Forward vs. reverse KL divergence in language model knowledge distillation: An empirical investigation

Anonymous ACL submission

Abstract

Classical analyses portray forward KL (FKL) as mass-covering and reverse KL (RKL) as mode-seeking, but these intuitions arise from continuous toy settings. We revisit this dichotomy in *discrete*, token-level language-model distillation under a controlled teacher–student setup (LLaMA-360M → BabyLlama-58M) with 10M tokens. We use this training setup to compare FKL and RKL across phases and head–tail strata, using teacher-side probability masks and frequency slices. Empirically, both directions converge to similar end states, yet *early* behaviors diverge: FKL aligns head tokens faster, while RKL suppresses tail errors sooner. These effects are magnified at linguistically critical loci and systematically moderated by domain-level properties: long-form text favors FKL improvements, short-utterance domains favor RKL. Our findings clarify *when and where* KL direction matters in practice, and motivate phase-aware or hybrid divergences that exploit complementary strengths under constrained compute budgets.

1 Introduction

In machine learning, knowledge distillation (KD) compresses a large teacher model into a smaller student model by aligning their predictive distributions, typically using Kullback–Leibler (KL) divergence (Hinton et al., 2015). In language modelling, this strategy is particularly useful when exploring the linguistic abilities of very small models. A prime example is the BabyLM Challenge (Warstadt et al., 2023; Hu et al., 2024), a shared task on data-efficient pretraining of small language models (LMs) on developmentally (cognitively) plausible corpora, enforcing strict text budgets (10M or 100M words). Indeed, in this challenge, distillation of small LMs is a common strategy adopted by numerous participating teams (e.g., Timiryasov and Tastet, 2023; Tastet and Timiryasov, 2024; Nair et al., 2024).

A key choice to be made in the KD process is the direction of the KL divergence: whether to measure divergence from the teacher to the student (forward KL) or vice versa (reverse KL). Forward KL (FKL) is often described as *mean-seeking* (or zero-avoiding): it heavily penalizes missing teacher probability mass and therefore encourages the student to cover *all* regions with non-negligible teacher probability, while reverse KL (RKL) is *mode-seeking* (or zero-forcing): it penalizes allocating probability where the teacher's probability is very small, pushing the student to concentrate on high-probability modes and avoid low-probability regions (Bishop and Nasrabadi, 2006; Murphy, 2012). This intuition originates from analyses on continuous toy distributions, such as Gaussians or mixtures thereof, and has guided practical choices of loss functions across multiple ML domains (Agarwal et al., 2024; Gu et al., 2024).

The application of this intuition to language modelling is, however, not straightforward, since next-token distributions in language models follow Zipf-like laws with heavy heads and long tails (Piantadosi, 2014; Holtzman et al., 2019; Takahashi and Tanaka-Ishii, 2017). Indeed, Wu et al. (2025) have recently challenged the direct transfer of the classical mean-/mode-seeking narrative to language model KD. They explicitly note two premises under which that narrative is typically derived, neither of which holds in the context of language modelling: (i) the student distribution is unimodal (often Gaussian); (ii) both teacher and student are continuous. In contrast, KD for LLMs is a discrete, token-level process with Zipf-like long tails. Using toy data, Wu et al. (2025) show *theoretically* that with sufficient optimization, FKL and RKL converge to the same fixed point; yet, in the early regime, the two directions emphasize different regions: FKL aligns faster on the head, while RKL attends more to the tail. This is an interesting finding, but what precisely happens *in practice* with FKL vs. RKL in

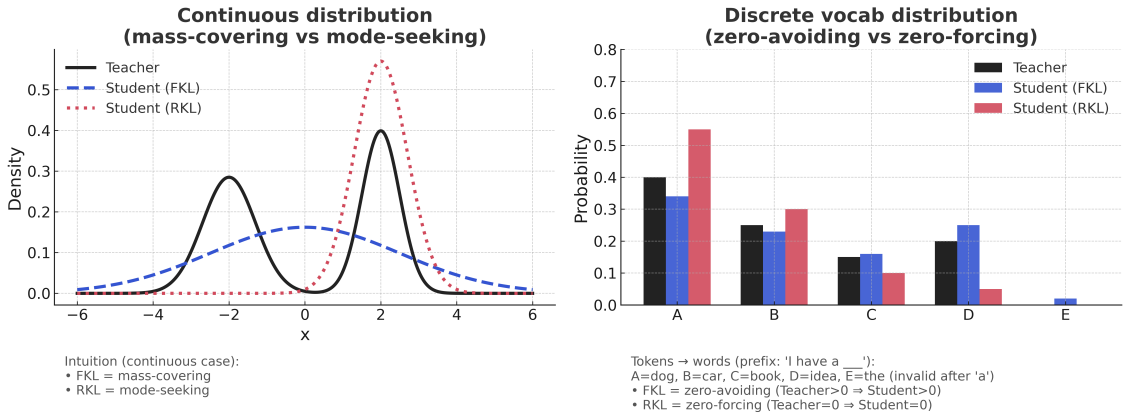

Figure 1: Intuition for forward vs. reverse KL. (**Left**) Continuous distributions: FKL (blue) spreads probability mass (mass-covering/mean-seeking behavior), while RKL (red) sharpens peaks (mode-seeking behavior). (**Right**) Discrete Zipf-like token distributions on the teacher rank axis: FKL yields a less concentrated distribution (more tail mass), whereas RKL produces a more peaked distribution (more head concentration). In softmax LMs, non-masked tokens satisfy $p(i) > 0$; thus "zero-avoiding/zero-forcing" should be interpreted as penalties under *support mismatch* and as differing emphasis on near-zero regions, rather than literal zeros.

the distillation process of language models is not yet uncovered.

Building on Wu et al. (2025)'s observations based on toy data, we study *empirically* the behaviour of FKL and RKL in the context of an actual model distillation process with real language data both in training and at test time, leveraging the BabyLM Challenge setup. We define head and tail on the teacher token axis in two complementary views. The *global view* reflects corpus-level frequency (high- vs. low-frequency types), while the *local view* reflects position-wise teacher probabilities (high- vs. low-mass regions within each conditional distribution). Using a controlled teacher–student setup on the 10M ('strict-small') BabyLM corpus, we study FKL vs. RKL in *discrete* KD and ask: *(RQ1)* When and where does the KL direction matter in discrete LM KD (early vs. late; head vs. tail; critical tokens)? *(RQ2)* How large are these effects and how do they evolve across training steps and text features (different domains, sentence length)? *(RQ3)* What practical guidance follows for choosing or scheduling divergences under compute constraints?

Our contributions are four-fold: (1) To our knowledge, this is the **first empirical study** to systematically examine the behavioural differences between forward and reverse KL in *real* language-model distillation with *real* data and models, beyond toy settings. (2) We propose two complementary **head–tail frameworks**—a global corpus-level view and a local teacher-side view—and use them to characterize when and where FKL and RKL diverge. Under the global view, we discover that FKL and RKL exhibit *opposite preferences* with respect to sentence length: long-form domains favor FKL, while short-utterance domains favor RKL. (3) We **focus on linguistically critical tokens** to probe grammar-sensitive loci, showing that early-phase head–tail asymmetries are amplified at these loci. (4) Under the local view, we design **three diagnostics**—per-mass KL decomposition, head-deficit/tail-leakage, and asymmetry indices—to interpret the dynamics of FKL and RKL within head and tail regions, confirming previous theoretical claims about phase-specific head/tail preferences.

