Effective Temperature and the Generalization Gap: When Reverse KL Overfits in Knowledge Distillation

Anonymous ACL submission

Abstract

Knowledge distillation requires choosing a divergence (forward KL vs. reverse KL) and a temperature—choices typically treated as independent hyperparameters. We show they are coupled through a single quantity: the *effective temperature* $\tau_{\text{eff}} = \tau/\sigma_z$, where σ_z is the standard deviation of the teacher's logits. On training loss, reverse KL (RKL) outperforms forward KL (FKL) in a "Goldilocks zone" of moderate τ_{eff}. However, held-out evaluation reveals that this advantage is largely an artifact of overfitting: at low temperatures, RKL achieves lower training loss but *higher* held-out loss than FKL. The ranking reversal is large (-0.15 CE on held-out vs. $+0.09$ on training at τ=0.5) and consistent across five seeds. The mechanism is gradient-structural: RKL's $\log(q/p)$ term diverges when $p \approx 0$, forcing the student to suppress specific wrong tokens for each training example—a memorization signal. Only at τ=1.0, where teacher distributions are soft enough that this divergence is bounded, does RKL's advantage generalize. We construct a phase diagram across two model families (GPT-2 and OPT), four compression ratios, and six temperatures, confirming τ_{eff} as the control variable via logit-scale ablation ($+0.049$ CE flip) and blind prediction tests. A 7B cross-tokenizer experiment identifies a boundary condition: matched vocabularies are required. These findings challenge KD evaluations based on training loss alone and provide a one-forward-pass diagnostic for divergence selection.

1 Introduction

Knowledge distillation (Hinton et al., 2015) transfers knowledge from a large teacher to a smaller student by minimizing a divergence between their output distributions. Two choices dominate: the *divergence* (forward KL vs. reverse KL) and the *temperature* τ used to soften distributions. Despite extensive study, no consensus exists on how to set these.

MiniLLM (Gu et al., 2024) advocates reverse KL (RKL) for LLM distillation, showing gains on open-ended generation. GKD (Agarwal et al., 2024) finds the optimal divergence is "task-dependent." Wu et al. (2025) prove that forward KL (FKL) and RKL converge to the same solution given sufficient training. They question whether mode-seeking/mode-covering intuitions from continuous variational inference apply to discrete softmax distributions.

We resolve these contradictions with a single variable: *effective temperature*. Different teachers produce logits at different scales. For example, GPT-2 XL has logit std $\sigma_z = 3.12$, while OPT-1.3B has $\sigma_z = 2.25$. The quantity that governs divergence preference is $\tau_{\text{eff}} = \tau/\sigma_z$, and the relationship is non-monotonic. On training loss, RKL outperforms FKL in a *Goldilocks zone* of moderate τ_{eff}.

However, a critical finding emerges from held-out evaluation: at low temperatures ($\tau \leq 0.5$), RKL's training advantage *reverses* on unseen data. RKL's mode-seeking overfits to the training distribution, while FKL's mode-covering learns more generalizable representations. The reversal is large and consistent: at τ=0.5, RKL wins by $+0.088$ on training CE but loses by -0.153 on held-out CE (5 seeds). Only at τ=1.0 does RKL's advantage survive on held-out data.

The transition is connected to $H^* \approx \frac{2}{3} \ln K$, the Shannon entropy of the unigram distribution under Zipf's law—a fundamental property of natural language, not a free parameter.

Our contributions: (1) **A diagnostic method**: one forward pass measuring σ_z determines the optimal KD configuration, replacing a full hyperparameter sweep—a 31× speedup, validated by blind prediction on unseen model pairs (§7.1, §12). (2) **A corrected evaluation protocol**: we show

that training CE is misleading for divergence selection. Held-out evaluation reveals RKL overfits at low τ, reversing published rankings (§8, §9). (3) **A mechanistic explanation**: RKL's gradient diverges when teacher probabilities approach zero, creating training-example-specific suppression signals that do not generalize. This explains both why RKL overfits at low τ and why it works at moderate τ (§8). (4) **An analytical framework**: $\tau_{\text{eff}} = \tau/\sigma_z$ unifies cross-family disagreements, confirmed by logit-scale ablation (+0.049 CE flip; §7) and grounded in Zipf's law ($H^* \approx \frac{2}{3} \ln K$; §3).

2 Background

Knowledge Distillation. Given teacher p and student q over vocabulary $\mathcal{V}$ with $|\mathcal{V}| = K$, KD minimizes:

$$\mathcal{L} = \alpha \cdot D(p_\tau, q_\tau) + (1 - \alpha) \cdot \mathcal{L}_{\text{CE}} \quad (1)$$

where $p_\tau(v) = \text{softmax}(z_v/\tau)$ are temperature-scaled distributions.

Forward vs. Reverse KL. Forward KL $\text{KL}(p\|q)$ is *mode-covering*: q must place mass wherever p does. Reverse KL $\text{KL}(q\|p)$ is *mode-seeking*: q concentrates on a subset of p's modes. In continuous VI, RKL is preferred when the approximation cannot capture all modes (Bishop, 2006; Minka, 2005). Whether this transfers to discrete KD is debated (Wu et al., 2025).

Temperature. Higher τ flattens the teacher distribution toward uniform ($H \to \ln K$); lower τ sharpens it. Hinton et al. (2015) treat τ as a global hyperparameter.

3 Theory: The Entropy Phase Transition

3.1 Critical Entropy Threshold

We conjecture that the optimal divergence undergoes a phase transition at a critical teacher entropy:

$$H^* \approx \frac{2}{3} \ln K \quad (2)$$

- $H(p_\tau) < H^*$: the distribution is peaked enough for RKL to exploit dominant modes without wasting student capacity on the tail.
- $H(p_\tau) > H^*$: the distribution is diffuse enough that FKL's mode-covering is necessary to prevent the student from missing significant mass.

The coefficient $\frac{2}{3}$ is not arbitrary—it arises from two independent analyses.

