

GateKD

Confidence-Gated Closed-Loop Distillation for Robust Reasoning

Teerapong Panboonyuen

Chulalongkorn University · MARSAIL · PBYAIL

teerapong.pa@chula.ac.th

kaopanboonyuen.github.io/GateKD



ASSOCIATION FOR
COMPUTATIONAL
LINGUISTICS

Open-Loop Distillation Trusts the Teacher Blindly

Reasoning distillation transfers Chain-of-Thought ability from large teachers to compact students. Existing methods — including mentor-based approaches — assume teacher supervision is uniformly reliable.



Noisy rationales

Hallucinated intermediate reasoning steps are distilled wholesale into the student, regardless of teacher uncertainty.



Static teacher-student interaction

Soft labels, hidden states, and attention maps are aligned uniformly, even when the teacher itself is unstable.

THE CENTRAL QUESTION

Should all teacher reasoning be trusted equally during distillation?

We answer this negatively — and propose GateKD, which treats the teacher as a dynamic gatekeeper, not a static oracle.

From Sourcing Better Supervision to Trusting It Selectively

GateKD complements mentor-based distillation by changing not WHERE supervision comes from, but WHEN and HOW it is transferred.

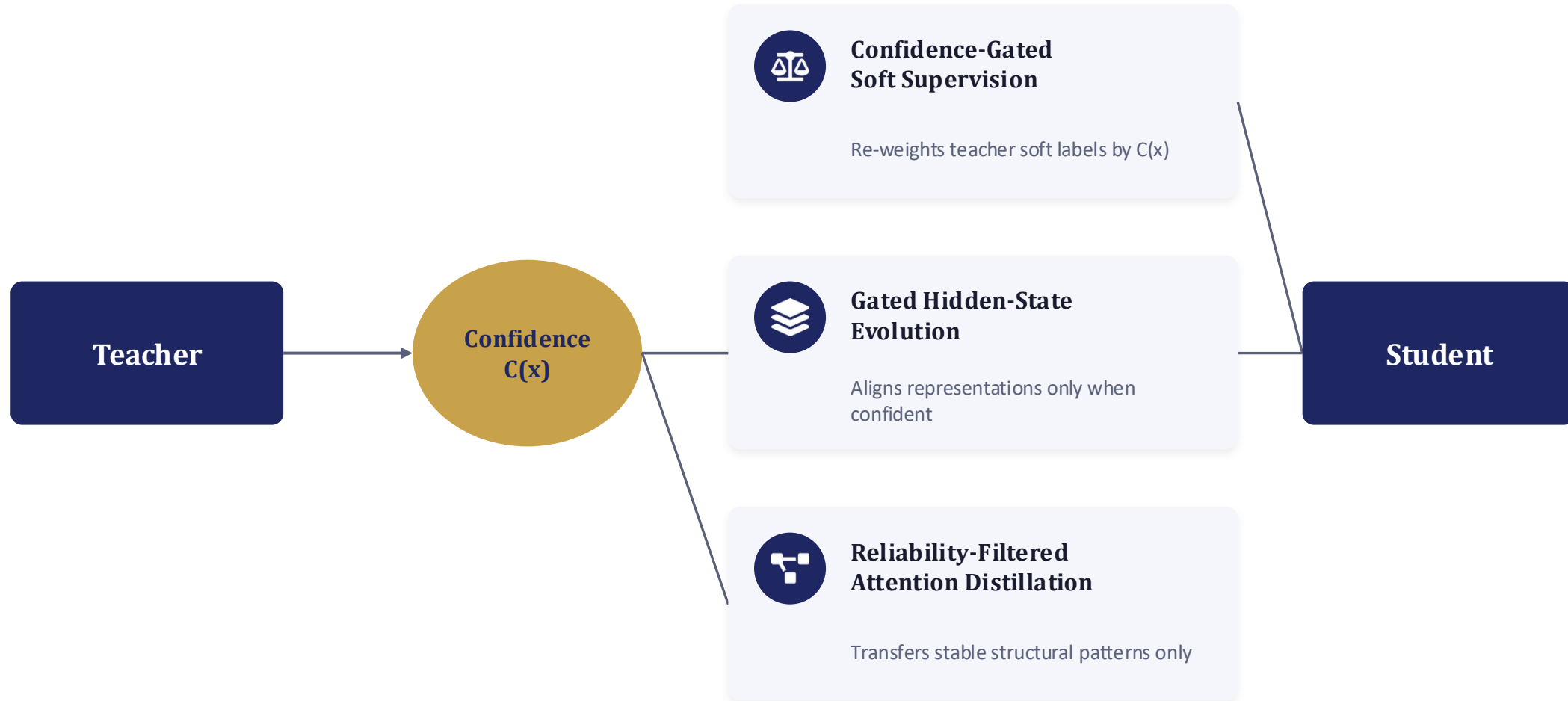
Mentor-KD (open-loop)

