WHEN BETTER PREDICTION MAKES WORSE GENERATION: TEACHER-DOMAIN ALIGNMENT IN KNOWLEDGE DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Teacher-domain alignment can reverse whether knowledge distillation improves generation, even when distillation improves held-out prediction. We establish this through a controlled intervention: adapting one GPT-2 teacher on PubMed changes its student's embedding-fidelity effect from harmful to helpful while keeping the corpus, student, tokenizer, optimization, and generation protocol fixed. The adapted-minus-unadapted Fréchet improvement replicates across three training seeds (-0.043 ± 0.006, mean $\pm$ SD). Across exactly shared audio, text, and image token spaces, prediction gains coexist with different generation outcomes: distant teachers degrade audio fidelity or increase text repetition, whereas aligned teachers improve fidelity. A web teacher also worsens protein fidelity at every positive weight tested. We introduce **Domain-Anchored KD (DA-KD)**, a supervised training-time gate that retains teacher KL when the observed target token lies in the teacher's top-K. DA-KD improves fidelity over standard KD across GPT-2, Pythia, and SmolLM2, also improves over from-scratch training in the reported text comparisons, and reduces repetition without additional parameters or inference overhead. Aligned and null controls delimit when gating is useful. A two-risk analysis under quadratic surrogates explains how prediction and generation can favor different distillation weights. Experiments fit within a 16 GB GPU budget; code and result summaries are available at `https://anonymous.4open.science/r/domain-aligned-distillation-3894/`.

1 INTRODUCTION

Knowledge distillation transfers a teacher's output distribution to a smaller student, commonly selecting its weight by held-out, teacher-forced likelihood. Generation instead evaluates the student's own rollouts. These targets need not agree (Theis et al., 2016): on a specialized domain, a generic teacher can improve next-token prediction while shifting generated samples away from the target corpus.

We isolate this effect with exactly shared vocabularies and repeatable sweeps, avoiding projection or re-tokenization confounds. Small from-scratch students share discrete audio-codec, text BPE, or ImageGPT-cluster vocabularies with larger pretrained MusicGen, GPT-2, or ImageGPT teachers, enabling index-aligned distillation on one consumer GPU.

Our central claim is that teacher–domain alignment can change the sign of distillation's generation benefit, so prediction gains alone cannot certify generative transfer. Adapting one teacher on a fixed corpus moves the same student from worse to better embedding fidelity than scratch across three training seeds. Figure 1 organizes the regimes. Cross-modal controls establish the phenomenon's breadth; three text families validate DA-KD, which retains teacher supervision where the observed token supports it.

Our contributions:

- Controlled teacher adaptation reverses the fidelity effect while fixing corpus, student, tokenizer, optimization, and evaluation, with consistent endpoint gains across three training seeds (Section 7).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

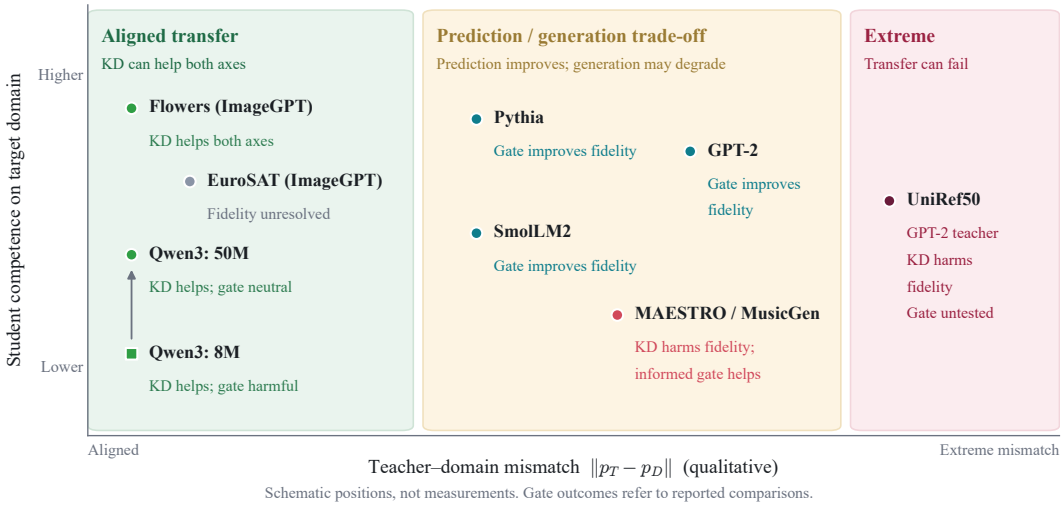

Figure 1: **When does distillation help generation?** Horizontal: teacher–domain distance $\|p_T - p_D\|$; vertical: student competence on the target domain. Aligned transfer can help both axes; intermediate mismatch trades prediction for generation; extreme mismatch can defeat transfer. Qwen's 8M/50M budgets separate student fit from alignment. Positions and boundaries are qualitative, not fitted; Pythia/SmolLM2 are unordered. Labels distinguish gate gains, unresolved effects, and untested cases.

