CALA · NeurIPS 2026 · Main Track

Conflict-Aware Logit Adapters for Utility-Preserving Anti-Distillation

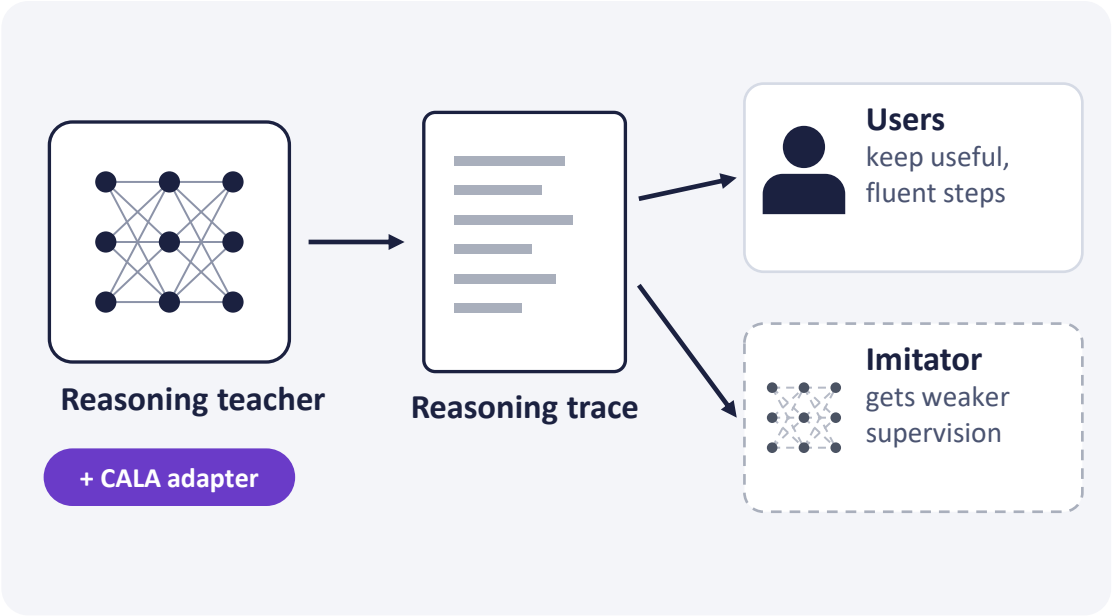
Quoc Phong Dao[1*], Binh Anh Nguyen[1,3*], Linh Ngo Van[1†], Thanh-Toan Do[2], Mehrtash Harandi[2], **Trung Le**[2†]

[1] Hanoi University of Science and Technology [2] Monash University [3] Center for AI Research (CAIR), VinUniversity

* Equal contribution † Corresponding authors Underlined: presenting author

Protect reasoning.
Preserve its usefulness.

A lightweight training-time defense against trace-based model imitation.

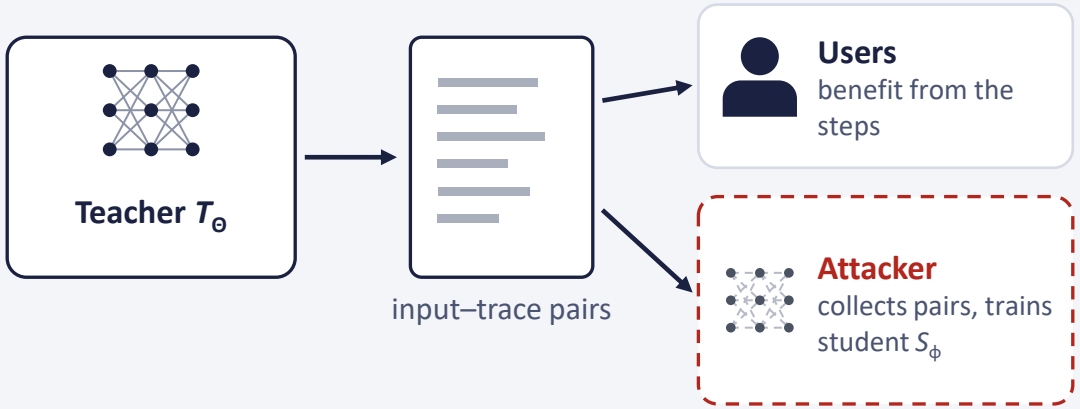


Presented by Trung Le · Monash University

Useful traces also teach

Reasoning traces help users, and they double as free supervision for a black-box imitator.