- Introduces a task-specific mentor to augment rationales & soft labels
- Improves low-resource reasoning distillation
- Treats all mentor/teacher signals as equally trustworthy
- Static pipeline — no reliability awareness at training time

GateKD (closed-loop)

- Estimates teacher reliability via predictive entropy
- Gates soft labels, hidden states, and attention jointly
- Suppresses low-confidence, hallucination-prone signals
- Reliable supervision dominates as training proceeds

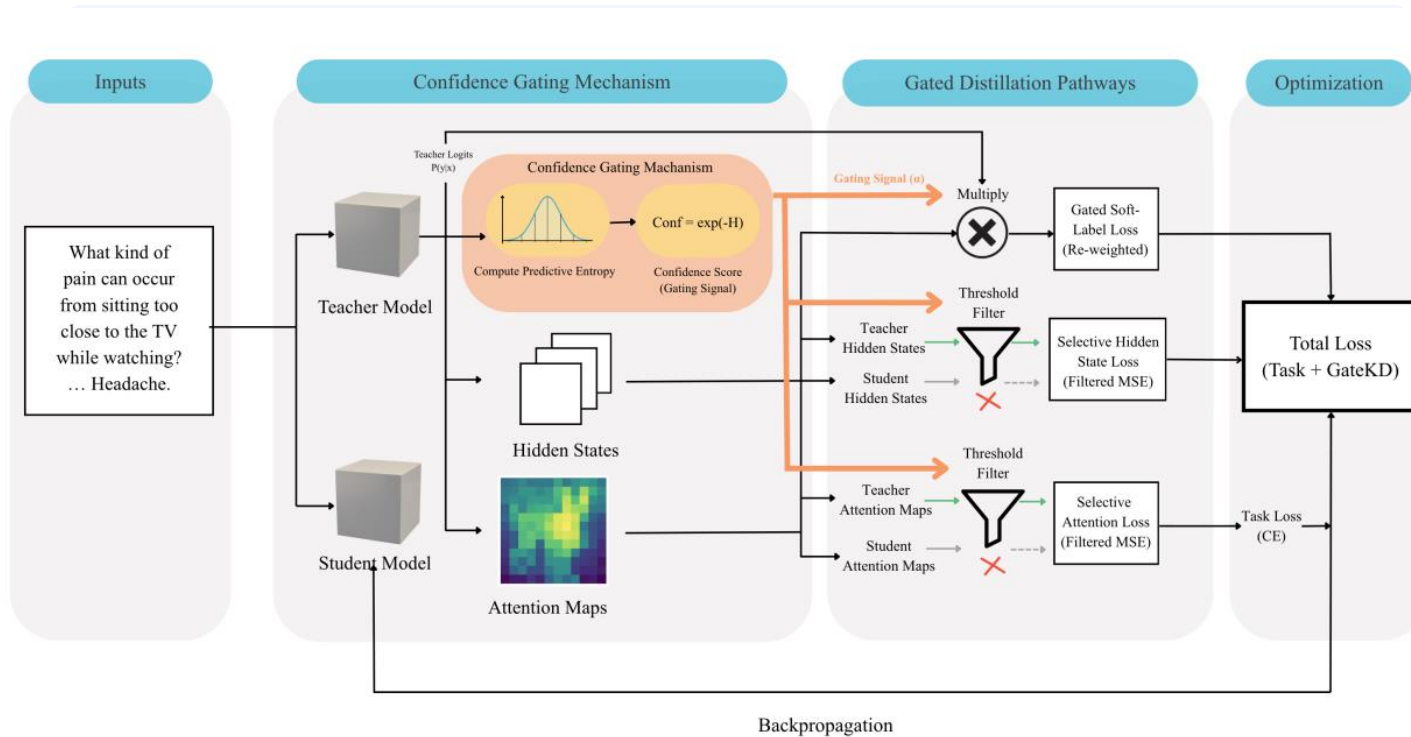
GateKD: One Confidence Signal, Three Gated Pathways



Teacher confidence — estimated via predictive entropy — acts as a unified gating signal across all three distillation pathways, forming an implicit closed feedback loop.