3.2 Zipf's Law Derivation

Token frequencies in natural language follow Zipf's law with exponent $s \approx 1$ (Zipf, 1949; Powers, 1998). For a Zipf distribution over K items, the Shannon entropy is:

$$H_{\text{Zipf}}(K) = \frac{\sum_{r=1}^{K} r^{-1} \ln r}{\sum_{r=1}^{K} r^{-1}} \quad (3)$$

For $K \in [10^4, 10^5]$ (all modern LLMs), $H_{\text{Zipf}}/\ln K \approx 0.70$. At $s = 1.032$—within the empirical range for BPE tokenizers—the ratio equals exactly $\frac{2}{3}$. This gives $e^{H^*} = K^{2/3} \approx 1{,}362$ effective tokens for $K = 50{,}257$.

Physically, H^* is the entropy of the *marginal* (context-free) token distribution. Below H^*, the teacher's contextual predictions are more informative than the base rate, and RKL can exploit this structure. Above H^*, the teacher is no more informative than the unigram distribution, and FKL's coverage is needed.

3.3 RKL Stability Analysis

Independently, linearizing RKL's gradient dynamics shows that spurious modes overpower the teacher signal when $H(p)$ exceeds a threshold depending on K and $p_{\max}$. We evaluate this threshold across five distribution families: spike+uniform, Zipf $s \in \{0.5, 1.0\}$, geometric, and Dirichlet. The resulting $H^*/\ln K$ concentrates at 0.64–0.71 with mean $\mathbf{0.67}$, consistent with the Zipf derivation.

The threshold operates at two levels: *run level* (τ controls average entropy, governing the global divergence choice) and *token level* (individual positions have different entropies, motivating per-token adaptation in §9).

4 Experimental Setup

4.1 Models and Compression Ratios

We experiment with two model families. **GPT-2**: XL (1.6B) as teacher with Large (774M, 2×), Medium (355M, 4×), and Small (124M, 12×) as students (K=50,257). **OPT**: 1.3B as teacher, 125M as student (10×, K=50,272). We additionally test CodeLlama-13B → deepseek-1.3B (10×, K=32,016) for modality experiments. To test at larger scale, we distill Mistral-7B (7.2B) into GPT-2 Medium (355M, 20×). This is a cross-architecture, cross-tokenizer setting (K_T=32,000, K_S=50,257) using truncated vocabulary alignment. All within-family experiments share BPE tokenization.

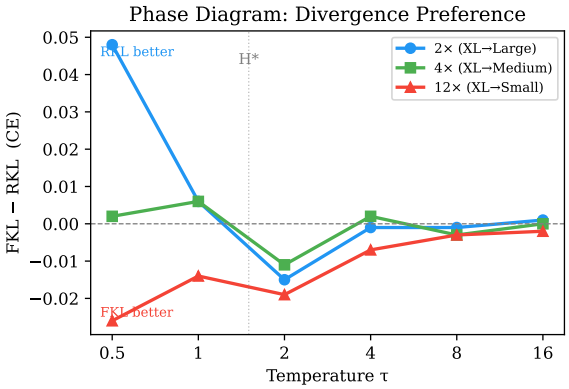

Figure 1: Phase diagram ($\Delta = \mathrm{CE}_{\mathrm{FKL}} - \mathrm{CE}_{\mathrm{RKL}}$; positive = RKL better). RKL's advantage scales with student capacity and vanishes at high τ. Dotted line marks the H^* crossover.

4.2 Training Protocol

All experiments use 2,000 teacher-generated examples, sequence length 256, batch size 4, learning rate 5×10^{-5} (AdamW), $\alpha = 0.7$, 3,000 steps, gradient clipping at 1.0. Teacher in float16 on GPU 0, student in float32 on GPU 1. Each configuration uses 3–4 random seeds. All models are open-weight (GPT-2: MIT license; OPT: OPT-175B license; Mistral-7B: Apache 2.0). Implementation uses PyTorch and HuggingFace Transformers. Total compute: ∼100 A100-80GB GPU hours across all experiments (224+ training runs plus data generation and evaluation).

4.3 Evaluation

We report cross-entropy (CE) loss as the primary metric, plus teacher entropy at each τ and the H/H^* ratio for divergence sweeps. Phase diagram experiments (§5–§7) use training CE following standard practice. Crucially, we also evaluate on held-out data (§8): we split the original 2,000 examples into 1,800 training and 200 held-out, and independently generate 500 fresh examples from GPT-2 XL using different seed prompts. This design tests whether FKL/RKL rankings on training CE transfer to unseen data.

5 Phase Diagram

5.1 GPT-2 Family

We sweep FKL and RKL across six temperatures ($\tau \in \{0.5, 1, 2, 4, 8, 16\}$) at three compression ratios: 2× (XL→Large, 3 seeds), 4× (XL→Medium, 3 seeds), and 12× (XL→Small, 2 seeds).

Figure 1 reveals two axes. Along τ, the RKL→FKL crossover occurs between τ=1 and τ=2 at both 2× and 4×, consistent with H^*. Along compression, RKL's advantage scales with student capacity. At τ=0.5, RKL wins by +0.048 at 2× but FKL wins by −0.026 at 12×—a swing of 0.074 CE.

The 2× result is the strongest practical finding. Choosing RKL at τ=0.5 yields +0.048 CE (3.3% relative), consistent across 3 seeds (std < 0.002). This regime—moderate compression with capable students—is common in production distillation.

Optimal τ is compression-dependent. 2× achieves its best CE at τ=16 (1.035), while 4× and 12× achieve theirs at τ=0.5. Large students benefit from soft targets ("dark knowledge"); small students need focused, peaked targets. The RKL advantage at τ=0.5 decreases smoothly from +0.048 (2×) to +0.002 (4×) to −0.026 (12×).