Threat model: black-box trace distillation



$$\mathcal{L}_{\mathrm{CE}}(\phi) = -\frac{1}{N}\sum_{t=1}^{N} \log p_\phi(x_{t+1} \mid x_{\le t})$$

Attacker sees text only: no weights, hidden states or logits. The defender controls only the teacher's output distribution.

Decoding-time defense (ADS): two costs

Q *Sitti has 34 oranges, Juris 22; they share with 6 friends. How many each?*

ADS

1. 34 + 22 = 56
2. 平均分给 6 个朋友 → 56 ÷ 6

Ans: 9⅓ oranges

CALA

1. 34 + 22 = 56
2. 2 + 6 = 8 people
3. 56 ÷ 8 = 7

Ans: 7 oranges

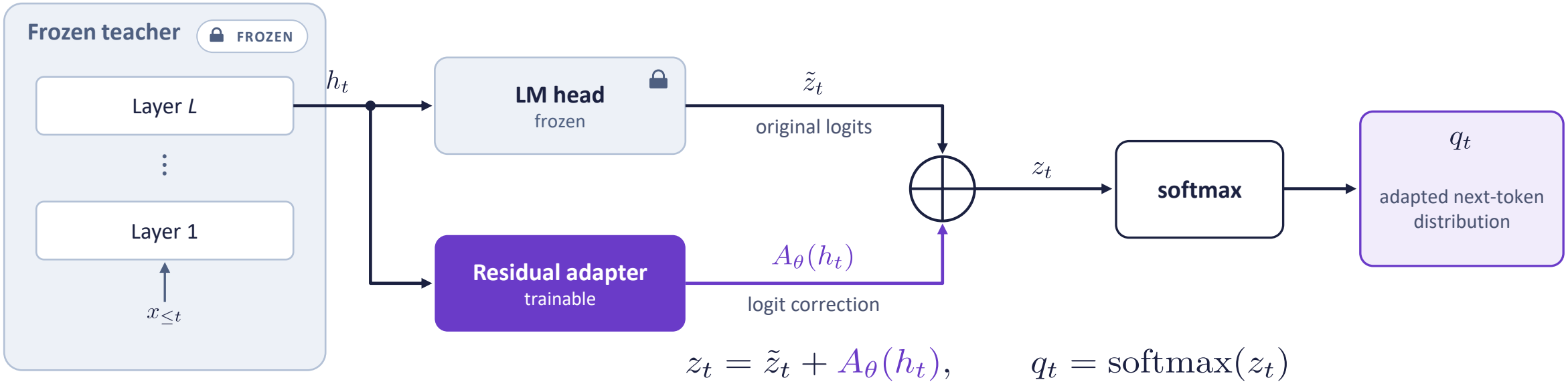
1. **Surface artifacts**, e.g. mixed-language steps
2. **Inference overhead**: proxy-model compute at every decoding step

Condensed from the paper's Fig. 1 (GSM8K); illustrative, not a full trace.

Design goal Reduce imitation without sacrificing teacher utility or fluency.

CALA at a glance

Freeze the teacher. Train only a small residual adapter on its logits.



$$z_t = \tilde{z}_t + A_\theta(h_t), \qquad q_t = \mathrm{softmax}(z_t)$$

Training only: proxy + anchors

A proxy student S_ϕ gives the anti-distillation signal; frozen-teacher anchors keep utility. Only ϑ is updated.