2 Background and Motivation

In continuous density estimation, FKL is often labeled mean-seeking or mass-covering, while RKL is mode-seeking or zero-forcing, a dichotomy grounded in information-geometry and support-mismatch analyses (Amari and Cichocki, 2010; Bishop and Nasrabadi, 2006; Murphy, 2012). For *discrete* next-token distributions in LMs, however, this framing is only approximate: softmax normalization yields full (nonzero) support over tokens, and empirical probabilities are heavily skewed by Zipf-like head–tail structure rather than true zeros (Piantadosi, 2014; Holtzman et al., 2019). A more faithful reading in token spaces is that RKL penalizes allocations in near-zero-under-teacher regions

more steeply, whereas FKL is more sensitive to missing teacher mass, see Figure 1.

Building on this intuition, recent KD studies have experimented beyond a fixed FKL objective. Some use RKL to mitigate overestimation of low-probability regions and report gains on instruction-following tasks (Gu et al., 2024). Others extend the loss family (e.g., f-divergences, α-divergence, α–β divergence) to interpolate between FKL-like and RKL-like behaviors and to separately control hardness vs. confidence concentration (Agarwal et al., 2024; Wang et al., 2025). At a finer granularity, token-wise mixtures (Jung et al., 2025) adaptively weight FKL/RKL per token, leveraging the observation that FKL tends to boost underestimated tokens while RKL suppresses overestimated ones.

Importantly for our study, Wu et al. (2025) challenge a literal transfer of the 'mean- vs. mode-seeking' narrative to token-level KD: with sufficient optimization, FKL and RKL converge to the same optimum, while in early training they emphasize different regions—FKL faster aligns the head, RKL the tail. These insights, however, are largely established on toy analyses, leaving open how such behaviors evolve over time on real language data, how they vary by token frequency strata, and how strongly they manifest at linguistically critical loci.

To probe these questions in a realistic yet controlled regime, we ground our study in the BabyLM challenge, a shared task on data-efficient pretraining of small LMs under strict text budgets (e.g., 10M or 100M words) with unified evaluation (Warstadt et al., 2023; Hu et al., 2024). Across recent editions, many competitive systems explicitly adopt teacher→student KD under these constraints (Timiryasov and Tastet, 2023; Tastet and Timiryasov, 2024; Nair et al., 2024), making BabyLM a natural testbed for studying KL-direction effects in small-model distillation. In parallel, minimal-pair grammatical paradigms such as BLiMP are widely used for targeted evaluation (Warstadt et al., 2020); they isolate token-level decisions where small probability reallocations can flip a model's preference, and thus are well-suited to diagnose head–tail alignment.

Building on theoretical and toy analyses (Wu et al., 2025) showing that FKL and RKL share the same global optimum but differ in their early emphasis, we ask how these dynamics manifest in practice for token-level KD—how the two directions diverge across training phases, frequency strata, and linguistically critical loci. Understanding these patterns is key to interpreting the transient behaviors often observed during distillation. We examine them empirically in a controlled teacher–student setup, as described below.

3 Method

3.1 Preliminaries

Knowledge distillation. KD transfers the teacher's conditional distribution $p_t(\cdot)$ to a smaller student $q_t(\cdot)$ over a vocabulary $\mathcal{V}$ of size $|\mathcal{V}|$. A standard choice of loss is the KL divergence in one of the two directions:

$$\begin{aligned} D_{\mathrm{FKL}}(p_t \| q_t) &= \sum_{i=1}^{V} p_t(i) \, \log \frac{p_t(i)}{q_t(i)}, \\ D_{\mathrm{RKL}}(q_t \| p_t) &= \sum_{i=1}^{V} q_t(i) \, \log \frac{q_t(i)}{p_t(i)}. \end{aligned} \tag{1}$$

Their local gradients differ:

$$\begin{aligned} \frac{\partial}{\partial q_{t,i}} D_{\mathrm{FKL}} &= -\frac{p_{t,i}}{q_{t,i}}, \\ \frac{\partial}{\partial q_{t,i}} D_{\mathrm{RKL}} &= \log \frac{q_{t,i}}{p_{t,i}} + 1. \end{aligned} \tag{2}$$

These gradient forms reveal complementary tendencies: FKL places larger penalties where the teacher assigns non-negligible mass but the student underestimates it, encouraging broad coverage of the teacher support; RKL instead penalizes overestimation in low-probability regions, driving the student to focus on high-confidence tokens. Before studying FKL's and RKL's behaviour at token level and across training phases, we need to provide a working definition of head and tail in the context of language modelling.

Head and Tail. We define "head" and "tail" in two complementary ways, both grounded in the *teacher* and evaluated on the same *teacher token axis* (identical tokenizer/vocabulary across models): a *global* corpus-frequency partition, paralleling Anand et al. (2024), and a *local* teacher-side nucleus (Top-p) mask inspired by Wu et al. (2025). Threshold choices and formal definitions are given below.

- **Global (corpus-level) teacher view**. Let $\mathcal{V}$ be the teacher vocabulary ($|\mathcal{V}| \approx 16$k). Using the teacher's pretraining corpus, denote the empirical count of type $w \in \mathcal{V}$ by $c(w)$, and its normalized

frequency by:

$$f(w) \;=\; \frac{c(w)}{\sum_{u\in\mathcal{V}} c(u)}. \tag{3}$$

Sorting all tokens by $f(\cdot)$ in non-increasing order gives their ranks $r(w) \in \{1, \ldots, |\mathcal{V}|\}$. This allows us to define fixed (context-agnostic) frequency slices for the head and tail, respectively:

$$\begin{aligned} H^{\text{freq}}_{\text{Top-1k}} &= \{\, w \in \mathcal{V} : r(w) \le 1000 \,\}, \\ T^{\text{freq}}_{>10\text{k}} &= \{\, w \in \mathcal{V} : r(w) > 10{,}000 \,\}, \end{aligned} \tag{4}$$

and analogously the mid-frequency band 1k–10k. These rank-based slices underlie the Top-1k / 1k–10k / >10k analyses in our results.

- **Local (token-level) teacher view.** At position t with left context $x_{<t}$, let the teacher's conditional distribution be $p_t(\cdot) = p(\cdot \mid x_{<t})$. We sort its probabilities in descending order by defining a permutation π_t such that:

$$p_t\big[\pi_t(1)\big] \ge p_t\big[\pi_t(2)\big] \ge \cdots \ge p_t\big[\pi_t(|\mathcal{V}|)\big] \tag{5}$$

For a cumulative-mass threshold $\alpha \in (0, 1)$ (nucleus/top-p), define the minimal prefix size as follows:

$$K(t, \alpha) \;=\; \min\Big\{\, k \in \{1, \ldots, |\mathcal{V}|\} : \sum_{j=1}^{k} p_t\big[\pi_t(j)\big] \ge \alpha \Big\}. \tag{6}$$

The head and tail sets are then computed as:

$$\begin{aligned} H_t(\alpha) &= \{\pi_t(1), \ldots, \pi_t(K(t, \alpha))\}, \\ T_t(\alpha) &= \mathcal{V} \setminus H_t(\alpha). \end{aligned} \tag{7}$$

This teacher-side, context-dependent masking follows the intuition of Wu et al. (2025): for each t, the mask depends only on the teacher p_t and remains invariant across checkpoints or students.