- We establish prediction–generation divergence under exactly shared vocabularies, separating changes in embedding fidelity from changes in repetition (Sections 5.1, 6 and 8).
- We identify *which* supervision-selection criterion matters. Comparing four rules at matched retention – the observed token (DA-KD), the student's own proposal, teacher confidence, and random – separates them on Pythia and on audio even where GPT-2 cannot, and only the observed-token rule transfers (Sections 9, M and O). DA-KD is the minimal form of that rule: one comparison against the teacher's top-K, no extra parameters, forward passes, or inference cost. It improves fidelity over dense KD in three teacher families, beats from-scratch in each, and composes with decoder tuning, while on-policy distillation costs 1.75–2.53$\times$ the training compute for no fidelity gain (Table 15).
- We provide aligned, null, and extreme boundary cases across vision, audio, protein, and modern-scale medical text, with the scope of each stated (Sections 8 and 10).

2 Related Work

Codec language models. Discrete neural codecs (Défossez et al., 2023) enabled autoregressive audio LMs: AudioLM (Borsos et al., 2023), MusicLM (Agostinelli et al., 2023), and MusicGen (Copet et al., 2023). We adopt MusicGen's token space wholesale, which is what makes its decoder an index-aligned teacher (full lineage in Section C).

Knowledge distillation. Soft targets act as both transfer signal and regularizer (Hinton et al., 2015; Menon et al., 2021; Mobahi et al., 2020). Forward KL is mode-covering (Minka, 2005), whereas reverse KL is mode-seeking. Our ablation tests token-level KL direction on teacher-forced prefixes (§5.1), not the on-policy sequence objective of MiniLLM (Gu et al., 2024). Calls to account for distillation's losses beyond aggregate performance (Wang, 2026) motivate broader evaluation. We jointly measure prediction and domain-specific generation fidelity under teacher mismatch. Section C situates the sequence-level, reverse-KL, and teacher-side literature.

Selective distillation. Student-uncertainty selection improves predictive and downstream performance (Tavor et al., 2026). SelecTKD verifies student-proposed tokens against the teacher's top-K and evaluates generated responses (Huang et al., 2025). DA-KD (Section 9) instead tests the *observed domain target token*, studying fidelity under teacher mismatch. We run both criteria under

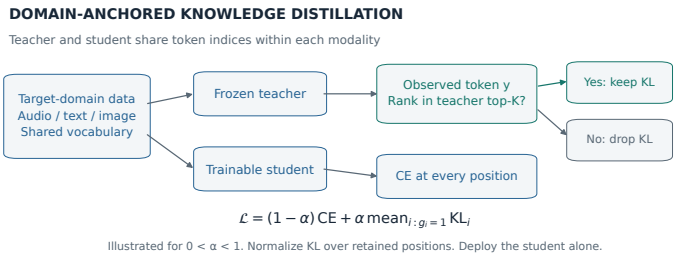

Figure 2: Method overview (modality-agnostic). Three token spaces, audio codec, text BPE, image cluster tokens, share a vocabulary with a larger pretrained teacher, so the frozen teacher's logits $z^{\mathcal{T}}$ are position-aligned with the student's $z^{\mathcal{S}}$ index-for-index: distillation needs no projection or alignment. The objective interpolates data cross-entropy with the teacher KL at weight α; *Domain-Anchored KD* gates the KL where the observed token passes the teacher's top-K test (Section 3.3). We read out two axes, *prediction* (held-out likelihood) and *generation* (primed-sample Fréchet fidelity), whose divergence is the paper's subject.

one protocol: they are indistinguishable on GPT-2, but the student-proposal rule is significantly worse on Pythia (Section O). We also train the on-policy objective rather than only citing it: it reaches DA-KD's repetition but not its fidelity, at 1.75–2.53× the training compute.

Evaluation. We report likelihood and FAD (Kilgour et al., 2019) jointly (CLAP embeddings (Wu et al., 2023), following Gui et al. (2024)); their divergence is the measurement problem studied here.

3 Method

3.1 Token space and index-aligned distillation

A neural audio codec emits Q residual codebook tokens per acoustic frame at rate f. We adopt the EnCodec-32 kHz (Défossez et al., 2023) vocabulary exactly: $Q{=}4$ codebooks of $V_c{=}2048$ codes at $f{=}50$ Hz. Following MusicGen (Copet et al., 2023), codebook k is delayed by k steps. This delay pattern lets all Q tokens of a step be predicted in parallel from summed codebook embeddings, factorizing the joint with within-step conditional independence:

$$d_k[s] = \begin{cases} c_k[s-k] & k \le s < T+k, \\ \langle \mathrm{PAD} \rangle & \text{otherwise,} \end{cases} \qquad p(c) = \prod_s \prod_k p_\theta(d_k[s] \mid d_{<s}) \,. \tag{1}$$

Here $d_{<s}$ contains only earlier delayed steps. Because tokenization, vocabulary, and delay pattern match the pretrained MusicGen-small decoder (297M) exactly, the teacher is an *index-aligned* target: its logits $z^{\mathcal{T}} \in \mathbb{R}^{V_c}$ align index-for-index with the student's $z^{\mathcal{S}}$. No latent projections, re-tokenization, or sequence alignment are needed; extracting the teacher's prior costs one frozen forward pass per batch. The text modality (§6) reuses this structure with the shared GPT-2 BPE vocabulary.