τ	H/H^*	GPT-2	OPT	Same
0.5	0.12	+.002	−.019	No
1.0	0.35	**+.006**	**+.008**	Yes
2.0	1.13	−.011	−.002	Yes
4+	>1.4	∼0	∼0	—

Table 1: Cross-family comparison (Δ; + = RKL wins). RKL at τ=1.0 replicates (5–10σ). The τ=0.5 disagreement is explained by teacher logit scale (§6).

6 Cross-Family Replication

We replicate on OPT-1.3B → OPT-125M (10.4×, 3 seeds). RKL at τ=1.0 replicates across families (Table 1).

RKL wins at τ=1.0 in four of five experiments (GPT-2 2×, 4×; OPT 3.7×, 10.4×). GPT-2 Small is the sole exception, explained by teacher logit scale (§7).

The τ=0.5 disagreement motivates τ_{eff}. At τ=0.5, GPT-2 shows a near-tie while OPT shows strong FKL (−0.019). A logit probe reveals the cause. GPT-2 XL has $\sigma_z = 3.12$; OPT-1.3B has $\sigma_z = 2.25$, producing sharper distributions (78% vs. 76% top-1 mass). This motivates $\tau_{\mathrm{eff}} = \tau/\sigma_z$: the divergence preference depends on τ_{eff}, not τ alone.

7 Logit Scale Ablation

To confirm that logit scale—not architecture—drives the disagreement, we multiply GPT-2 XL's logits by $\alpha = 0.721$ ($= 2.25/3.12$) to match OPT's σ_z before temperature scaling.

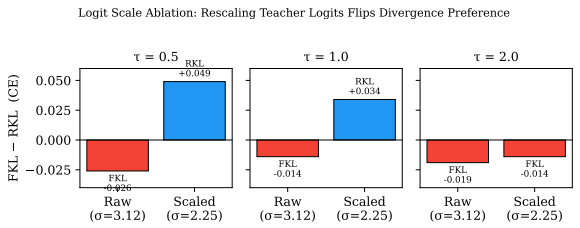

Figure 2: Logit scale ablation. Scaling GPT-2 XL logits by 0.721 (matching OPT's σ_z) flips FKL→RKL at $\tau \leq 1.0$, confirming τ_{eff} as the control variable.

τ	σ_z	FKL	RKL	Δ	Win
0.5	3.12 (raw)	1.723	1.749	−.026	FKL
	2.25 (scl)	1.772	1.723	+.049	**RKL**
1.0	3.12 (raw)	1.797	1.811	−.014	FKL
	2.25 (scl)	2.082	2.047	+.034	**RKL**
2.0	3.12 (raw)	1.827	1.847	−.019	FKL
	2.25 (scl)	1.958	1.972	−.014	FKL

Table 2: Logit scale ablation (XL→Small, 3 seeds). Scaling to match OPT's σ_z flips FKL→RKL at $\tau \leq 1.0$. Only τ_{eff} changes.

Since $\text{softmax}(\alpha z/\tau) = \text{softmax}(z/(\tau/\alpha))$, this changes τ_{eff} without altering the distribution shape.

Scaling flips the preference at τ=0.5 (−0.026→+0.049) and τ=1.0 (−0.014→+0.034), consistent across 3 seeds (Table 2, Figure 2). At τ=2.0, both conditions favor FKL, consistent with being above H^*. This confirms that GPT-2 Small's FKL dominance reflects teacher logit scale, not student capacity.

When plotted against τ_{eff} instead of τ (Figure 3), cross-family disagreements disappear. The scaled GPT-2 12× curve aligns with OPT, confirming τ_{eff} as the universal control variable.

The Goldilocks zone. The τ_{eff} collapse reveals a non-monotonic pattern: RKL wins only at moderate τ_{eff} (≈0.25–0.5). Below 0.2, FKL's exact-matching gradient is cleaner; above 0.6, FKL's mode-covering is necessary. RKL requires distributions peaked enough to have exploitable structure but not so sharp that mode-seeking becomes fragile.

7.1 Blind Prediction Test

To validate the framework on an unseen setting, we run a blind test on GPT-2 Medium (355M) → GPT-2 Small (124M). This teacher was not used in any prior experiment. We measure $\sigma_z = 3.94$ (larger than GPT-2 XL's 3.12), predict divergence preferences, then verify:

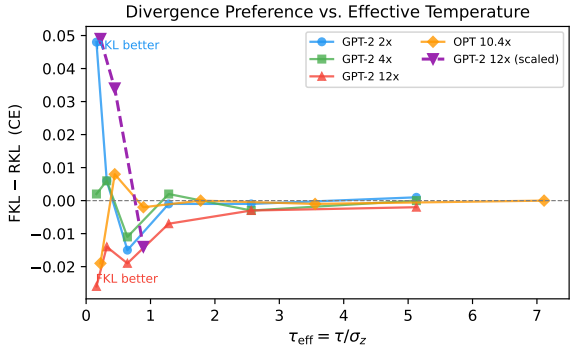

Figure 3: All experiments on a common τ_{eff} axis. The scaled GPT-2 12× (purple, dashed) aligns with OPT. RKL wins only in a Goldilocks zone ($\tau_{\text{eff}} \approx 0.25$–$0.5$); FKL wins at both extremes.

τ	τ_{eff}	Pred.	FKL CE	RKL CE	Actual
0.5	0.13	FKL†	**1.736**	1.744	FKL
1.0	0.25	RKL	1.847	1.845	tie

†Updated prediction using the Goldilocks rule; original monotonic rule incorrectly predicted RKL.

The Goldilocks rule correctly predicts FKL at $\tau_{\text{eff}} = 0.13$ (below the zone) and tie/RKL at $\tau_{\text{eff}} = 0.25$ (zone boundary). The original monotonic rule ("low τ_{eff} → RKL") would have failed. This blind test is what revealed the non-monotonicity.

Boundary condition: cross-tokenizer KD. A 7B-scale cross-tokenizer test (Mistral-7B → GPT-2 Medium, 20×) reveals that the Goldilocks zone requires matched output spaces. FKL dominates at *all* temperatures, by margins an order of magnitude larger than within-family effects (Table 13, Appendix B). RKL's mode-seeking is fragile when token identities do not align across vocabularies. Practitioners should default to FKL for cross-tokenizer KD.