Logging and residual. For efficiency we export the teacher's top-P probabilities and aggregate the remainder into a residual bin:

$$r_t \;=\; 1 \;-\; \sum_{i \in \text{Top-}P} p_t(i), \tag{8}$$

Unless otherwise stated, we use the teacher-side cumulative probability mass (CPM, nucleus/top-p) threshold α=0.5 for all main results; analyses under higher thresholds (0.8, 0.9, 0.95) are provided in the Appendix D.

3.2 Global View Metrics

This section defines the win-rate metric used in §4.2–4.3 below. Here, positions are first stratified by the *teacher's* corpus-frequency ranks as described above: Top-1k, 1k–10k, and >10k. Within each frequency stratum, we compare students by their absolute probability error on the ground-truth next token y_t, measured under the teacher's vocabulary axis.

Per-position absolute-error kernel. Let y_t denote the next ground-truth token at position t. For each model $m \in \{\text{FKL}, \text{RKL}\}$, we measure its absolute probability error on that token:

$$e^{(m)}(t) \;=\; \big|\, p_t(y_t) - q_t^{(m)}(y_t) \,\big|, \tag{9}$$

where $p_t(y_t)$ and $q_t^{(m)}(y_t)$ are the teacher's and the student's probabilities assigned to the true next token. A win for model m at t is recorded when its error is strictly smaller than the other model's:

$$\begin{aligned} \text{Win}^{\text{FKL}}(t) &= \mathbb{1}\big[e^{(\text{FKL})}(t) < e^{(\text{RKL})}(t)\big], \\ \text{Win}^{\text{RKL}}(t) &= \mathbb{1}\big[e^{(\text{RKL})}(t) < e^{(\text{FKL})}(t)\big]. \end{aligned} \tag{10}$$

We attribute the span to the frequency bin of its onset token y_t (by the teacher rank $r(y_t)$).

Bin- and phase-level aggregation. For a frequency bin $b \in \{\text{Top-1k}, \text{1k–10k}, >\text{10k}\}$ at checkpoint c, define:

$$\begin{aligned} \text{WinRate}_b^{\text{FKL}}(c) &= \frac{\sum_{t\in b(c)} \text{Win}^{\text{FKL}}(t)}{\sum_{t\in b(c)} 1}, \\ \text{WinRate}_b^{\text{RKL}}(c) &= \frac{\sum_{t\in b(c)} \text{Win}^{\text{RKL}}(t)}{\sum_{t\in b(c)} 1}. \end{aligned} \tag{11}$$

Phase summaries average these rates over selected checkpoints. We also compare early vs. final stages and correlate the resulting differences with average sentence length.

3.3 Local View Metrics

At each position t, the teacher p_t and a student q_t assign probabilities over the shared vocabulary (teacher's tokenizer). We construct a **teacher–side cumulative–mass** head–tail mask on the teacher axis: sort tokens by $p_t(i)$ and take the smallest prefix that covers α of teacher mass as the head H_t; the tail T_t contains all remaining tokens *plus a residual bin* to preserve total probability. Write,

for a slice $S \in \{H, T\}$:

$$P_t(S) = \sum_{i \in S_t} p_t(i), \quad Q_t(S) = \sum_{i \in S_t} q_t(i), \tag{12}$$

All sums and masks are on the teacher token axis and shared across students. We use $\varepsilon = 10^{-9}$ to avoid $\log 0$.

Per-mass KL (size-free within a slice). To remove dependence on slice probability mass, we compute a per-mass KL on a *slice* $S \in \{H, T\}$ (head or tail), with the teacher as the reference:

$$\mathrm{PM}_S^{\mathrm{FKL}}(t) = \frac{\sum_{i \in S_t} p_t(i) \, \log \frac{p_t(i)}{\max(q_t(i), \varepsilon)}}{P_t(S)}. \tag{13}$$

This quantity admits a simple decomposition:

$$\mathrm{PM}_S^{\mathrm{FKL}}(t) = D_{\mathrm{KL}}\big(p_t|S \,\|\, q_t|S\big) + \log \frac{P_t(S)}{\max(Q_t(S), \varepsilon)}. \tag{14}$$

The first term is the *shape* mismatch (conditional KL between the normalized distributions within the slice), and the second term is the *mass-allocation* mismatch (total teacher vs. student probability placed on the slice). Lower is better: $\mathrm{PM}_S^{\mathrm{FKL}}(t) = 0$ iff both the conditional shapes match and $Q_t(S) = P_t(S)$. The score increases when the conditional shapes diverge and/or when the student under-allocates slice mass ($Q_t(S) < P_t(S)$); it can be negative when $Q_t(S) > P_t(S)$ and the mass-ratio term dominates the conditional KL. We report $\mathrm{PM}_H^{\mathrm{FKL}}(t)$ and $\mathrm{PM}_T^{\mathrm{FKL}}(t)$ for both students under the same teacher p_t.

Quality mismatch: Head Deficit and Tail Leakage. To isolate total-mass misallocation independent of within-slice shape, we use *head deficit (HD)* and *tail leakage (TL)*. HD is the shortfall of the student's head mass relative to the teacher; TL is the excess of the student's tail mass relative to the teacher:

$$\mathrm{HD}(t) = \max\big(P_t(H) - Q_t(H),\, 0\big), \quad \mathrm{TL}(t) = \max\big(Q_t(T) - P_t(T),\, 0\big). \tag{15}$$

By mass conservation, $P_t(H) - Q_t(H) = Q_t(T) - P_t(T)$, so the absolute quantities coincide: $\mathrm{HD}(t) = \mathrm{TL}(t)$. This equality holds for the absolute definitions; the relative versions generally differ due to different normalizers. We retain both names for interpretability (head shortfall vs. tail overshoot). For scale-free views, we report the *relative* forms:

$$\mathrm{HD}_{\mathrm{rel}}(t) = \frac{\mathrm{HD}(t)}{\max(P_t(H), \varepsilon)}, \quad \mathrm{TL}_{\mathrm{rel}}(t) = \frac{\mathrm{TL}(t)}{\max(P_t(T), \varepsilon)}. \tag{16}$$

Interpretation and bounds. $\mathrm{HD}(t) = 0$ iff $Q_t(H) \geq P_t(H)$; $\mathrm{TL}(t) = 0$ iff $Q_t(T) \leq P_t(T)$. Both $\mathrm{HD}(t)$ and $\mathrm{TL}(t)$ lie in $[0, 1]$. Moreover, $\mathrm{HD}_{\mathrm{rel}}(t) \in [0, 1]$, while $\mathrm{TL}_{\mathrm{rel}}(t) \in [0, \infty)$ and can exceed 1 when $P_t(T)$ is small.