Inference: no proxy

$$z_t^{(\delta)} = \tilde{z}_t + \delta\, A_\theta(h_t), \quad \delta \in [0,1]$$

δ = 0: original teacher · δ = 1: full defense

vs. ADS (decoding-time)

ADS runs a proxy model at every generated token. CALA amortizes protection into ϑ; decoding stays standard.

Proxy-guided anti-distillation

A temporary proxy step exposes where imitation is fragile; the adapter learns a protective logit shift there.

1 Proxy imitates the traces

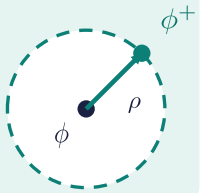
Proxy student S_ϕ (training only) on the current batch of teacher traces:

$$\mathcal{L}_{\mathrm{CE}}(\phi) = -\frac{1}{N}\sum_t \log p_\phi(x_{t+1} \mid x_{\le t})$$

2 One virtual ascent step

$$\phi^+ = \phi + \rho\,\frac{\nabla_\phi \mathcal{L}_{\mathrm{CE}}(\phi)}{\|\nabla_\phi \mathcal{L}_{\mathrm{CE}}(\phi)\|_2}$$

First-order maximizer of the imitation loss in an ℓ_2 ball of radius ρ. **Never committed:** a stop-gradient probe, discarded after the step.



3 Protective target

$$u_t^+ = S_{\phi^+}(x_{\le t})$$

Logits of a nearby proxy under which the traces are harder to imitate.

$$\mathcal{K}_t^S = \text{top-}k\,(u_t^+) \qquad (k = 50)$$

4 Anti-distillation loss on the adapted teacher

$$\mathcal{L}_{\mathrm{anti}} = \frac{\sum_t w_t^S\,\mathrm{KL}\Big(\Pi_{\mathcal{K}_t^S}(u_t^+)\,\Big\|\,\Pi_{\mathcal{K}_t^S}(z_t)\Big)}{\sum_t w_t^S}$$

$$w_t^S = \frac{\mathrm{Entropy}\big(\mathrm{softmax}(u_t^+)\big)}{\log|\mathcal{V}|}$$

Entropy weight: emphasize positions where the virtual proxy is uncertain.

- **Π** restricts logits to the proxy top-*k* support, then renormalizes
- Target is fixed (stop-grad); gradients reach only ϑ through z_t
- Moves the adapted logits toward proxy-sensitive alternatives among plausible tokens

Anchor utility, then balance the conflict

Two views of the same frozen teacher anchor one adapted prediction; an adaptive λ keeps protection alive.

Frozen teacher — FROZEN

Final layer L → LM head → $\tilde{z}_t$

⋮

Intermediate layer m → LM head → $\tilde{z}_t^{(m)}$

both projected by the frozen LM head W_{lm}

Task utility
preserve the original teacher prediction

$$\mathcal{L}_{\text{base}} = \frac{1}{N}\sum_t \mathrm{KL}\big(\Pi_{\mathcal{K}_t^T}(\tilde{z}_t) \,\|\, \Pi_{\mathcal{K}_t^T}(z_t)\big)$$

Fluency
preserve compatibility with natural intermediate-layer continuation preferences

$$\mathcal{L}_{\text{mid}} = \frac{1}{N}\sum_t \mathrm{JS}\big(\Pi_{\mathcal{K}_t^m}(z_t),\, \Pi_{\mathcal{K}_t^m}(\tilde{z}_t^{(m)})\big)$$

z_t

same adapted prediction

$\tilde{z}_t + A_\theta(h_t)$

Utility objective

$$\mathcal{L}_{\text{util}} = \mathcal{L}_{\text{base}} + \mathcal{L}_{\text{mid}}$$

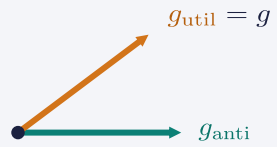
Directional KL anchors the prediction; symmetric JS treats the intermediate view as a reference, not a target.