Step 1 — Estimating Teacher Confidence

Teacher reliability is approximated using predictive entropy over the output distribution. Lower entropy implies higher confidence.



INTUITION

Confident teacher

low entropy · signal trusted

Uncertain teacher

high entropy · signal suppressed

GateKD softly suppresses, rather than hard-discards, low-confidence supervision — preserving smooth optimization.

Step 2 — Three Confidence-Gated Distillation Losses



Soft Supervision

$$L = C(x) \cdot \text{CE}(p^T, \log p^S)$$

Reweights soft-label cross-entropy by confidence, softly suppressing unreliable predictions.



Hidden-State Evolution

$$L = \sum_k \mathbb{I}[C_k > C] \parallel h_k^S - \varphi(h^T) \parallel^2$$

Aligns intermediate representations only when batch-relative confidence exceeds the mean.



Attention Distillation

$$L = \sum_k \mathbb{I}[C_k > C] \parallel A_k^S - A_k^T \parallel^2$$

Transfers structural reasoning patterns selectively, suppressing noisy attention maps.

Overall objective: $L = L_{\text{task}} + \lambda_1 L_{\text{gate-soft}} + \lambda_2 L_{\text{gate-hid}} + \lambda_3 L_{\text{gate-att}}$

Experimental Setup



Backbones

T5 & Flan-T5 — small (80M), base (250M), large (780M) — large variants as task-specific teachers



Benchmarks

Commonsense: CSQA, StrategyQA · Logical: Shuffled Objects · Symbolic: Last Letter Concatenation



Baselines

Vanilla-KD, MCC-KD, Mentor-KD, and zero-shot CoT GPT-4o-mini as an upper-bound reference



Protocol

Identical data splits & optimization across methods; results averaged over 5 random seeds