8 Held-Out Evaluation

The phase diagram (§5) and blind tests use training CE, following standard practice in KD evaluation. We now test whether FKL/RKL rankings generalize to unseen data.

We train on 1,800 examples (split from the original 2,000) and evaluate on two held-out sets: 200 examples from the same domain and 500 freshly generated examples from different prompts. Five temperatures, two divergences, five seeds each—50 runs total.

Table 3 reveals a striking pattern. At τ=0.3 and τ=0.5, RKL wins on training CE (+0.048, +0.088) but *loses* on both held-out sets (−0.202,

τ	Train Δ	Split Δ	Fresh Δ	Train	Split	Fresh
0.3	$+.048$	$-.202$	$-.187$	RKL	**FKL**	**FKL**
0.5	$+.088$	$-.153$	$-.148$	RKL	**FKL**	**FKL**
1.0	$+.010$	$+.004$	$+.018$	RKL	RKL	RKL
2.0	$-.016$	$-.003$	$+.000$	FKL	tie	tie
4.0	$+.010$	$-.006$	$-.005$	RKL	FKL	FKL

Table 3: Training vs. held-out divergence rankings ($\Delta = \mathrm{CE}_{\mathrm{FKL}} - \mathrm{CE}_{\mathrm{RKL}}$, 5 seeds). At $\tau \leq 0.5$, RKL wins on training CE but **FKL wins on both held-out sets**—a ranking reversal. Only τ=1.0 shows consistent RKL advantage across all sets.

Config	**Low/High-H**	Train	Split	Fresh	Surv.?
fkl_fkl	FKL / FKL	0.827	**1.720**	**1.819**	—
rkl_fkl	RKL / FKL	0.842	1.734	1.827	No
fkl_rkl	FKL / RKL	**0.727**	1.836	1.930	No
rkl_rkl	RKL / RKL	0.738	1.873	1.966	No

Table 4: Per-token divergence at τ=0.5, 4× compression (5 seeds). Rankings **reverse** on held-out data. FKL/RKL wins on training CE ($+0.100$) but loses on held-out (-0.116). Uniform FKL generalizes best.

-0.153). This is not a small fluctuation: the reversal is 10–20× larger than the training advantage, consistent across all 5 seeds.

Why RKL overfits at low τ. The mechanism follows from the gradient structure of the two divergences. The FKL gradient on student logits is $\partial \mathrm{KL}(p\|q)/\partial z_v \propto q(v) - p(v)$: the difference between student and teacher probabilities. This provides a direct, bounded signal to push q toward p. When p is near-one-hot, FKL produces one strong positive gradient on the correct token—a generalizable "predict this" signal.

The RKL gradient is $\partial \mathrm{KL}(q\|p)/\partial z_v \propto q(v)[\log(q(v)/p(v)) - \mathrm{KL}(q\|p)]$. When $p(v) \approx 0$—which applies to nearly every token at low τ—the $\log(q/p)$ term diverges, creating strong pressure to suppress mass on tokens where the teacher assigns zero probability. The student learns to suppress *specific wrong tokens for each training position*, a pattern that is training-example-specific rather than generalizable.

Three diagnostics confirm this mechanism empirically. First, the RKL student's output effective rank (number of tokens with meaningful probability) is **4.4** on training data vs. **8.3** for FKL—RKL concentrates on half as many tokens, consistent with memorization of specific token patterns. Second, gradient analysis shows FKL pushes the correct token's logit *upward* 30% of the time, while RKL does so only 19%; RKL's gradient is dominated by pushing *down* wrong tokens (81%) rather than pushing up the right one. Third, the generalization gap confirms the asymmetry: RKL's gap (held-out minus training CE) is 1.40 at τ=0.3, vs. FKL's 1.15—a 22% larger overfitting gap.

Cross-family replication. The overfitting pattern replicates on OPT-1.3B → OPT-125M (10.4×, Table 10, Appendix). At τ=0.5 ($\tau_{\text{eff}} = 0.22$), RKL wins training by $+0.040$ but loses held-out by -0.132. Even at τ=1.0 ($\tau_{\text{eff}} = 0.45$ for OPT, near the zone exit), RKL loses on held-out split (-0.013). This is consistent with the τ_{eff} framework: OPT's lower σ_z (2.25 vs. 3.12) shifts τ_{eff} higher at the same τ, pushing τ=1.0 past the Goldilocks boundary.

Why RKL works at τ=1.0 (GPT-2). At moderate τ, the teacher puts meaningful mass on 5–20 tokens. The $\log(q/p)$ term in RKL's gradient stays bounded because $p(v)$ is no longer near-zero for most of the student's probability mass. Both FKL and RKL students produce similar effective ranks (∼30 tokens) at τ=1.0, confirming that the concentration asymmetry disappears. RKL wins on all three GPT-2 evaluation sets ($+0.004$ to $+0.018$) at $\tau_{\text{eff}} = 0.32$. At $\tau \geq 2.0$, distributions are near-uniform and both divergences give similar gradients.

Implications. Published KD comparisons that evaluate FKL vs. RKL on training loss alone may report misleading rankings. The Goldilocks zone, as measured on training CE, overstates RKL's advantage. On held-out data, RKL's genuine advantage is confined to a narrow regime: $\tau_{\text{eff}} \approx 0.3$ with same-tokenizer distillation. The cross-family OPT results confirm this is not architecture-specific.

9 Per-Token Divergence Does Not Generalize

The phase diagram suggests applying different divergences to different tokens based on teacher confidence. We test this at τ=0.5: splitting tokens at the median teacher entropy and assigning FKL to confident tokens, RKL to uncertain tokens (the configuration that won on training CE in our initial experiments with $+0.027$ CE over uniform FKL).