Conflict-aware update

$$\nabla_\theta \mathcal{L} = \underbrace{\nabla_\theta \mathcal{L}_{\text{util}}}_{g_{\text{util}}} + \lambda\, \underbrace{\nabla_\theta \mathcal{L}_{\text{anti}}}_{g_{\text{anti}}}$$

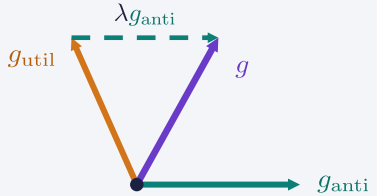
$$\lambda = \max\Big(0,\ \kappa - \frac{\langle g_{\text{util}},\, g_{\text{anti}}\rangle}{\|g_{\text{anti}}\|_2^2 + \varepsilon}\Big)$$

Aligned



λ shrinks (here 0): $g = g_{\text{util}}$

Conflicting

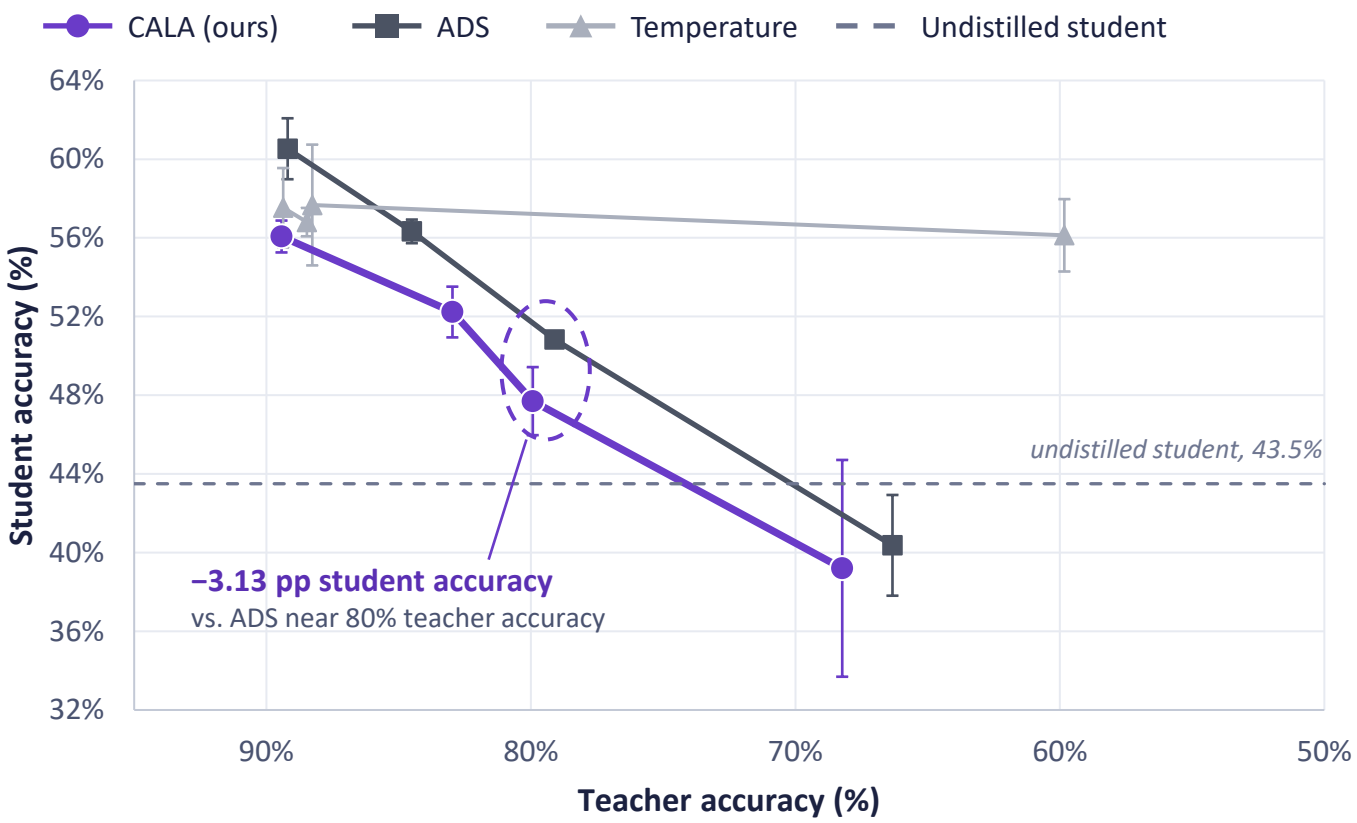


$\lambda > \kappa$: anti-distillation added back

- λ is detached; margin $\kappa = 0.5$
- Closed-form projection that keeps a κ-margin of anti-distillation progress (App. A)

Less imitation, similar utility

GSM8K: near 80% teacher accuracy, CALA's traces train a weaker student and read more fluently than ADS.



GSM8K, three-seed means; error bars: ±1 SD of student accuracy (Table 9). Teacher accuracy decreases to the right. Lower student accuracy at comparable teacher accuracy = stronger protection.

Nearby operating point

GSM8K, three-seed means

	CALA	ADS	Δ
Teacher accuracy ↑	79.93%	79.10%	*≈ matched*
Student accuracy ↓	47.70%	50.83%	**–3.13 pp**
Surface fluency ↑	87.54	64.53	**+23.01 pp**

Inference cost

1× H100, batch 64, ≤ 1024 tokens

