all real-world reasoning behaviors, particularly in open-ended or interactive environments.

D.2 Ethical Considerations

GateKD is designed to improve the reliability of reasoning distillation and reduce the transfer of hallucinated or unstable reasoning patterns. As such, it contributes positively to the development of safer and more dependable small language models. However, several ethical considerations warrant discussion.

First, GateKD does not eliminate biases present in the teacher or mentor models. While confidence-based gating suppresses unstable reasoning, it does not explicitly address social, cultural, or demographic biases encoded in model representations. Careful dataset curation and complementary bias-mitigation techniques remain necessary when deploying distilled models in real-world applications.

Second, the improved plausibility and coherence of distilled reasoning may increase user trust in model outputs. This underscores the importance of clearly communicating model limitations and avoiding deployment in high-stakes decision-making contexts without appropriate human oversight.

Finally, all experiments in this work are conducted on publicly available benchmarks and models, and no personal or sensitive data are used. GateKD does not introduce mechanisms for data memorization or privacy leakage beyond those already present in the underlying models.

Overall, we view GateKD as a step toward more robust and responsible reasoning distillation, while acknowledging that confidence-aware supervision should be combined with broader safety, fairness, and accountability practices in practical deployments.

Tracking Shuffled Objects

Question: Alice, Bob, and Claire are playing a game. At the start of the game, they are each holding a ball: Alice has a orange ball, Bob has a white ball, and Claire has a blue ball. \n\nAs the game progresses, pairs of players trade balls. First, Alice and Bob swap balls. Then, Bob and Claire swap balls. Finally, Alice and Bob swap balls. At the end of the game, Alice has the \nWhich choice is true? Answer choices: (A) orange ball, (B) white ball, (C) blue ball.

Answer: C

Role	Content	Prediction
Teacher	At the start of the game: Alice has an orange ball, Bob has a white ball, and Claire has a blue ball. After the first swap between Alice and Bob: Alice has a white ball, Bob has an orange ball, and Claire still has a blue ball. After the second swap between Bob and Claire: Alice still has a white ball, Bob has a blue ball, and Claire has an orange ball. After the final swap between Alice and Bob: Alice has an orange ball, Bob has a blue ball, and Claire has a white ball. **Therefore, at the end of the game, Alice has the orange ball. So, the correct answer is (A) orange ball.**	✗ A
Mentor	At the start of the game: Alice has an orange ball, Bob has a white ball, and Claire has a blue ball. After the first swap between Alice and Bob: Alice has a white ball, Bob has an orange ball, and Claire still has a blue ball. After the second swap between Bob and Claire: Alice still has a white ball, Bob has a blue ball, and Claire has an orange ball. After the final swap between Alice and Bob: Alice has a blue ball, Bob has a white ball, and Claire has an orange ball. **Therefore, at the end of the game, Alice has the blue ball. So, the correct answer is (C) blue ball.**	✓ C

Figure 3: **Qualitative comparison on shuffled object tracking.** The teacher produces a fluent but incorrect reasoning trace, leading to a wrong final answer. In contrast, the mentor generates a consistent intermediate state transition and arrives at the correct solution. GateKD selectively suppresses unreliable teacher reasoning and reinforces stable mentor supervision via confidence-aware gating.

Date Understanding

Question: 2015 is coming in 36 hours. What is the date a month ago in MM/DD/YYYY? \nWhich choice is true? Answer choices: (A) 11/29/2014, (B) 03/07/2015, (C) 12/05/2014, (D) 12/20/2014, (E) 11/23/2014, (F) 11/28/2014.

Answer: A

Role	Content	Prediction
Teacher	1. Determine the current date: 12/31/2014 2. Subtract one month from the current date: 12/31/2014 - 1 month = 11/30/2014 3. **The correct answer choice is (F) 11/30/2014.**	✗ F
Mentor	First, we need to determine today's date. Since 2015 is coming in 36 hours, today's date is December 30, 2014. Next, we need to subtract one month from December 30, 2014. This would bring us to November 30, 2014. **Therefore, the correct answer is (A) 11/29/2014.**	✓ A

Figure 4: **Error correction on date understanding.** The teacher incorrectly infers the current date and propagates the error through subsequent steps. The mentor exhibits lower predictive entropy and maintains a coherent temporal reasoning chain, resulting in the correct answer. GateKD prioritizes such low-entropy supervision, preventing error amplification during distillation.

SVAMP

Question: Melissa scored a total of 21 points playing some games. Is she scored 7 points in each game. How many games did she play?

Answer: 3.0

Role	Content	Prediction
Teacher	Melissa played 1 game. To see why, divide the total number of points by the number of points she scored in each game: 21 7 = 1 **So Melissa played 1 game.**	✗ 1.0
Mentor	We can start by using algebra to solve the problem. Let's call the number of games Melissa played \"g\". According to the problem, Melissa scored a total of 21 points playing some games, so we can write an equation: 21 = g We can simplify this by dividing both sides by 7: 21 7 = g Now we can solve for g by dividing both sides by 7: 21 - 7 = g This simplifies to: g = 3 **Therefore, Melissa played 3 games.**	✓ 3.0

Figure 5: **Robust arithmetic reasoning on SVAMP.** The teacher prematurely collapses the reasoning process, yielding an incorrect solution. The mentor follows a structured algebraic derivation with consistent intermediate steps. GateKD gates intermediate supervision based on confidence, enabling the student to acquire correct symbolic reasoning patterns.

Last Letter Concatenation

Question: Take the last letters of each words in \"Max Mikey Cynthia Holly\" and concatenate them.

Answer: xyay

Role	Content	Prediction
Teacher	Max (x) Mikey (y) Cynthia (a) Holly (y) Concatenating the last letters of each word, **we get \"xyab\".**	✗ xyab
Mentor	1. Max (x) Mikey (y) Cynthia (a) Holly (y) Concatenating the last letters gives **us \"xyay\".**	✓ xyay

Figure 6: **Illustrative failure case motivating confidence-gated distillation.** The example shows a simple string reasoning task—concatenating the last letters of each word in "Max Mikey Cynthia Holly." Although the teacher explicitly enumerates intermediate steps, it produces an incorrect conclusion ("xyab") due to unstable reasoning. In contrast, the mentor yields a consistent and correct trace ("xyay"). This discrepancy highlights that surface-level step-by-step reasoning does not guarantee correctness. GateKD exploits confidence and stability signals from the mentor to selectively distill reliable intermediate representations, filtering out misleading teacher reasoning.