Rethinking the Role of Temperature in Large Language Model Distillation

Hoang-Chau Luong and Lingwei Chen

Golisano College of Computing and Information Sciences, Rochester Institute of Technology, Rochester, NY, USA

cl6300@rit.edu · lwcics@rit.edu

Code on GitHub

Background & Motivation

- **KD temperature** τ softens the teacher distribution and exposes "dark knowledge" in non-target tokens (Hinton et al., 2015).
- Yet LLM distillation **fixes τ = 1** (MiniLLM, GKD, DistiLLM, ABKD, …), removing temperature from the design space.
- **Reverse KL (RKL)** is widely reported to beat **forward KL (FKL)** — but only under τ = 1.
- FKL weights its loss by teacher probabilities: a sharp, low-entropy teacher hides exactly the non-target mass FKL needs.

Is RKL intrinsically superior for LLM distillation, or has FKL been systematically underestimated?

Three research questions

1. **How does temperature distinguish FKL from RKL?**
   → τ changes what FKL learns, but mainly how RKL learns.
2. **When does FKL outperform RKL?**
   → At τ = 1 RKL wins; at higher τ FKL consistently wins.
3. **Does τ broadly improve KL-based objectives?**
   → Yes — simple KL losses + τ rival recent SOTA methods.

RQ1 · How Does τ Distinguish FKL from RKL?

Setup. Teacher / student logits z^t, z^s over a vocabulary of size V:

$$p^\tau = \mathrm{softmax}(z^t/\tau), \quad q^\tau = \mathrm{softmax}(z^s/\tau)$$

$$\nabla_{z_i^s}\mathcal{L}_{\mathrm{FKL}} = \tfrac{1}{\tau}\left(q_i^\tau - p_i^\tau\right) \qquad \nabla_{z_i^s}\mathcal{L}_{\mathrm{RKL}} = \tfrac{q_i^\tau}{\tau^2}\left(\Delta_i - \mathbb{E}_{j\sim q^\tau}[\Delta_j]\right)$$

with $\Delta_i = z_i^s - z_i^t$ (Prop. 2.2)

High-τ regime (Prop. 2.1)

As τ → ∞, FKL and RKL share the same gradient:

$$\nabla_{z_i^s}\mathcal{L}_{\mathrm{FKL}} = \nabla_{z_i^s}\mathcal{L}_{\mathrm{RKL}} = \tfrac{1}{V\tau^2}\left(z_i^s - z_i^t\right) + \mathcal{O}(\tau^{-3})$$

→ both reduce to logit matching; the difference must come from practical τ.

Low-τ regime (practical)

FKL: τ reshapes the *target* itself, moving mass to non-dominant tokens:

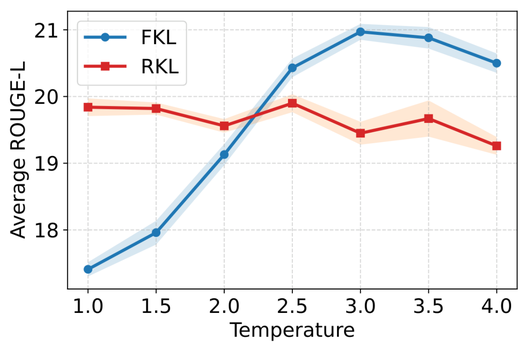
$$\frac{\partial p_i^\tau}{\partial \tau} = \frac{p_i^\tau}{\tau^2}\left(\mathbb{E}_{j\sim p^\tau}[z_j^t] - z_i^t\right)$$

RKL: driven by logit gaps; τ mostly *rescales* the gradient ($1/\tau^2$) and smooths the student-side weights.

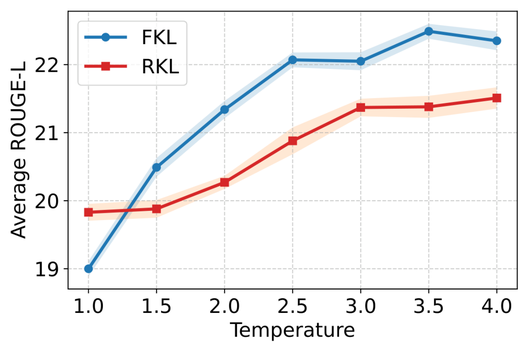
Insight. Temperature changes ***what*** FKL learns, but mainly ***how*** RKL learns: softened targets enrich FKL with non-target signal, while RKL mostly gets gradient rescaling.

RQ2 · When Does FKL Outperform RKL?

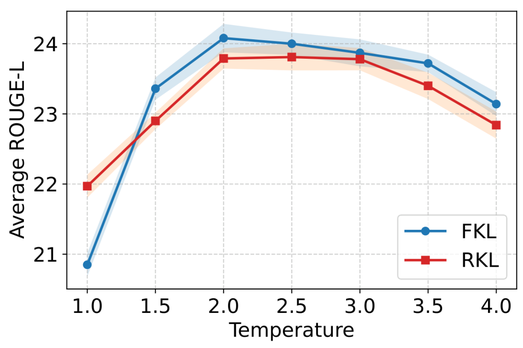
Off-policy distillation, average ROUGE-L over 5 instruction-following benchmarks (Dolly, Self-Instruct, Vicuna, Super-NI, UnNI), 5 seeds.