Table 4 shows a complete ranking reversal at τ=0.5. On training CE, FKL/RKL is the best configuration (0.727). On held-out data, it is the *worst* after uniform RKL (1.836 vs. 1.720 for uniform

Config	τ	Div	CE	Δ
split τ + RKL	0.5/2.0	RKL	**1.352**	—
split τ + FKL	0.5/2.0	FKL	1.354	+0.002
uniform τ + RKL	1.0	RKL	1.479	+0.127
uniform τ + FKL	1.0	FKL	1.484	+0.132

Table 5: Temperature vs. divergence on training CE (4 seeds, std ≤ 0.001). Adaptive τ accounts for 97% of improvement. Per-seed detail in Table 11 (Appendix).

FKL). Any configuration using RKL on any subset of tokens inherits the overfitting problem from §8.

At τ=1.0, the picture changes: all four configurations produce similar held-out CE (range < 0.004 on split, < 0.017 on fresh), with uniform RKL slightly best. Per-token switching does not improve over uniform divergence at either temperature. This confirms that the overfitting is τ-dependent, exactly as the gradient mechanism predicts: at τ=0.5, RKL's $\log(q/p)$ diverges and memorizes; at τ=1.0, it stays bounded and the small RKL advantage generalizes—but per-token selection adds no value beyond choosing the right uniform divergence.

10 Ablation: Temperature Splitting

Temperature dominates divergence at $4\times$. We cross binary temperature (uniform τ=1.0 vs. split τ_{low}=0.5, τ_{high}=2.0) with divergence (FKL vs. RKL) at $4\times$ compression (4 seeds). Temperature splitting reduces training CE by 0.128 (8.6%), while divergence adds only 0.002–0.005 (Table 5). At $2\times$, divergence contributes more (+0.048 CE).

Modality is a brittle proxy. Splitting by modality (NL vs. code) helps CodeLlama (H_{NL}=0.75; −0.019 CE) but hurts GPT-2 XL (H_{NL}=2.10; +0.054 CE). Entropy-based splitting is principled; modality-based splitting is not (Table 6, Appendix).

11 Related Work

Divergence Choice in Knowledge Distillation. The choice of divergence for knowledge distillation (KD) has received renewed attention as distillation becomes the dominant paradigm for deploying large language models. Hinton et al. (2015) introduced KD with forward KL (FKL) divergence and a fixed temperature, treating both as hyperparameters. Subsequent work has explored alternatives, but no consensus has emerged on *when* each divergence is preferable.

MiniLLM (Gu et al., 2024) advocates reverse KL (RKL) for LLM distillation, demonstrating gains on open-ended generation tasks. Their setting involves large compression ratios and tasks where teacher outputs are relatively peaked. Our phase diagram explains this finding: peaked teacher distributions (low entropy) favor RKL's mode-seeking behavior. However, MiniLLM does not characterize *when* RKL loses its advantage.

GKD (Agarwal et al., 2024) studies on-policy distillation with generalized divergences and finds that the optimal divergence is task-dependent. We provide the missing explanatory variable: what GKD observes as "task-dependent" is actually *entropy-dependent*. Tasks with peaked teacher posteriors favor RKL; tasks with diffuse posteriors favor FKL. Our entropy threshold H^* unifies these scattered observations into a single prediction.

Wu et al. (2025) prove that FKL and RKL converge to the same solution given sufficient training, and propose Adaptive KL (AKL), a fixed convex combination. They further argue that the classical mode-seeking/mode-covering characterization from continuous variational inference does not directly apply to discrete softmax distributions. We agree on asymptotic equivalence but show that the *finite-training* regime—which is the regime that matters in practice—has exploitable structure. Specifically, the teacher's token-level entropy determines which divergence converges faster, yielding a measurable advantage that persists across practical training budgets.

Wen et al. (2023) study a family of f-divergences for sequence-level KD, identifying a quality–diversity tradeoff: divergences closer to RKL produce higher-quality but less diverse outputs. Our work provides a mechanistic explanation for this tradeoff via the entropy phase transition: low-entropy tokens benefit from RKL, while high-entropy tokens benefit from FKL.

BDD (Anonymous, 2025) combines FKL and RKL with a fixed blending coefficient $\alpha \approx 4.0$. A fixed blend is suboptimal because the ideal FKL/RKL ratio changes with the teacher's per-token entropy.

Temperature in Knowledge Distillation. Temperature scaling has been part of KD since Hinton et al. (2015), who observed that higher temperatures produce softer probability distributions that transfer more information about inter-class similarities. Most subsequent work treats temperature as a global hyperparameter tuned on a validation set.

EA-KD (Su et al., 2025) uses the teacher's

entropy to *reweight* training samples, assigning higher weight to tokens where the teacher is more uncertain. This is complementary to our approach: EA-KD modulates *how much* to learn from each token, while we modulate *how* to learn (which divergence and what temperature).

Several works have explored learned or adaptive temperature (Li et al., 2023; Jin et al., 2023), typically parameterizing τ as a function of training progress or layer depth. Our temperature selection is instead grounded in information-theoretic quantities: the optimal τ for a given token depends on where the teacher's entropy falls relative to H^*.

Entropy and Information Theory in KD. Information-theoretic perspectives on KD have a long history. Ba and Caruana (2014) and Lopez-Paz et al. (2016) analyze KD through the lens of mutual information, showing that the student learns a compressed representation of the teacher's knowledge. Our work differs in focus: rather than analyzing *what* information is transferred, we study *how* the transfer mechanism (divergence choice) should depend on the teacher's output entropy.

The critical threshold $H^* \approx \frac{2}{3} \ln K$ arises from the entropy of the unigram distribution under Zipf's law. Zipf's law is well-established in natural language (Zipf, 1949; Powers, 1998): word frequencies follow a power law $f(r) \propto r^{-s}$ with $s \approx 1$. For a vocabulary of size K, the entropy of the Zipfian unigram distribution scales as $\frac{2}{3} \ln K$ when $s \approx 1$. This connects our threshold to a fundamental property of language: H^* is not a free parameter but emerges from the statistical structure of natural language.