Main Results — GateKD Outperforms Open-Loop Baselines

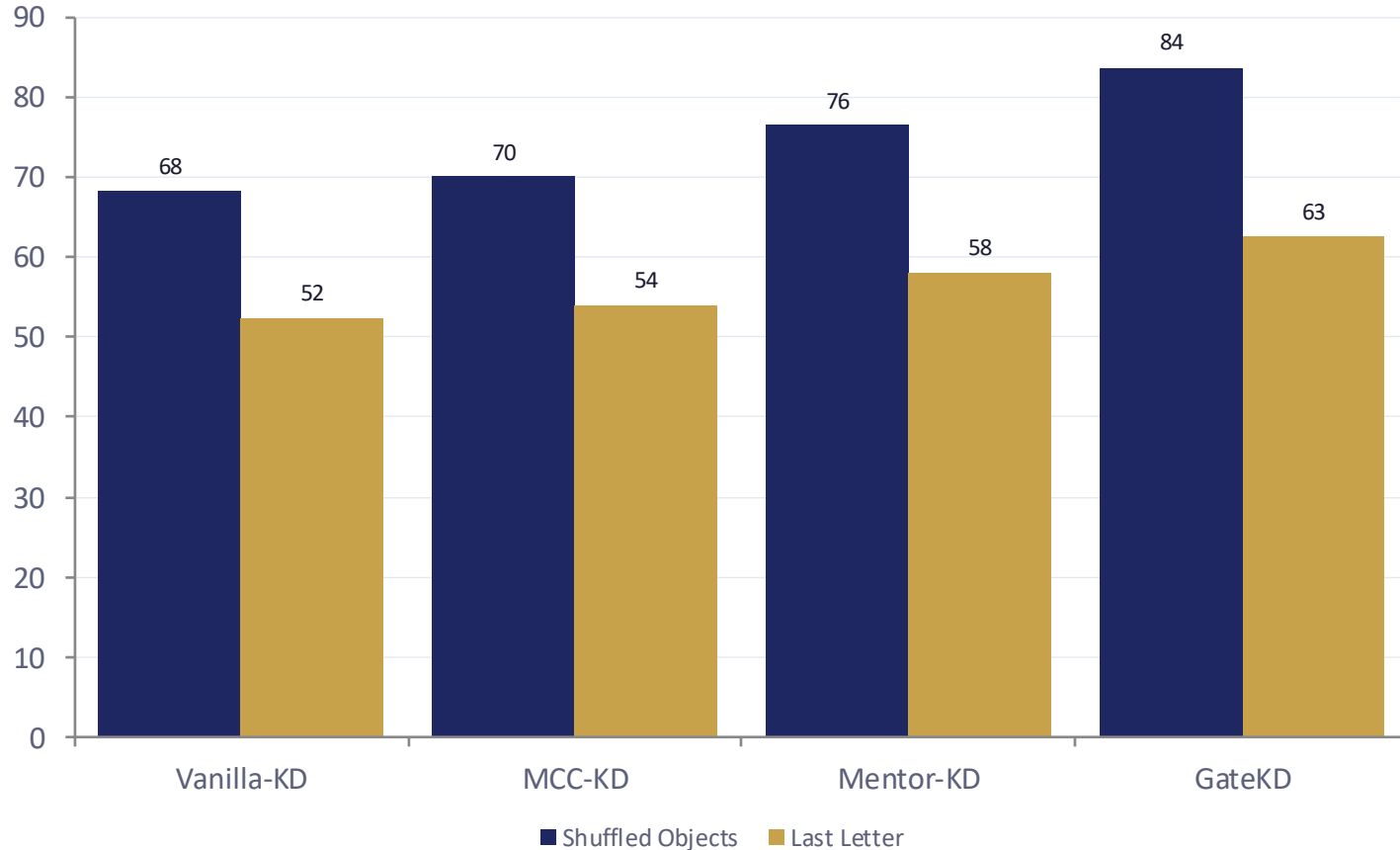
T5 backbones, accuracy (%)

Model	Method	CSQA	SQA	Shuffled	Last Letter
T5-large (780M)	Vanilla-KD	69.3	58.6	88.1	69.0
T5-base (250M)	Vanilla-KD	61.9	55.2	78.4	56.1
T5-base (250M)	Mentor-KD	64.2	57.6	84.9	61.0
T5-base (250M)	GateKD (ours)	66.8	59.9	90.6	65.7
T5-small (80M)	Vanilla-KD	55.4	48.9	63.7	49.6
T5-small (80M)	Mentor-KD	58.6	51.8	72.9	55.2
T5-small (80M)	GateKD (ours)	61.3	54.6	80.8	60.1

GateKD improves over Mentor-KD by up to +4.9 pts (logical) and +4.7 pts (symbolic) on T5-small — gains grow as student capacity shrinks.