Asymmetry between head and tail. To summarize which slice contributes more under each objective, we use *asymmetry indices (raw)*:

$$A^{\mathrm{FKL}}(t) = \frac{\mathrm{FKL}_H(t) - \mathrm{FKL}_T(t)}{\mathrm{FKL}_H(t) + \mathrm{FKL}_T(t) + \varepsilon}, \quad A^{\mathrm{RKL}}(t) = \frac{\mathrm{RKL}_T(t) - \mathrm{RKL}_H(t)}{\mathrm{RKL}_H(t) + \mathrm{RKL}_T(t) + \varepsilon}. \tag{17}$$

Here $\mathrm{FKL}_S(t) = \sum_{i \in S_t} p_t(i) \log \frac{p_t(i)}{\max(q_t(i), \varepsilon)}$ and $\mathrm{RKL}_S(t) = \sum_{i \in S_t} q_t(i) \log \frac{q_t(i)}{\max(p_t(i), \varepsilon)}$. Positive A^{FKL} means FKL error is concentrated in the head (as expected); positive A^{RKL} means RKL error is concentrated in the tail.

Per-mass asymmetry (scale-free). To remove dependence on slice probability mass, we also report the *per-mass* versions:

$$A^{\mathrm{FKL\text{–}PM}}(t) = \frac{\mathrm{PM}_H^{\mathrm{FKL}}(t) - \mathrm{PM}_T^{\mathrm{FKL}}(t)}{\mathrm{PM}_H^{\mathrm{FKL}}(t) + \mathrm{PM}_T^{\mathrm{FKL}}(t) + \varepsilon}, \quad A^{\mathrm{RKL\text{–}PM}}(t) = \frac{\mathrm{PM}_T^{\mathrm{RKL}}(t) - \mathrm{PM}_H^{\mathrm{RKL}}(t)}{\mathrm{PM}_H^{\mathrm{RKL}}(t) + \mathrm{PM}_T^{\mathrm{RKL}}(t) + \varepsilon}, \tag{18}$$

where $\mathrm{PM}_S^{\mathrm{FKL}}(t) = \frac{\mathrm{FKL}_S(t)}{P_t(S)}$ and $\mathrm{PM}_S^{\mathrm{RKL}}(t)$ is defined analogously for fair, scale-free comparison on the teacher axis. Positive signs have the same interpretation as the raw versions, and both per-mass indices lie in $(-1, 1)$.

Aggregation. At each position t, the teacher-side mask (H_t, T_t) is constructed from p_t. All metrics for all students at that position are computed under this same teacher-defined mask, ensuring comparability across objectives. The residual bin, if present, is assigned to the tail T_t.

4 Experiments and Results

4.1 Experimental Setup

Models. We follow the training setup of Tastet and Timiryasov (2024) and use LLaMA-360M as the teacher and BabyLlama-58M as the student, trained with either FKL or RKL. All evaluations are performed on the teacher's tokenizer (vocabulary size $|\mathcal{V}|$=16k) to ensure a shared token axis.

Data. We train models on the 10M-word data set from the BabyLM Challenge 2024 (Choshen et al., 2024). Evaluation is performed on two types of data. First, we consider the test portion of the BabyLM data set, covering the six text domains also present in the training data, to measure the head and tail behavior on the in-domain data (Table 1). Second, we consider a subset of grammatical sentences from BLiMP (Warstadt et al., 2020), a commonly used benchmark that was also used in the BabyLM evaluation, to assess out-of-distribution generalisation.

Training and phases. We use Adam optimizer, same schedule for both students, temperature T=2.0 at eval, with checkpoints saved along the training trajectory up to 16,520 steps. The phases are fixed by step count: **Early** – up to 2k training steps; **Mid** – intermediate checkpoints; **Final** – last 3 checkpoints.

4.2 BabyLM Test: Domain and Length Effects

Using the *global view* defined in §3.1, we evaluate six text domains from the BabyLM test data, comprising ∼1.1M sentences (Table 1). We report (i) per-domain win rates by frequency bin (Head: Top-1k; Mid: 1k–10k; Tail: >10k) at **early** and **final** checkpoints, and (ii) correlations between average sentence length and the change in FKL win rate over the training time (i.e., Final minus Early) across domains.

More formally, for each position t with ground-truth token y_t, we compute the absolute-error kernel as in Eq. (9), derive per-position win indicators as in Eq. (10), and aggregate wins within each frequency bin b at checkpoint c using Eq. (11) to obtain $\mathrm{WinRate}_b^{\mathrm{FKL}}(c)$ and $\mathrm{WinRate}_b^{\mathrm{RKL}}(c)$. We report **Early** and **Final** checkpoints and analyze (i) per-domain, per-bin win rates at each phase and (ii) the change $\Delta\mathrm{WinRate}_b^{\mathrm{FKL}} = \mathrm{WinRate}_b^{\mathrm{FKL}}(\mathrm{Final}) - \mathrm{WinRate}_b^{\mathrm{FKL}}(\mathrm{Early})$ versus average sentence length. All computations use the teacher's tokenizer and teacher-side frequency bins from §3.1.

Table 1: Sentence length statistics across BabyLM test domains. AvgTok = mean tokens per sentence; Med = median; P90/P95 = 90th/95th percentile of sentence length (in tokens, measured with the teacher tokenizer).

Domain	#Sent	AvgTok	Med	P90	P95
childes	538,980	5.01	4	9	11
open_subtitles	344,632	5.66	5	11	14
bnc_spoken	89,904	10.37	6	24	35
gutenberg	54,954	43.76	28	95	140
simple_wiki	49,580	26.22	12	68	90
switchboard	20,000	8.36	6	18	24

Results. Figure 2 shows strong correlations: *FKL* gains more advantage as sentence length increases, and this is true for all three frequency bins. Long-form domains (*Gutenberg*, *SimpleWiki*) favor FKL improvements across frequency bins, while short conversational domains (*CHILDES*, *OpenSubtitles*) favor RKL (Figure 3).

In other words, sentence length robustly moderates which direction improves faster: FKL gains more in long-form domains, RKL gains more in short-utterance domains.

4.3 BLiMP: Critical-Token Probes

We use the *global view* in §3.1 (corpus-level teacher frequency ranks on the teacher token axis: Head=Top-1k, Mid=1k–10k, Tail>10k). We probe two representative BLiMP paradigms with clear token-level contrasts: `relational_noun` and `irregular_past_participle_verb`. They are chosen based on the learners' final-checkpoint accuracy on these two paradigms: on `relational_noun`, RKL shows substantially higher accuracy compared to FKL (64.4 vs. 31.1), while on `irregular_past_participle_verb`, both learners exhibit similar accuracy (80.2 and 82.6), respectively.

Additionally, we identify *critical tokens* – i.e., tokens which distinguish grammatical sentences from ungrammatical ones in BLiMP minimal pairs. These tokens are important from the evaluation perspective, because a learner's accuracy on a BLiMP paradigm largely depends on its ability to correctly predict such critical tokens. Critical tokens are obtained by minimal-pair alignment (see §C.2) and mapped to the 16k teacher tokenizer.

