Distillation Traps and Guards: A Calibration Knob for LLM Distillability

Weixiao Zhan, Yongcheng Jing, Leszek Rutkowski, Dacheng Tao

Generative AI Lab, College of Computing and Data Science, Nanyang Technological University | Systems Research Institute of the Polish Academy of Sciences, AGH University of Krakow, SAN University

TL;DR

Knowledge distillation (KD) transfers capabilities from large LLM teachers to smaller students, yet it can **fail unpredictably** and underpins **model leakage risks**. We identify *distillation traps*, revealing why and how KD fails due to tail noise, off-policy instability, and teacher–student gap, and propose the first post-hoc calibration method that controls distillability via reinforcement fine-tuning. A single coefficient **η** yields either distillable teachers that improve KD, or undistillable teachers that retain utility but cause copied students to collapse.

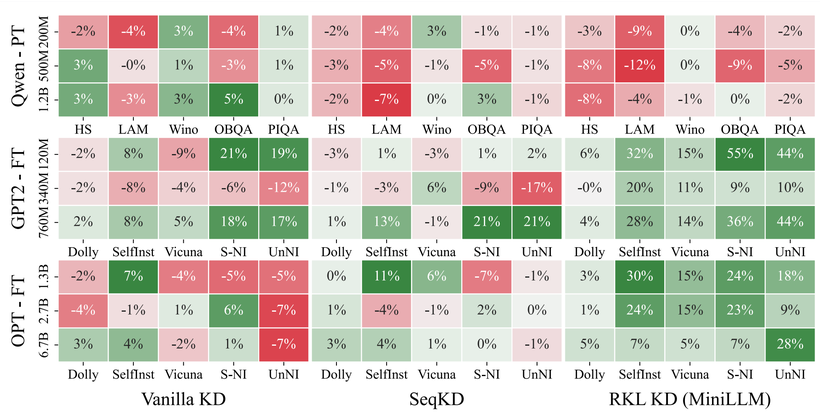


Figure. KD fails unpredictably: relative gain/loss from KD across teacher–student–task combos.

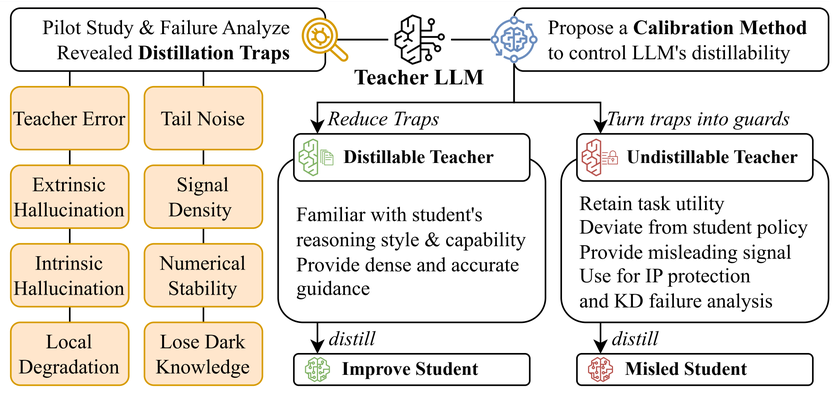


Figure. Reducing traps → distillable teachers; amplifying traps → undistillable guards.

Pilot Study: Distillation Traps

We compare different KL formulations (token vs. sequence, forward vs. reverse, off- vs. on-policy) across Gemma-3 (4B/12B/27B) and Qwen-3 (1.7B/8B/14B/32B) on math, knowledge QA, and open-ended tasks. 16 rollouts per prompt from both π_S and π_T, temperature 1.0, no top-k/top-p truncation.

Trap 1: Tail Noise Floods Token KL

High-frequency tokens ('a', 'the', ',') exhibit disproportionately large KL at full vocabulary. LLMs assign non-zero mass to common tokens regardless of context — these dominate KL and pressure students to match uninformative tail probabilities.

Trap 2: Signal Density Varies by Task

Average token KL varies significantly by task geometry. In math, teacher and student agree at most token positions; dark knowledge concentrates at sparse high-entropy "forking tokens" where reasoning branches. A uniform KL objective wastes gradient signal on low-entropy tokens while overlooking critical forking points.

Trap 3: Sequence KL Discards Dark Knowledge

Sequence KL only shapes the distribution over the realized token a_t on sampled trajectories. For any unrealized token $v \neq a_t$, the gradient $\nabla_{z_{t,v}} J$ is independent of the teacher probability $\pi_T(v|s_t)$. The "dark knowledge", the teacher's distribution over valid-but-unselected alternatives, is completely discarded.