Mode-Seeking vs. Mode-Covering. The distinction between forward and reverse KL divergence originates in variational inference (VI). Bishop (2006) provides the classical treatment: minimizing $\mathrm{KL}(q\|p)$ (RKL) yields mode-seeking approximations, while minimizing $\mathrm{KL}(p\|q)$ (FKL) yields mode-covering approximations. Minka (2005) extends this to the broader family of α-divergences.

These characterizations were developed for continuous distributions and low-dimensional latent spaces. Wu et al. (2025) question their applicability to KD, where distributions are discrete and high-dimensional ($K > 30$K). We show that the mode-seeking/mode-covering distinction *does* apply to discrete KD, but with a crucial caveat: the behavior is entropy-dependent. Below H^*, teacher distributions are sufficiently peaked that mode-seeking (RKL) exploits their structure. Above H^*, distributions are diffuse enough that mode-covering (FKL) is necessary. At extreme compression ($> 8\times$), limited student capacity makes mode-covering universally preferable. This entropy-dependent phase transition is, to our knowledge, the first characterization of when classical VI intuitions transfer to discrete KD.

12 Discussion and Conclusion

Training loss is not enough. Prior work evaluates FKL vs. RKL on training loss (Gu et al., 2024; Wu et al., 2025), implicitly assuming that training rankings transfer to unseen data. Our results show they do not. At low τ, RKL achieves lower training loss but higher held-out loss—and this affects not just uniform divergence selection (§8) but also per-token divergence switching (§9). Any configuration using RKL at low temperatures overfits, regardless of how selectively it is applied. This means conclusions drawn from training CE alone—including much of the FKL/RKL literature—may be unreliable.

The one-forward-pass diagnostic. Our framework replaces a full hyperparameter sweep (36 runs, ∼18 GPU-hours) with a single forward pass measuring σ_z (∼5 min)—a **31× speedup**. The recipe is simple: (1) If teacher and student use different tokenizers, use FKL at low τ. (2) Otherwise, measure σ_z and compute τ_{eff}. (3) Use FKL as the default divergence. Consider RKL only at $\tau_{\text{eff}} \approx 0.3$ (e.g., τ=1.0 for GPT-2 XL), where it provides a small but robust advantage on held-out data (+0.004–0.018 CE). (4) Use low τ ($\leq$0.5) for $\geq 4\times$ compression; high τ for $\leq 2\times$.

Explaining prior work. MiniLLM's RKL advantage likely reflects training-set overfitting; held-out evaluation may narrow or reverse the gap. GKD's "task-dependent" divergence reflects different σ_z profiles across tasks. Wu et al.'s asymptotic convergence is compatible: the Goldilocks zone is a finite-training phenomenon. Our framework unifies these: the "optimal divergence" depends on τ_{eff}—but only when evaluated on held-out data.

Summary. We introduce $\tau_{\text{eff}} = \tau/\sigma_z$ as the control variable for KD configuration. A logit-scale ablation (+0.049 CE flip) and blind prediction tests validate the framework. Held-out evaluation reveals that training CE overstates RKL's advantage:

at low τ, RKL's mode-seeking overfits to training-specific token patterns, while FKL's mode-covering learns more generalizable representations. This overfitting affects both uniform and per-token RKL application. RKL's genuine advantage is confined to $\tau_{\text{eff}} \approx 0.3$ with same-tokenizer distillation. A 7B cross-tokenizer test identifies a further boundary condition: matched vocabularies are required. The framework reduces KD configuration to a one-forward-pass diagnostic, with FKL as the safe default.

Limitations

Our within-family experiments use teachers up to 1.6B parameters. A 7B cross-tokenizer test confirms σ_z generalizes but reveals that the Goldilocks zone requires matched vocabularies. Within-family validation at 7B+ scale remains open. H^* is a conjecture supported by five distribution families and three architectures, not a theorem. Divergence effect sizes on held-out data are small ($+0.004$–0.018 CE at τ=1.0 for GPT-2), though consistent across seeds. The held-out overfitting pattern is confirmed on two model families (GPT-2 and OPT), but additional families and larger-scale within-family experiments would strengthen the finding. Per-token divergence switching does not generalize at τ=0.5 (§9); at τ=1.0, it provides no advantage over uniform RKL. Downstream task evaluation (e.g., generation quality, factuality) would further strengthen the findings.

Ethical Considerations

This work studies divergence selection in knowledge distillation, a standard model compression technique. We identify no direct risks from the findings. The τ_{eff} diagnostic could reduce the computational cost of KD hyperparameter tuning, which has a modest positive environmental impact. All models used are publicly released under open licenses. No human subjects or private data were involved.

Acknowledgements

An AI assistant (Claude, Anthropic) was used for writing assistance, including paraphrasing, drafting, and code generation for experiment scripts. All content was reviewed, verified, and approved by the authors.

References

Rishabh Agarwal, Nino Vieillard, Piotr Stanczyk, Sabela Roux, Matthieu Geist, and Olivier Bachem. 2024. On-policy distillation of language models: Learning from self-generated mistakes. In *ICLR*.

Anonymous. 2025. Balance divergence for knowledge distillation. *Pattern Recognition*. ArXiv:2501.07804.

Jimmy Ba and Rich Caruana. 2014. Do deep nets really need to be deep? *NeurIPS*.

Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning*. Springer.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. 2024. MiniLLM: Knowledge distillation of large language models. In *ICLR*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. In *NIPS Deep Learning Workshop*.

Haoran Jin and 1 others. 2023. Adaptive temperature scheduling for knowledge distillation. *arXiv preprint*.

Zheng Li, Xiang Li, Lingfeng Yang, and 1 others. 2023. Curriculum temperature for knowledge distillation. *AAAI*.

David Lopez-Paz, Léon Bottou, Bernhard Schölkopf, and Vladimir Vapnik. 2016. Unifying distillation and privileged information. In *ICLR*.