At each (critical or non-critical) position t we compute the per-position absolute-error kernel and

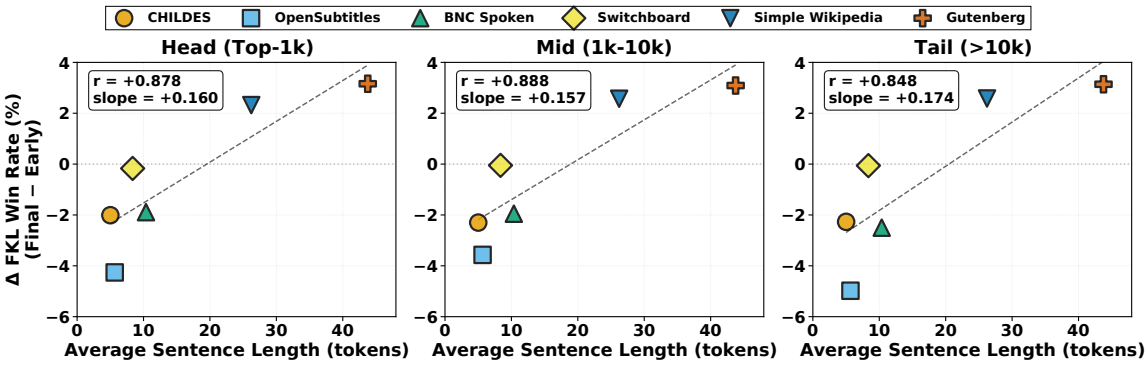

Figure 2: Sentence-length correlation. FKL win-rate change by frequency bins vs. average sentence length across domains. From left to right: **Head** (Top-1k), **Mid** (1k–10k), **Tail** (>10k). Points are domains; dashed lines are least-squares fits. Δ FKL Win Rate denotes the change from Early to Final. RKL exhibits the complementary trend (not shown).

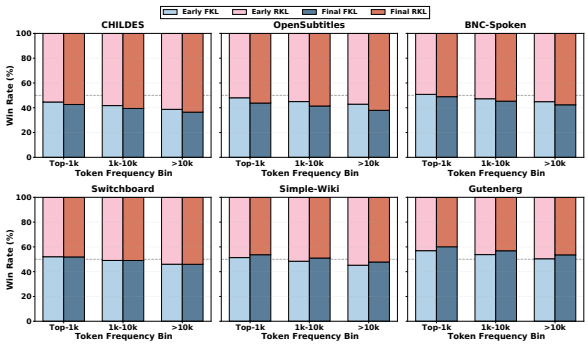

Figure 3: Within-domain frequency bins. FKL win rates by teacher frequency bin (Top-1k / 1k–10k / >10k), Early vs. Final.

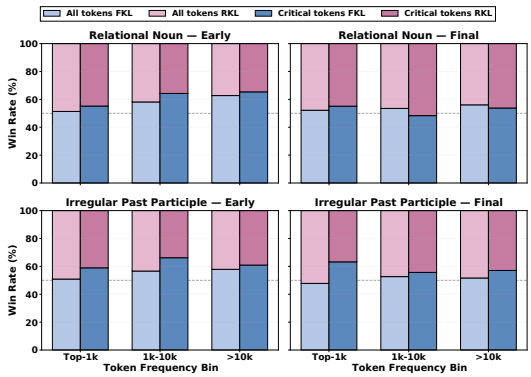

Figure 4: BLiMP frequency bins: all vs. critical tokens. Critical tokens magnify early head–tail differences.

win indicators exactly as in §3.1, Eqs. (9)–(10), and aggregate bin-level win rates with Eq. (11). We report **Early/Final** phase summaries (§4.1) and contrast *critical* versus *all-token* aggregates within each frequency bin.

Results. Figure 4 shows that FKL dominates in most bins at both checkpoints. At **Early**, all three bins (Top-1k / 1k–10k / >10k) are above 50% for FKL in both paradigms, with even larger FKL shares when restricting to *critical-only* positions relative to the all-token baseline. From **Early** to **Final**, bin-wise gaps shrink overall. The **Mid** bin (1k–10k) exhibits the largest change: in `relational_noun`, Mid shifts toward RKL by the Final checkpoint, whereas in `irregular_past_participle`, FKL remains ahead across bins and phases. In the **Tail** bin (>10k) we observe catch-up without overtake: RKL narrows the gap (e.g., `relational_noun` critical-only FKL ∼65% → ∼52%; `irregular_pp` all-token FKL ∼58–59% → ∼51%), but an RKL advantage does not appear in any panel. These trends are consistent with the critical-token inventory (Appendix C): only 11.9% of loci fall beyond 10k, while 64.5% lie in the Mid range, where most phase dynamics occur.

4.4 Phase Dynamics with Teacher-Side Head/Tail Masks

We now adopt the *local view* of §3.1, i.e., teacher-side cumulative-probability masks $H_t(\alpha), T_t(\alpha)$ on the teacher token axis. Unlike the *global view* in §4.2–§4.3 (based on corpus frequency), the local view conditions on the position-specific teacher distribution p_t. This method isolates how quickly each objective fixes where the teacher places probability mass (head vs. tail) and is therefore the right lens for studying early-phase learning dynamics and the gradient sensitivities of FKL and RKL.

Following §3.3, we compute: (i) per-mass KL on Head vs. Tail, $\mathrm{PM}^{\mathrm{FKL}}_{H/T}(t)$; (ii) Head Deficit (HD) and Tail Leakage (TL) to quantify total-mass misallocation and (iii) *asymmetry indices* A^{FKL} and

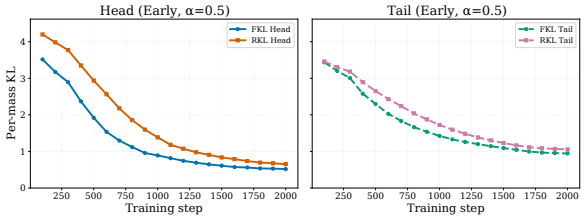

Figure 5: Per-mass KL across training phases (BLiMP–`relational_noun`, CPM α=0.5). FKL aligns head mass faster; RKL reduces tail KL earlier; both converge near Final.

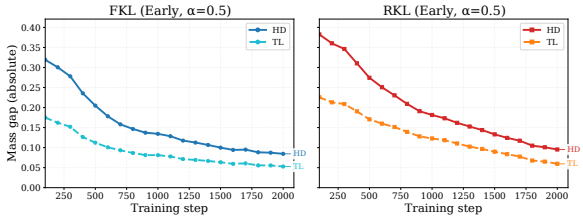

Figure 6: Head-Deficit and Tail-Leakage (BLiMP–`relational_noun`, CPM α=0.5). Early FKL overspreads (larger TL), while RKL starts conservative on head (larger HD); both are corrected over time.

A^{RKL} that summarize head–tail error concentration. We save *dense early checkpoints* (50–2000 steps) to resolve very-early dynamics. Comprehensive $\alpha \in \{0.8, 0.9, 0.95\}$ experiments are provided in the Appendix D. Results are discussed below.

Per-mass KL trajectories. Figure 5 shows a clear split: FKL reduces *Head* per-mass KL faster, while RKL suppresses *Tail* per-mass KL sooner; the gaps then narrow as training proceeds, indicating *transient* direction-specific effects. This matches the kernel-level gradient intuition (§3.3): FKL's $-p/q$ emphasizes correcting missing teacher mass (head-first), whereas RKL's $\log(q/p)+1$ penalizes off-support allocation more steeply (tail-denoising first).