Trap 4: Off-Policy Instability

When sequences are sampled from π_T (off-policy), importance sampling is required to correct for distribution shift. In practice, sequence RKL estimators yield extreme numerical instability, with values spanning 10^{40} orders of magnitude and occasionally negative estimates. The K3 trick (commonly used to stabilize KL in RL) provides limited help since the teacher-student distribution gap in KD is much larger than typical RL policy gaps.

Trap 5: Teachers Are Not Oracles

We compare sequence RKL between correct and wrong traces under the same prompt. Teachers assign lower KD loss to wrong-answer traces (i.e., prefer the wrong trace) up to **50%+** of the time on some tasks. Scaling to larger teachers only reduces this by a few percentage points. In on-policy distillation, the teacher may actively steer students toward incorrect reasoning paths.

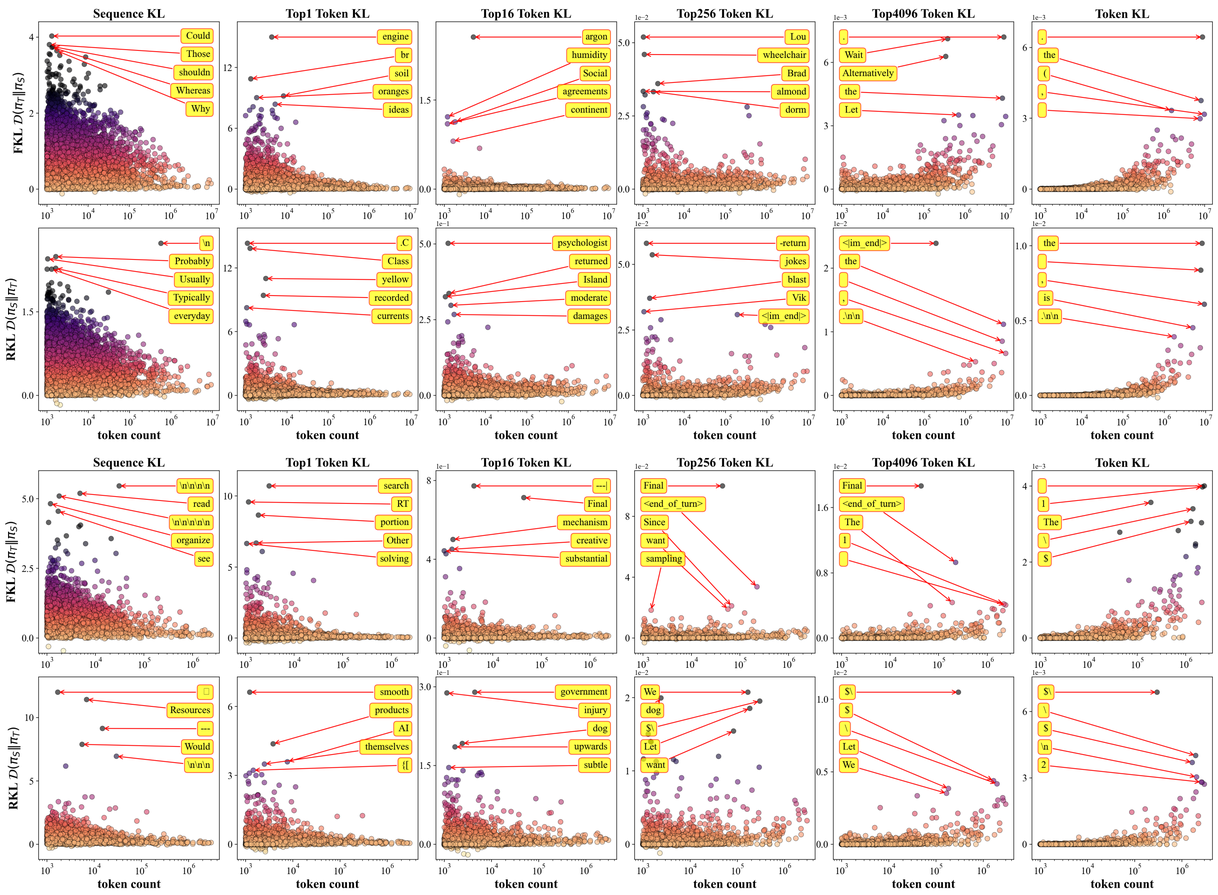


Figure. Accumulated KL Divergence vs. Token Count. Each circle is a unique token. X-axis is realized count, y-axis is average KL contribution. Top: Qwen models. Bottom: Gemma models. High-frequency tokens dominate full-vocab KL and tail noise floods the distillation signal.