Flan-T5 Backbones — Consistent Gains Across Scales

Shuffled-Objects (logical) and Last-Letter (symbolic) accuracy, FlanT5-small



+3.8

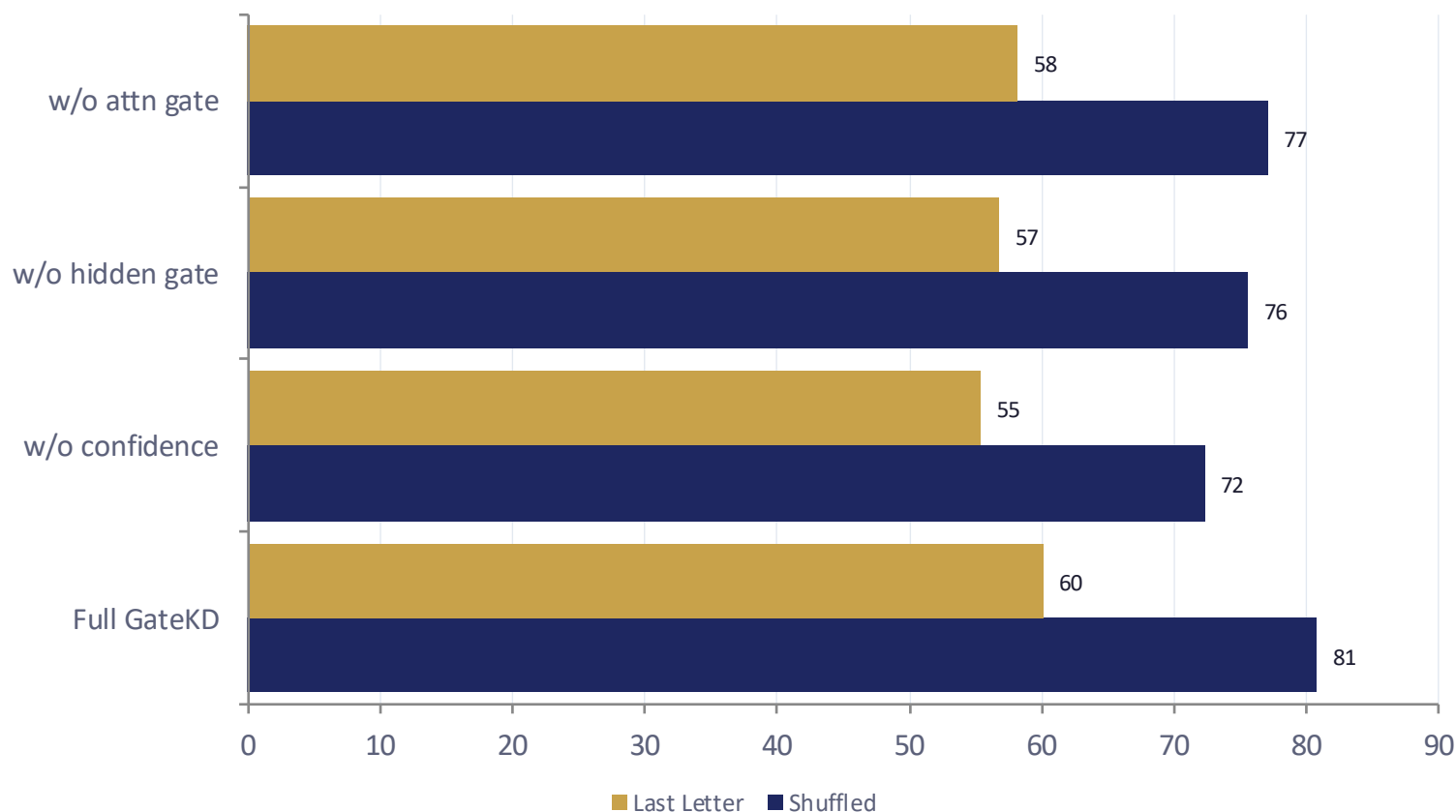
points on Shuffled Objects vs. Mentor-KD (FlanT5-small)

+4.4

points on Last Letter vs. Mentor-KD (FlanT5-small)

Ablation — Every Gating Component Matters

T5-small, logical & symbolic reasoning accuracy (%)



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

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 an 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 \n\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

Qualitative Insight — Suppressing Speculative Reasoning

StrategyQA: “Could a house be built on an asteroid?”



Teacher — low confidence

Generates a fluent, verbose trace exploring speculative construction technologies; concludes the house is “theoretically possible.”

Prediction: “Yes” — INCORRECT

Conflates theoretical possibility with practical feasibility — a hallmark of overconfident speculative reasoning.

Mentor → GateKD-distilled student

Emphasizes material limits, gravity, and sustainability; explicitly separates technical feasibility from practical realizability.

Prediction: “No” — CORRECT

GateKD detects the teacher's low confidence and gates supervision toward this constraint-aware, expert-aligned trajectory.



Takeaways



Selective, not uniform, trust

GateKD treats the teacher as a dynamic gatekeeper — confidence determines how much each signal is trusted.



Closed-loop gating across pathways

A single confidence signal jointly modulates soft labels, hidden states, and attention.



Largest gains where it matters most

Consistent improvements over open-loop baselines, with the biggest wins on logical & symbolic reasoning and smaller students.

kaopanboonyuen.github.io/GateKD · teerapong.pa@chula.ac.th