Mass misallocation (HD/TL). Figure 6 complements this result: early FKL exhibits larger TL (overspread into the tail), while early RKL shows larger HD (head-conservative); both artifacts shrink over time. Consistent with mass conservation, the HD and TL curves are numerically aligned , confirming that per-mass trajectories in the previous paragraph are mirrored by total-mass reallocation.

Asymmetry indices. Figure 7 summarizes slice contributions over time: FKL error concentrates more in the *head* (positive A^{FKL}) and RKL error concentrates more in the *tail* (positive A^{RKL}) within the *very-early* window; both indices decay toward zero later. Per-mass asymmetry shows smaller amplitudes than the raw version, indicating that normalizing by slice mass mitigates slice-size confounds and yields a more stable cross-setting comparison. Across Figures 5–7, early differences arise because the FKL gradient $-p/q$ emphasizes correcting missing teacher mass (favoring high-p_t tokens), whereas RKL's $\log(q/p)+1$ term more steeply penalizes off-support allocation. Thus, FKL accelerates head alignment and RKL

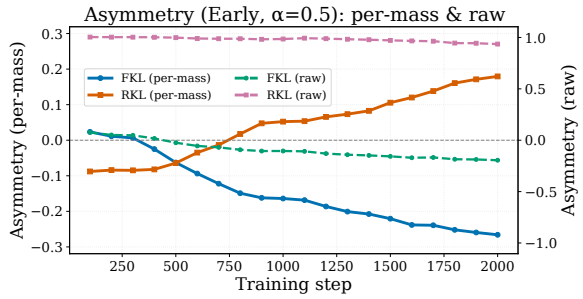

Figure 7: Asymmetry indices over time (BLiMP–`relational_noun`, CPM α=0.5). Direction biases peak early and decay toward zero, indicating transient effects.

cleans up residual tail noise. By the final phase, both approach near-equilibrium, suggesting that KL direction primarily affects *learning dynamics*, not final optima.

5 Conclusion

In this paper, we observed that the magnitude and persistence of direction-specific effects vary with corpus properties such as sentence length: longer, narrative domains favor early FKL gains, while shorter, dialogue-like domains favor RKL. To quantify these dynamics, we introduced teacher-side head–tail partitions (global frequency bins and local nucleus masks) and slice-aware diagnostics—per-mass KL, head deficit / tail leakage, and asymmetry indices—that isolate region-specific alignment behavior. Empirically, FKL and RKL differ most in the early phase: FKL aligns faster on high-probability head tokens, whereas RKL suppresses low-probability tail errors sooner, with the gap diminishing as training converges. These results empirically confirm the theoretical prediction of Wu et al. (2025) that both divergences converge to the same optimum but emphasize different probability regions early in training.

6 Limitations

Results are based on a single teacher–student pair trained on a 10M-token English corpus, so behavior may differ for larger, multilingual, or instruction-tuned models. Analyses rely on the teacher tokenizer and specific head–tail partitions (global frequency bins, local thresholds), and magnitudes may shift under alternative tokenizations or thresholds. Our diagnostics measure distributional alignment rather than downstream utility, and early-phase patterns can depend on the training schedule; the observed moderation by sentence length is correlational.

References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. 2024. On-policy distillation of language models: Learning from self-generated mistakes. In *The twelfth international conference on learning representations*.

Shun-ichi Amari and Andrzej Cichocki. 2010. Information geometry of divergence functions. *Bulletin of the polish academy of sciences. Technical sciences*, 58(1):183–195.

Suraj Anand, Michael A Lepori, Jack Merullo, and Ellie Pavlick. 2024. Dual process learning: Controlling use of in-context vs. in-weights strategies with weight forgetting. *arXiv preprint arXiv:2406.00053*.

Christopher M Bishop and Nasser M Nasrabadi. 2006. *Pattern recognition and machine learning*, volume 4. Springer.

Leshem Choshen, Ryan Cotterell, Michael Y Hu, Tal Linzen, Aaron Mueller, Candace Ross, Alex Warstadt, Ethan Wilcox, Adina Williams, and Chengxu Zhuang. 2024. [call for papers] the 2nd babylm challenge: Sample-efficient pretraining on a developmentally plausible corpus. *arXiv preprint arXiv:2404.06214*.

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

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.

Michael Y Hu, Aaron Mueller, Candace Ross, Adina Williams, Tal Linzen, Chengxu Zhuang, Ryan Cotterell, Leshem Choshen, Alex Warstadt, and Ethan Gotlieb Wilcox. 2024. Findings of the second babylm challenge: Sample-efficient pretraining on developmentally plausible corpora. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 1–21.

Seongryong Jung, Suwan Yoon, DongGeon Kim, and Hwanhee Lee. 2025. Todi: Token-wise distillation via fine-grained divergence control. *arXiv preprint arXiv:2505.16297*.

Kevin P Murphy. 2012. *Machine learning: a probabilistic perspective*. MIT press.

Aakarsh Nair, Alina Hancharova, Mayank Kumar, and Ali Gharaee. 2024. Babylm challenge: Experimenting with self-distillation and reverse-distillation for language model pre-training on constrained datasets. In *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, pages 28–36.

Steven T Piantadosi. 2014. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, 21(5):1112–1130.

Shuntaro Takahashi and Kumiko Tanaka-Ishii. 2017. Do neural nets learn statistical laws behind natural language? *PloS one*, 12(12):e0189326.

Jean-Loup Tastet and Inar Timiryasov. 2024. Babyllama-2: Ensemble-distilled models consistently outperform teachers with limited data. *arXiv preprint arXiv:2409.17312*.

Inar Timiryasov and Jean-Loup Tastet. 2023. Baby llama: knowledge distillation from an ensemble of teachers trained on a small dataset with no performance penalty. *arXiv preprint arXiv:2308.02019*.

Guanghui Wang, Zhiyong Yang, Zitai Wang, Shi Wang, Qianqian Xu, and Qingming Huang. 2025. Abkd: Pursuing a proper allocation of the probability mass in knowledge distillation via α-β-Divergence. *arXiv preprint arXiv:2505.04560*.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 2023. Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, pages 1–34, Singapore. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Taiqiang Wu, Chaofan Tao, Jiahao Wang, Runming Yang, Zhe Zhao, and Ngai Wong. 2025. Rethinking Kullback-Leibler divergence in knowledge distillation for large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5737–5755, Abu Dhabi, UAE. Association for Computational Linguistics.

A License and Data Usage

All datasets (BabyLM and BLiMP) are released under research-friendly CC-BY or equivalent licenses and were used strictly for academic research. The BabyLLaMA and LLaMA models were used under Meta's research license; no redistribution was performed. All experiments comply with dataset terms of use, and no personally identifiable or offensive content was included.

B AI Assistance Disclosure

We used ChatGPT to assist with LaTeX formatting and data-to-table conversion. Specifically, AI assistance was limited to typesetting and layout tasks, including: (1) formatting mathematical equations in Section 3, (2) arranging numerical data into Table 1 in Section 4 (Experiments), and (3) formatting several tables in the Appendix. No text generation, analysis, or interpretation was performed by AI.