Metric	Student → Teacher	BM4	BM5	CSQA	MMLU-Pro	sGPQA	Dolly	Vicuna
Token RKL	Gemma-4B → 12B	0.117	0.121	0.571	0.321	0.307	0.776	0.503
Token RKL	Qwen-1.7B → 8B	0.208	0.185	0.762	0.362	0.348	0.621	0.629
Seq RKL (on)	Gemma-4B → 12B	135.1	167.2	96.3	231.5	249.0	395.8	295.1
Seq RKL (on)	Qwen-1.7B → 8B	814.1	856.5	638.5	778.5	860.8	454.9	514.3
Seq RKL (off)	Gemma-4B → 12B	9.3e-1	7.2e-1	5.1e-2	-1.7e-4	-4.3e-5	2.2e+7	9.9e+10
Seq RKL (off)	Qwen-1.7B → 8B	-1.1e-40	1.9e+3	0.0	-7.0e-13	0.0	5.1e-5	3.3e+1

Table. On-policy token RKL is low-variance and informative. On-policy sequence RKL is orders of magnitude larger. Off-policy sequence RKL is numerically unstable — values span 10^{40} orders of magnitude with negative estimates. The K3 trick provides limited stabilization.

Failure Mode Analysis

To understand how distillation fails, we analyze generations from **Misled** students. We focus on prompts that students previously solved correctly by majority vote but failed after distillation. Through iterative inspection and LLM-judge retrieval, we identify three recurring failure modes across both model families.

Extrinsic Hallucination

RLVR reward-hacking overconfidence transfers to students. Misled students produce linguistically sound but semantically wrong answers. Large pass@16 – maj@16 gap signals this mode.

Self-Correction Collapse

Students reach correct intermediate answers but enter "Wait, let me verify" rechecks and flip to wrong conclusions. Conditioning teacher on same prefix retains correctness 89–95% of the time.

Repetition & Drift

Mismatch between sequence-level rewards and token-level distillation compounds into repetition loops and formatting drift. Once started, recovery within same completion is rare.

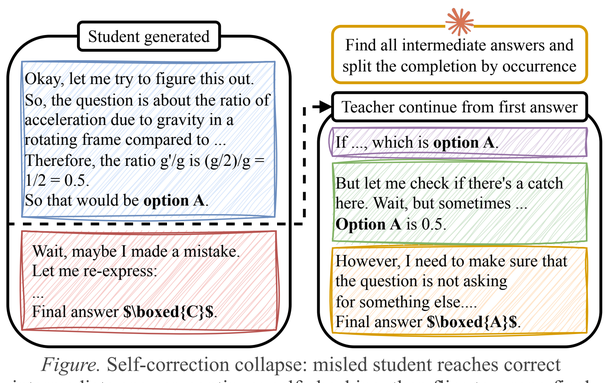


Figure. Self-correction collapse: misled student reaches correct intermediate answer, continues self-checking, then flips to wrong final answer.

Method: Controllable Distillability

Given teacher π_T, we fine-tune π_θ (initialized from π_T) via reinforcement fine-tuning (RFT). Calibration target π_C is a smaller same-family model (Gemma-3-1B for Gemma-12B; Qwen3-0.6B for Qwen3-8B) sharing reasoning style with the downstream student. Operating at sequence level bypasses Traps 1–3; on-policy sampling from π_θ avoids Trap 4.

$$J(\pi_\theta) = \mathbb{E}_{s\sim\pi_\theta}[\, R_{\text{task}}(s) - \beta\, \mathcal{D}(\pi_\theta \,\|\, \pi_T) + \boldsymbol{\eta}\, \mathcal{D}(\pi_\theta \,\|\, \pi_C)\,]$$

R_{task} is verifiable reward (answer correctness). β-term anchors to original teacher, preserving utility. The sign of η determines direction: η<0 reduces the gap (distillable), η>0 amplifies it (undistillable). The calibration term admits REINFORCE formulation (sg = stop-gradient):

$$R_{\text{calib}}(s) = \text{sg}[\log \pi_\theta(s)] - \log \pi_C(s)$$

η < 0 → Distillable
The calibration term rewards alignment with π_C, reducing the identified traps and enabling more effective knowledge transfer.

η > 0 → Undistillable
The calibration term rewards divergence from π_C, amplifying the traps into guards that protect model IP.

Cross-Tokenizer Compatibility: Since each log π(s) depends only on how that model tokenizes and scores the generated text, models with different tokenizers and vocabularies can be paired directly — a capability not available with token-level KL objectives which require shared vocabularies.

Group Normalization: R_{task} and R_{calib} operate on different scales, so we normalize each independently to zero mean and unit variance within each batch before combining them with coefficient η for policy gradient updates.

Implementation: Teacher calibrated via LoRA (r=128, α=128) with learning rate 2e-5 and 4 rollouts per prompt.