Tom Minka. 2005. Divergence measures and message passing. Technical report, Microsoft Research.

David M. W. Powers. 1998. Applications and explanations of Zipf's law. *Joint Conference on New Methods in Language Processing and Computational Natural Language Learning*.

Shangquan Su, Shan Li, and 1 others. 2025. Entropy-aware knowledge distillation. In *ICCV*. ArXiv:2311.13621.

Jianhao Wen, Pei Ke, Shaohan Huang, Binhang Li, and Furu Wei. 2023. f-divergence minimization for sequence-level training of neural machine translation. In *ACL*.

Taiqiang Wu, Chaofan Tao, Jiahao Wang, Zhe Yang, Dongliang Zhao, and Ngai Wong. 2025. Rethinking kullback-leibler divergence in knowledge distillation for large language models. In *COLING*.

George Kingsley Zipf. 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley.

A Additional Tables

B Full Per-Run Results

We report all individual training runs (224 total) across GPT-2 and OPT families. All runs use 3,000 steps, batch size 4, α=0.7, lr=5×10^{-5}.

Teacher	H_{NL}	H_{code}	Effect
CodeLlama-13B	**0.75**	1.72	-0.019
GPT-2 XL	2.10	1.58	$+0.054$

Table 6: Per-modality temperature effect. Sharpening helps only when $H < 1.0$.

Thr.	H^*	%high	Best	CE
p50	2.12	57%	τ+FKL	1.415
p75	3.55	29%	τ+FKL	1.392
p90	4.57	12%	τ+FKL	1.372

Table 7: Per-token divergence at multiple thresholds (3 seeds). Split τ + FKL wins everywhere; the "correct" direction loses by 0.006–0.018 CE.

B.1 GPT-2 Phase Diagram

τ	H/H^*	2×	4×	12×
0.5	0.12	$+.\mathbf{048}$	$+.002$	$-.027$
1.0	0.35	$+.\mathbf{006}$	$+.\mathbf{006}$	$-.014$
2.0	1.13	$-.015$	$-.011$	$-.019$
4.0	1.44	$-.001$	$+.002$	$-.007$
8.0	1.48	$-.001$	$-.003$	$-.003$
16.0	1.49	$+.001$	$-.000$	$-.002$

Table 8: GPT-2 phase diagram: $\Delta = \text{CE}_{\text{FKL}} - \text{CE}_{\text{RKL}}$ (+ = RKL better). 3 seeds for 2×/4×; 1 seed for 12×. Bold = $|\Delta| > 0.003$.

Per-seed detail (4×, selected τ). At τ=1.0: seeds 7/42/123 give FKL CE = 1.484/1.483/1.484 vs. RKL CE = 1.478/1.477/1.479. RKL wins all 3 seeds. At τ=0.5: seeds 7/42/123 give FKL CE = 1.355/1.352/1.354 vs. RKL CE = 1.351/1.352/1.352. Near-tie, all $|\Delta| < 0.003$. Standard deviations across seeds: ≤ 0.002 for all configurations.

Per-seed detail (2×, τ=0.5). Seeds 7/42/123 give FKL CE = 1.150/1.148/1.149 vs. RKL CE = 1.102/1.101/1.100. $\Delta = +0.048 \pm 0.001$. RKL wins all 3 seeds with the largest effect in the study.

B.2 OPT Cross-Family

OPT per-seed detail (τ=1.0, training CE). OPT-350M: seeds 7/42/123 give Δ = +0.011/+0.011/+0.009. All RKL. OPT-125M: seeds 7/42/123 give Δ = +0.008/+0.008/+0.008. All RKL. Maximum seed-to-seed std: 0.001 CE.

		OPT 3.7× (350M)		OPT 10.4× (125M)	
τ	H/H^*	FKL	RKL	FKL	RKL
0.5	0.08	**1.559**	1.582	**1.669**	1.688
1.0	0.23	1.668	**1.658**	1.761	**1.753**
2.0	0.76	1.586	1.587	1.711	1.714
4.0	0.98	1.491	1.493	1.623	1.623
8.0	1.00	1.493	1.492	1.614	1.615
16.0	1.00	1.493	1.493	1.612	1.611

Table 9: OPT cross-family: mean CE across 3 seeds. RKL at τ=1.0 replicates on both student sizes. Bold = $|\Delta| > 0.003$.

τ	τ_{eff}	Train Δ	Split Δ	Fresh Δ	Train	Split
0.3	0.13	$-.004$	$-.125$	$-.129$	FKL	FKL
0.5	0.22	$+.040$	$-.132$	$-.147$	RKL	**FKL**
1.0	0.45	$+.042$	$-.013$	$-.002$	RKL	**FKL**
2.0	0.89	$+.000$	$-.010$	$-.010$	TIE	FKL
4.0	1.78	$+.003$	$-.006$	$-.006$	RKL	FKL

Table 10: OPT-1.3B → OPT-125M held-out evaluation ($\sigma_z = 2.25$, 5 seeds). RKL overfitting replicates: at τ=0.5, RKL wins training by $+0.040$ but loses held-out by -0.132. Even at τ=1.0 ($\tau_{\text{eff}} = 0.45$, near zone exit), FKL wins on held-out split.

B.3 OPT Held-Out Evaluation

B.4 Temperature × Divergence Ablation (Executor C)

B.5 Per-Token Divergence Switching

B.6 Cross-Architecture Scale Test (7B)

Mistral-7B (7.2B, $\sigma_z = 2.89$) → GPT-2 Medium (355M), 20× compression, cross-tokenizer (K_T=32,000, K_S=50,257; logits truncated to shared 32K subspace). FKL dominates at every temperature (Table 13), including $\tau_{\text{eff}} = 0.35$ where the Goldilocks rule predicts RKL. RKL's mode-seeking concentrates on teacher tokens that may not correspond to coherent student tokens under a different BPE vocabulary. This makes concentration fragile. FKL's coverage objective does not depend on token identity.