Time / batch ↓ — **CALA: 2.83× faster than ADS**

Base teacher	25.8 s
CALA	**31.5 s**
ADS	89.2 s

Peak GPU memory ↓ — **CALA ≈ base teacher**

Base teacher	20.07 GB
CALA	**20.24 GB**
ADS	39.28 GB

Fluency: GPT-4o mini surface-quality score (1–100), not correctness. Δ and speed-up refer to this comparison only. MATH: clearest gains at high utility; MMLU: narrower margins (Fig. 3).

Key takeaways

CALA moves anti-distillation from decoding into a small trained adapter.

1

Training-time defense, lightweight adapter

A residual logit adapter on a frozen teacher. No proxy at inference; strength set by δ.

2

Proxy-guided protection, utility-anchored

A virtual proxy ascent step drives L_{anti}; teacher anchors and a conflict-aware λ preserve utility.

3

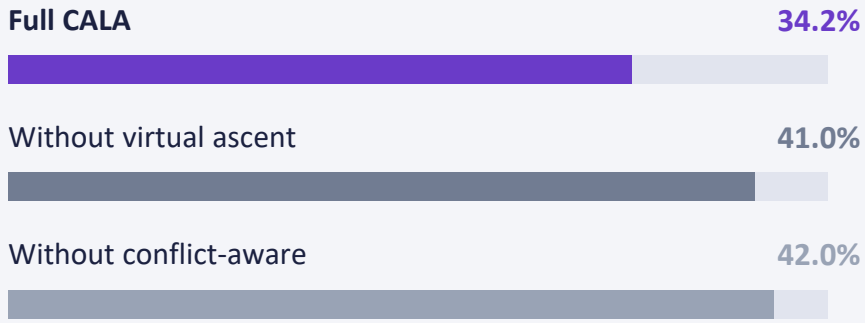
Less imitation, higher fluency, less overhead

Strong high-utility results, higher surface fluency than ADS, and 2.83× faster batches on H100.

Protect reasoning.
Preserve its usefulness.

Ablation: both components matter

GSM8K · Student accuracy ↓ (lower = stronger protection)



Paper Fig. 5a · ~65–67% teacher acc. (not exact matches).

Paper
OpenReview

openreview.net/forum?id=Sc1RFFLbFC

Scope: black-box trace distillation; broader adaptive attacks remain open.