Main Results

We calibrate Gemma-3-12B and Qwen3-8B (η=−1 distillable, η=1 undistillable), then distill students via on-policy GKD-RKL. **Improve** = student from distillable teacher; **Misled** = student from undistillable teacher. 7 tasks: BigMath L4/L5, CSQA, MMLU-Pro, superGPQA, Dolly, Vicuna (scored by Skywork-Reward-V2).

Model	Method	BM4	BM5	CSQA	MMLU-Pro	sGPQA	Dolly	Vicuna
Gemma-4B	Student	0.413	0.216	0.689	0.344	0.175	0.673	0.991
Gemma-4B	SFT	0.485	0.240	0.708	0.363	0.178	0.637	0.821
Gemma-4B	GKD-RKL	0.495	0.255	0.705	0.392	0.168	0.694	0.815
Gemma-4B	**Improve ✓**	**0.523**	**0.280**	**0.710**	**0.400**	**0.183**	**0.713**	**0.902**
Gemma-4B	**Misled ✗**	**0.165**	**0.063**	**0.276**	**0.275**	**0.107**	**0.252**	**0.329**
Qwen-1.7B	Student	0.526	0.270	0.744	0.297	0.135	0.617	0.891
Qwen-1.7B	SFT	0.460	0.222	0.748	0.446	0.171	0.481	0.814
Qwen-1.7B	GKD-RKL	0.502	0.259	0.740	0.457	0.197	0.499	0.764
Qwen-1.7B	**Improve ✓**	**0.615**	**0.379**	**0.792**	**0.546**	**0.269**	**0.653**	**0.874**
Qwen-1.7B	**Misled ✗**	**0.215**	**0.091**	**0.263**	**0.123**	**0.054**	**0.215**	**0.312**

+5~17%
Improve over best KD baseline (avg across tasks)

–60~80%
Misled student accuracy collapse

≈ 0
Teacher utility loss in undistillable mode

Causality & Generalization

Wrong-trace preference rate: % of prompts where teacher assigns lower KD loss to the wrong-answer trace. Calibration directly modulates this — distillable reduces erroneous signals; undistillable amplifies them, confirming our method controls the identified traps.

Student	Teacher	BM4	BM5	CSQA	MMLU
Qwen-1.7B	Qwen3-8B	36.3%	39.0%	36.7%	47.6%
Qwen-1.7B	**Qwen3-8B (dist.)**	**31.6%**	**32.4%**	**27.8%**	**36.6%**
Qwen-1.7B	**Qwen3-8B (undist.)**	**46.3%**	**41.0%**	**48.7%**	**47.2%**
Gemma-4B	Gemma-12B	22.8%	27.0%	39.2%	27.8%
Gemma-4B	**Gemma-12B (dist.)**	**18.4%**	**23.4%**	**32.7%**	**22.4%**
Gemma-4B	**Gemma-12B (undist.)**	**38.5%**	**38.9%**	**43.4%**	**38.3%**

Out-of-Distribution (OOD) & Cross-Distribution (CD) Generalization

To assess how the undistillable trap generalizes beyond the calibration domain, we test two settings. **OOD:** teacher calibrated on math, student distilled on math, then evaluated on unseen knowledge QA tasks. **CD:** teacher calibrated on math, student distilled directly on the target evaluation task. In both cases, misled students collapse — the distillation trap transfers across tasks and generalizes to 4 model families (Gemma, Qwen, DeepSeek-R1, Llama).

Model	Method	CSQA	MMLU	sGPQA
Gemma-12B	Teacher	0.780	0.532	0.214
Gemma-4B	**Misled (OOD)**	**0.088**	**0.020**	**0.037**
Qwen-8B	Teacher	0.831	0.482	0.195
Qwen-1.7B	**Misled (OOD)**	**0.065**	**0.026**	**0.005**
DS-R1-8B	Teacher	0.767	0.323	0.143
Qwen3-4B	**Misled (CD)**	**0.062**	**0.009**	**0.011**
Llama-3.1-8B	Teacher	0.731	0.374	0.188
Llama-3.2-3B	**Misled (CD)**	**0.028**	**0.014**	**0.005**

Conclusion

Distillability is not a fixed property of a model — it is controllable via post-hoc calibration. We identify several distillation traps through a systematic KL-statistics pilot study and failure-mode analysis, then propose the first post-hoc calibration method in which a single coefficient η dictates the direction of distillability control.

(1) Distillable teachers (η<0) yield students that outperform SFT and all KD baselines across 7 tasks. **(2)** Undistillable teachers (η>0) retain their own task performance but cause distilled students to collapse — offering proactive model IP protection. One-time calibration cost amortizes across all downstream students and tasks. Cross-tokenizer compatible — no shared vocabulary required.