C Entropy Distribution

C.1 Teacher Logit Scale

Table 15 reports σ_z for each teacher model. Notably, σ_z does not grow monotonically with model size.

C.2 Per-Token Entropy Distribution (GPT-2 XL, τ=1)

The per-token entropy of GPT-2 XL's output distribution at τ=1 is strongly right-skewed: the teacher

τ config	Div config	CE by seed				Mean	Std
		7	42	123	456		
0.5/2.0	RKL/FKL	1.351	1.352	1.351	1.353	**1.352**	0.001
0.5/2.0	RKL/RKL	1.351	1.352	1.353	1.352	1.352	0.001
0.5/2.0	FKL/RKL	1.355	1.352	1.354	1.354	1.354	0.001
0.5/2.0	FKL/FKL	1.355	1.353	1.354	1.354	1.354	0.001
1.0/1.0	RKL/RKL	1.479	1.478	1.480	1.478	1.479	0.001
1.0/1.0	RKL/FKL	1.479	1.478	1.478	1.479	1.479	0.001
1.0/1.0	FKL/RKL	1.485	1.482	1.484	1.484	1.484	0.001
1.0/1.0	FKL/FKL	1.484	1.483	1.484	1.485	1.484	0.001

Table 11: Executor C: all 32 individual runs (8 configs $\times$ 4 seeds). Temperature splitting (τ=0.5/2.0 vs. 1.0/1.0) accounts for >97% of CE improvement. Note: $H^*=\ln K$=10.82 in this experiment (before the $\frac{2}{3}$ correction), so 100% of tokens fall in the low-entropy regime and the high-H divergence is unused.

Low-H	High-H	H^*	CE by seed			Mean
		(p50)	7	42	123	
FKL	RKL	2.12	1.326	**1.325**	1.327	**1.326**
RKL	RKL	2.12	1.350	1.352	1.353	1.352
FKL	FKL	2.12	1.355	1.352	1.354	1.353
RKL	FKL	2.12	1.375	1.375	1.377	1.376

Table 12: Per-token divergence switching at τ=0.5, 4$\times$ compression, all 12 runs. Entropy split at p50 (H^*=2.12 nats, 42.8% low-entropy tokens). FKL-confident + RKL-uncertain wins all seeds with std $= 0.0008$.

is confident on most tokens and uncertain on few.

At τ=1, only $\sim$1% of tokens exceed $H^* = 7.25$ nats. This explains why Executor C's per-token divergence split was ineffective when using $H^* = \ln K = 10.82$: *zero* tokens crossed the threshold. After correcting to empirical percentile thresholds (p50 $= 2.12$ nats), per-token divergence switching becomes effective (Table 12).

The practical implication: per-token divergence switching should use the teacher's *empirical* entropy median, not the Zipf-derived H^*. H^* governs the run-level phase transition—which global divergence is optimal. The per-token split requires a threshold calibrated to the teacher's actual entropy distribution.

C.3 Effective Vocabulary Under Zipf's Law

For a vocabulary of $K = 50{,}257$ tokens following Zipf's law ($s \approx 1$), the effective vocabulary at H^* is:

$$K_{\text{eff}} = e^{H^*} = K^{2/3} \approx 1{,}362 \text{ tokens} \qquad (4)$$

The teacher's contextual predictions concentrate on $\sim$2.7% of the vocabulary on average. This is consistent with the observed entropy distribution: the median token has perplexity $e^{2.12} \approx 8$, meaning the teacher considers $\sim$8 plausible continuations.

τ	τ_{eff}	FKL	RKL	Δ	Actual	Pred.	
0.5	0.17	**1.573**	1.999	-0.426	FKL	FKL	✓
1.0	0.35	**1.650**	2.343	-0.693	FKL	RKL	$\times$
2.0	0.69	**1.986**	2.473	-0.487	FKL	FKL	✓
4.0	1.38	**1.790**	1.822	-0.031	FKL	FKL	✓

Table 13: Cross-architecture: Mistral-7B $\rightarrow$ GPT-2 Medium (20$\times$, 3 seeds, std $\leq$ 0.01). FKL dominates at all τ; the Goldilocks zone breaks under tokenizer mismatch.

		FKL CE by seed			RKL CE by seed		
τ	τ_{eff}	42	123	7	42	123	7
0.5	0.17	1.574	1.573	1.573	2.007	1.997	1.993
1.0	0.35	1.655	1.645	1.650	2.335	2.348	2.344
2.0	0.69	1.989	1.981	1.988	2.484	2.473	2.461
4.0	1.38	1.792	1.784	1.794	1.824	1.815	1.826

Table 14: Mistral-7B $\rightarrow$ GPT-2 Medium (20$\times$, cross-tokenizer): all 24 runs. FKL wins at every temperature. Max per-seed std: 0.010 (RKL at τ=2.0); FKL std ≤ 0.005.

Teacher	Params	σ_z	τ_{eff} at τ=1	K
GPT-2 Medium	355M	3.94	0.25	50,257
GPT-2 XL	1.6B	3.12	0.32	50,257
Mistral-7B	7.2B	2.89	0.35	32,000
OPT-1.3B	1.3B	2.25	0.44	50,272

Table 15: Teacher logit scale (σ_z) across model families. The same nominal τ produces different τ_{eff}. Mistral-7B has lower σ_z than GPT-2 XL despite being 4.5$\times$ larger.

Percentile	H (nats)	Regime
p10	0.15	Near-deterministic
p25	0.52	Low entropy
p50	2.12	Moderate
p75	3.55	High
p90	4.57	Very high
p95	5.50	Approaching H^*
p99	6.82	Near $H^* = 7.25$
$H^* = \frac{2}{3} \ln K$		$= 7.25$ nats
$\ln K$		$= 10.82$ nats

Table 16: GPT-2 XL per-token entropy percentiles at τ=1 (over >100K tokens from 2,000 generated examples). Even p99 falls below H^*, confirming that the run-level entropy threshold (Figure 1) governs divergence selection, not per-token entropy at τ=1.