C Critical Token Analysis

C.1 Executive Summary

Among 1,730 BLiMP critical tokens, **90.8% (1,571)** are single-subword and **9.2% (159)** span multiple subwords. The median frequency rank is 2,064, with 64.5% concentrated in the mid-frequency range (1k–10k). *relational_noun* shows more multi-subword tokens (12.7%) and higher-frequency targets, whereas *irregular_past_participle* focuses on single-subword, lower-frequency morphological variants.

C.2 Critical Token Identification and Mapping

Critical Token Identification For each BLiMP pair, we find the minimal character span where the two sentences diverge via longest common prefix/suffix alignment, and map it to token indices with the *teacher* tokenizer (16K BPE), yielding a single position or a contiguous subword span. If the locus covers multiple subwords $y_{t:t+n-1}$, we evaluate the *joint* probability under autoregressive factorization and take the absolute error:

Bin	Count	Percentage
Top-1k (high frequency)	408	**23.6%**
1k–10k (mid frequency)	1,116	**64.5%**
>10k (low frequency)	206	**11.9%**

Table 2: Frequency-bin distribution of critical tokens.

Category	Count	Percentage
Single-subword (1 subword)	1,571	**90.8%**
Multi-subword (2+ subwords)	159	**9.2%**

Table 3: Subword composition of BLiMP critical tokens.

$$
\begin{aligned}
P_{\text{crit}}^{(\mathrm{T})}(t) &= \prod_{j=0}^{n-1} p_{t+j}\big(y_{t+j} \mid y_{t:t+j-1}\big), \\
Q_{\text{crit}}^{(m)}(t) &= \prod_{j=0}^{n-1} q_{t+j}^{(m)}\big(y_{t+j} \mid y_{t:t+j-1}\big), \\
e_{\text{crit}}^{(m)}(t) &= \big|\, P_{\text{crit}}^{(\mathrm{T})}(t) - Q_{\text{crit}}^{(m)}(t) \,\big|.
\end{aligned}
\tag{19}
$$

We attribute each span to the frequency bin of its onset token y_t (teacher rank $r(y_t)$).

C.3 Core Tables

Note. Multi-subword loci predominantly include contractions (e.g., *doesn't*, *hasn't*) and polysyllabic verbs subject to tokenizer segmentation.
Interpretation. The mid-frequency concentration (1k–10k) explains why most head–tail dynamics in the results concentrate in the Mid bin; the smaller Tail share supports the "catch-up without overtake" pattern.

C.4 Descriptive Text (formerly separate tables)

Rank statistics. The rank distribution is long-tailed: mean 3,796.2; median 2,064.0; min/max 60/15,801; P25/P75 = 1,025.0/5,911.0; P90/P95 = 10,734.0/10,915.0; P99 = 15,801.0.

Subtask breakdown. *relational_noun*: 12.7% multi-subword; median rank 1,801; Top-1k/1k–10k/>10k = 31.9%/59.1%/9.0%. *irregular_past_participle*: 93.7% single-subword; median rank 4,416; Top-1k/1k–10k/>10k = 16.7%/69.0%/14.3%.

Single vs. multi-subword Multi-subword loci have a higher median rank (5,288 vs. 2,064) but lower P90/P95 (7,594/7,823 vs. 10,734/10,915),

Table 4: BLiMP critical token statistics (overall and by subtask).

Metric	Overall	Rel. Noun	Irreg. Past
Total tokens	1,730	788	942
Single-subword (%)	90.8	87.3	93.7
Multi-subword (%)	9.2	12.7	6.3
Median freq. rank	2,064	1,801	4,416
Top-1k (%)	23.6	31.9	16.7
1k–10k (%)	64.5	59.1	69.0
>10k (%)	11.9	9.0	14.3

reflecting a mix of very common contractions and rare polysyllabic words.

Most common / rare tokens Common items are auxiliaries and contractions (e.g., `do` appears 77 times, rank 221). At the rare end are morphological variants and tokenizer fragments (e.g., `woke` at rank 15,801).

Multi-subword examples Examples include `don't` → [`don`, `'t`] (rank 502), `hasn't` → [`hasn`, `'t`], and polysyllabic verbs such as `chuckle` → [`chuck`, `le`] (rank 6,643).

D Mask Configuration Sensitivity Analysis

In the main text we presented results using the CPM mask with $\alpha = 0.5$ (50% cumulative probability mass) as the boundary between Head and Tail. Here we systematically vary the mask definition to assess the robustness of our findings. We consider two mask families:

- **Cumulative Probability Mass (CPM):** $\alpha \in \{0.5, 0.8, 0.9, 0.95\}$. Tokens are sorted by teacher probability in descending order at each position; the Head comprises the smallest set whose cumulative probability $\geq \alpha$.
- **Top-K:** $k \in \{1, 3, 5, 10, 100, 1000\}$. The Head is simply the top-k tokens by teacher probability; the Tail is everything else.

For each mask configuration, we computed all AKL metrics (per-mass KL, head deficit / tail leakage, asymmetry indices) on the same two BLiMP paradigms (`relational_noun` and `irregular_past_participle_verbs`) throughout training. We aggregate final-checkpoint metrics and visualize trends to understand:

1. How mask size affects measured errors;
2. Whether FKL/RKL behavioral differences persist across definitions;
3. Which metrics are stable vs. sensitive to boundary choice.

D.1 Per-Mass KL Trends

CPM configurations (Figure 8). As α increases from 0.5 to 0.95, the Head region grows (capturing more of the teacher's probability mass) and the Tail shrinks. We observe:

- **Head per-mass KL** (*left panels*): Both FKL and RKL Head errors *decrease* monotonically with α. This is expected: a larger Head slice dilutes extreme probabilities with more moderate ones, reducing the conditional shape mismatch. Critically, *RKL Head error remains systematically lower than FKL Head error* across all α, confirming the mode-seeking behavior is robust to mask definition.
- **Tail per-mass KL** (*right panels*): Errors *increase* with α because the Tail becomes smaller and concentrates the hardest, lowest-probability tokens. Again, *FKL Tail error is consistently lower than RKL Tail error* across all α, reaffirming that FKL is more tolerant of tail under-modeling.
- **Task consistency**: Both paradigms show the same qualitative pattern, though `relational_noun` exhibits slightly higher absolute errors, consistent with main-text frequency-bin results.

D.2 Head Deficit & Tail Leakage Trends

These metrics quantify total probability mass misallocation (independent of within-slice shape). Recall that HD and TL are numerically equal by mass conservation; we report both for interpretability.

Absolute HD/TL (Figures 9.

- **CPM**: As α increases, the Head region grows and the student's shortfall (HD) *decreases* for both FKL and RKL. At $\alpha = 0.5$ (roughly half the mass in Head), HD ≈ 0.02–0.04; at $\alpha = 0.95$, HD drops below 0.01. This suggests that students have little trouble matching the teacher's aggregate allocation to large, high-mass regions, but struggle with fine-grained boundaries.