GPT-2 XL → GPT-2 Base



GPT-2 XL → GPT-2 Medium



OPT 6.7B → OPT 1.3B

On-policy check (GPT-2 Base)

+4.29 **FKL gain** 16.95 → 21.24 (τ = 3.5)

+0.44 **RKL gain** 19.97 → 20.41, then degrades

Insight. At τ = 1 RKL beats FKL, as prior work reports. Once τ softens the teacher, FKL gains **> 3 points** while RKL stays nearly flat — the ranking **reverses** across all model scales (best τ ≈ 2.5–3.5).

RQ3 · Does τ Broadly Improve KL-Based Distillation?

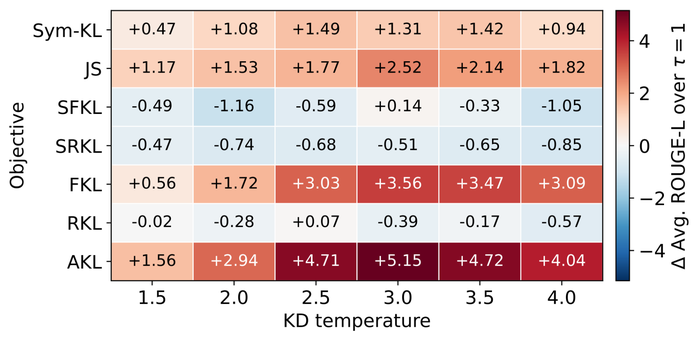
Average ROUGE-L (Dolly, Super-NI, UnNI): default τ = 1 vs. tuned τ

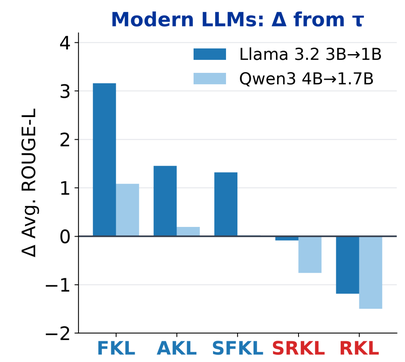
Objective	GPT-2 XL → GPT-2 Base			OPT 6.7B → OPT 1.3B		
	τ = 1	+ τ	Δ	τ = 1	+ τ	Δ
FKL	19.70	25.53	+5.83	25.00	29.08	+4.08
RKL	23.21	23.70	+0.49	25.17	28.63	+3.46
Sym-KL	22.17	24.17	+2.00	24.86	29.23	+4.37
JS	21.97	**25.63**	+3.66	25.67	29.48	+3.81
SFKL	23.31	25.60	+2.29	25.25	28.98	+3.73
SRKL	23.91	24.70	+0.79	24.57	29.24	+4.67
AKL	18.43	25.05	+6.62	24.74	**29.62**	+4.88
AB (ICML'25)		23.91			25.96	
DRKL		25.16			27.71	

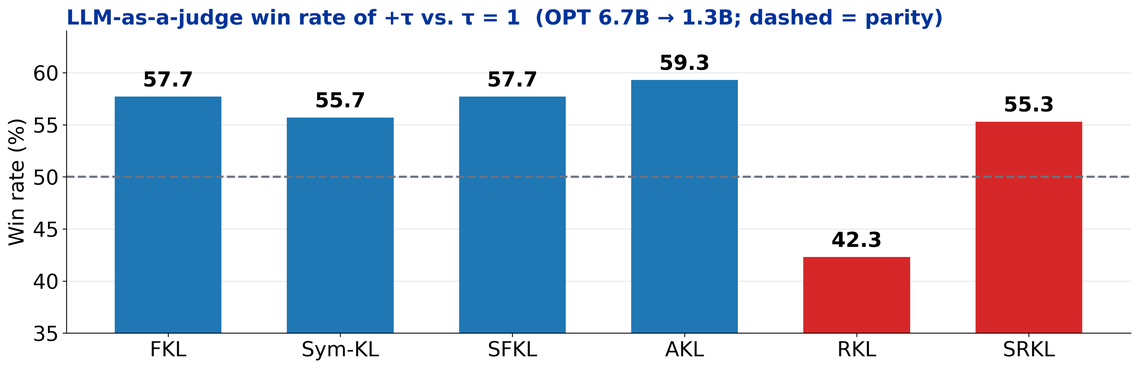
Every objective improves with τ. FKL + τ (25.53 / 29.08) overtakes recent SOTA AB and DRKL; JS + τ and AKL + τ are best overall.

Gains are largest for FKL-based losses (FKL +5.83, AKL +6.62 on GPT-2) vs. RKL +0.49, SRKL +0.79.

Δ Avg. ROUGE-L over τ = 1 across temperatures (GPT-2 Base)







Take-home: KD temperature is not a minor implementation detail. RKL's reported edge over FKL is an artifact of evaluating at τ = 1. Tune τ for every KL-based objective before comparing distillation methods.

Acknowledgment: NSF grant CNS-2549995 · RIT Research Computing

EMNLP2026 BUDAPEST