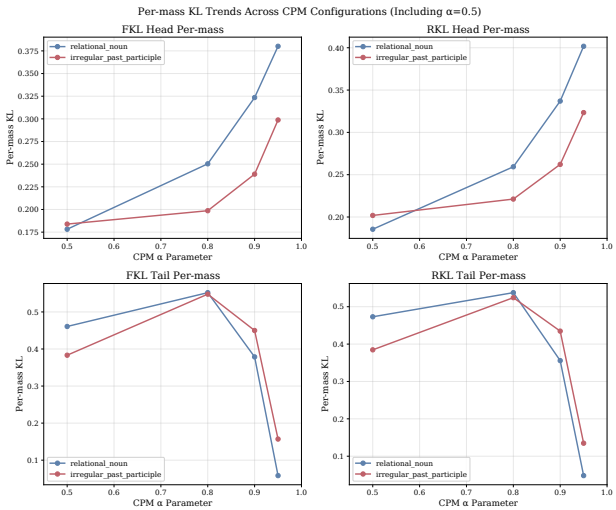

Figure 8: Per-mass KL vs. CPM α. Four panels show FKL/RKL errors in Head/Tail for both tasks. Head errors decrease with α; Tail errors increase. The FKL $>$ RKL gap in Head and RKL $>$ FKL gap in Tail are stable across all α.

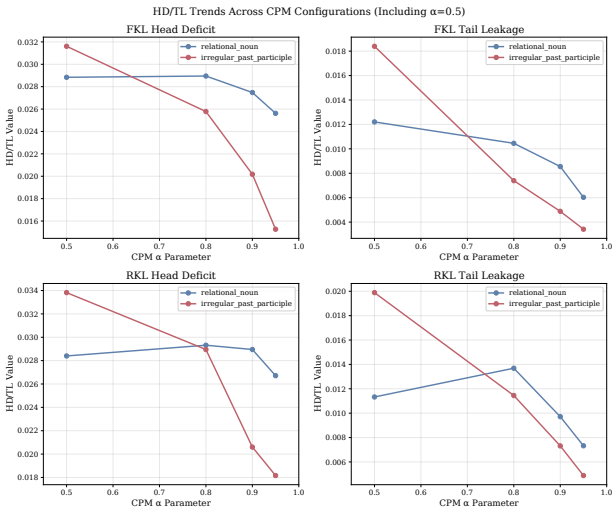

Figure 9: Absolute HD/TL vs. CPM α. Both FKL and RKL show decreasing deficits as the Head grows. FKL $\approx$ RKL within each task, indicating mass allocation is objective-neutral.

- **FKL vs. RKL**: In CPM masks, FKL and RKL HD values are nearly identical, indicating that mass allocation errors are *objective-agnostic*—both students place similar total mass on the Head, even though their conditional shapes differ. This decoupling of mass vs. shape is a key insight: the per-mass KL differences arise from *how* probability is distributed within Head/Tail, not *how much* is allocated.

Relative HD/TL (Figures 10. Normalizing by $P(H)$ or $P(T)$ yields scale-free views:

- **CPM**: Relative HD increases slightly with α (since $P(H) \to 1$ and any shortfall becomes a larger fraction). Relative TL decreases with α (the Tail shrinks, so even a fixed leakage is a smaller fraction of the tiny tail mass).

- The FKL $\approx$ RKL equality holds in relative terms as well, reinforcing that quality mismatch is objective-neutral.

D.3 Asymmetry Index Trends

The asymmetry index quantifies whether KL error concentrates in Head or Tail: $A^{\text{FKL}} = (\text{FKL}_H - \text{FKL}_T)/(\text{FKL}_H + \text{FKL}_T)$, and analogously for RKL. Positive values indicate Head-dominant error; negative values indicate Tail-dominant error.

CPM configurations (Figure 11).

- **FKL asymmetry**: Strongly positive at all α (0.4–0.6 for `relational_noun`, 0.3–0.5 for `irregular_past_participle`), confirming FKL error is Head-concentrated. Asymmetry *increases* slightly with α: as the Head grows to dominate the mass, the Head vs. Tail imbalance becomes more pronounced.

- **RKL asymmetry**: Negative at all α (-0.2 to -0.4), confirming RKL error is Tail-concentrated. Asymmetry becomes *more negative* (i.e., more Tail-skewed) as α increases, because the shrinking Tail concentrates RKL's penalty on under-modeled low-probability tokens.

- **Crossover stability**: The FKL/RKL sign flip is consistent across all α, demonstrating that the mode-seeking/mean-seeking dichotomy is a *robust property of the objective functions*, not an artifact of boundary choice.

D.4 Summary and Implications

Our mask-sensitivity analysis reveals:

1. **Robustness**: The qualitative FKL/RKL differences (mode-seeking vs. mean-seeking) are *stable across all mask definitions*. Whether we use CPM $\alpha \in [0.5, 0.95]$, FKL consistently shows higher Head per-mass error and positive asymmetry, while RKL shows higher Tail per-mass error and negative asymmetry.

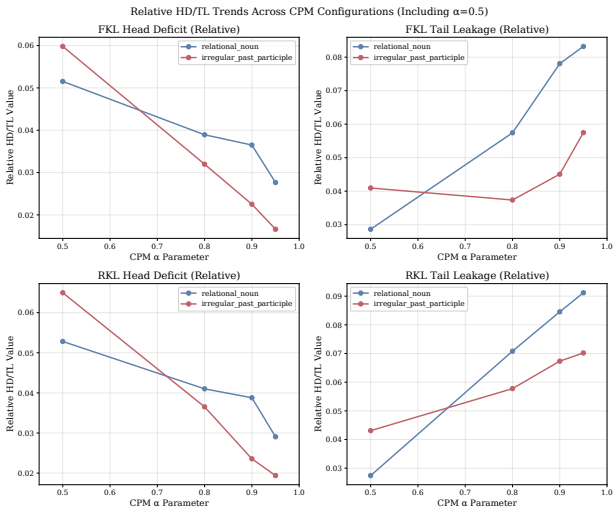

Figure 10: Relative HD/TL vs. CPM α. Scale-free view: relative HD increases with α, relative TL decreases. FKL $\approx$ RKL.

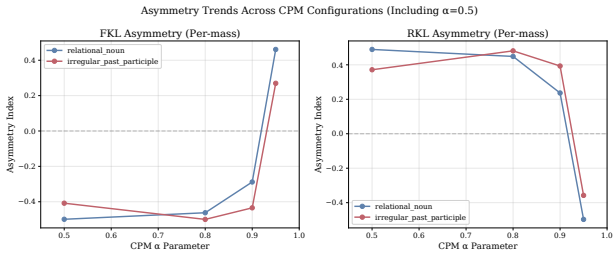

Figure 11: Asymmetry index (per-mass) vs. CPM α. FKL is always positive (Head-skewed); RKL is always negative (Tail-skewed). Magnitude increases with α for both.

2. **Quantitative trends**: Per-mass errors scale predictably with slice size (Head errors decrease as the Head grows; Tail errors increase). Mass allocation errors (HD/TL) are nearly objective-neutral, confirming that FKL/RKL differ in *shape fitting within slices*, not in total mass placement.

3. **Practical choice**: $\alpha = 0.5$ (used in main text) sits in the middle of the CPM range and balances Head/Tail sizes, making asymmetry differences most interpretable. The trends show no pathologies at this choice, and conclusions generalize to other α.

In sum, our main-text findings—that FKL prioritizes high-probability tokens (mode-seeking) while RKL spreads effort across the tail (mean-seeking)—are *not* artifacts of a single mask definition, but reflect fundamental properties of the divergence objectives. The mask trends provide a rich sensitivity analysis that strengthens confidence in our conclusions.