Backbone. We use xLSTM (Beck et al., 2024) solely as a memory-efficient backbone, so complete α sweeps fit one 16 GB GPU; the study is not about the architecture. A parameter-matched Transformer preserves every α-ordering (Table 5); recurrences and systems benchmarks are in Sections H and I.

3.2 The interpolated objective

With $\mathcal{M}$ the positions whose target is a real code (specials masked), the objective interpolates the empirical data distribution (masked cross-entropy) and the teacher's via a temperature-scaled forward KL:

$$\mathcal{L} = (1-\alpha)\,\mathcal{L}_{\mathrm{CE}} + \alpha\,\frac{\tau^2}{|\mathcal{M}|} \sum_{(s,k)\in\mathcal{M}} D_{\mathrm{KL}}\big(\mathrm{softmax}(z^{\mathcal{T}}_{s,k}/\tau) \,\big\|\, \mathrm{softmax}(z^{\mathcal{S}}_{s,k}/\tau)\big) \,. \tag{2}$$

We fix τ=1 and treat $\alpha \in [0, 1]$ as the primary dial (one no-grad teacher pass per batch, +2.1 GB, 1.31× step time). As proven in Appendix B, its unconstrained per-context minimizer is the α-mixture $(1-\alpha)p_D + \alpha p_T$ of the empirical data and the teacher's prior (Proposition 1). Its consequence for our two axes can be stated compactly:

Theorem 1. *Under quadratic prediction and generation risk surrogates around the static mixture target, with positive semidefinite geometries and zero cross-term between sampling noise and fixed teacher bias, the mixture error is $(1-\alpha)\varepsilon + \alpha b$. Writing V_j and B_j for variance and squared bias in geometry $j \in \{P, G\}$, with $V_j + B_j > 0$, the risks have minimizers $\alpha_j^* = V_j/(V_j + B_j)$. Increasing α improves prediction while worsening generation on a nonempty interval exactly when $V_P B_G > V_G B_P$; with zero teacher bias and positive variances, both improve together. The formal assumptions and proof are in Section B.1.*

Theorem 1 isolates a local trade-off under quadratic prediction and feature-error surrogates, consistent with distinct evaluation criteria (Theis et al., 2016). Complementarily, Proposition 2 bounds context drift in the idealized mixture student (cf. Ross et al., 2011): under common prompts and bounded conditional reverse KL, $D_{\mathrm{KL}}(q_{1:T} \| p_{1:T}) \le T\bar{R}_{\mathrm{cond}} + \frac{1}{2}\kappa\alpha\delta T(T-1)$, with teacher–data mismatch δ and conditional-KL bound κ defined there. Exact alignment makes these idealized trajectory laws coincide. The bound allows, but does not require, compounding drift; neither result predicts measured rollout Fréchet or repetition.

3.3 Domain-Anchored KD (DA-KD)

DA-KD is a *supervised, training-time* gate that conditions teacher KL on the observed next token. Standard KD (Equation (2)) applies teacher supervision uniformly, including at teacher-surprised positions where it can transfer domain mismatch. For context $c = (s, k)$, let $\mathcal{A}_K(c)$ contain the teacher's K most probable tokens; DA-KD tests whether $y_{s,k} \in \mathcal{A}_K(c)$ as an observable proxy for local teacher–target agreement (K is distinct from the codebook count Q). Entropy is unchanged by permuting token probabilities, whereas membership of the observed token in $\mathcal{A}_K(c)$ can change, so this label-conditioned test distinguishes agreement from confidence alone. Motivated by the bias–variance account in Section B.1, the gate selects teacher supervision from this one-bit agreement signal and existing logits, without a learned router; write $G = \sum_{(s,k)\in\mathcal{M}} g_{s,k}$:

$$\mathcal{L}_{\text{DA-KD}} = (1-\alpha)\,\mathcal{L}_{\mathrm{CE}} + \alpha\,\frac{\tau^2}{\max\{1, G\}} \sum_{(s,k)\in\mathcal{M}} g_{s,k}\, D_{\mathrm{KL}}\big(\mathrm{softmax}(z^{\mathcal{T}}_{s,k}/\tau) \,\big\|\, \mathrm{softmax}(z^{\mathcal{S}}_{s,k}/\tau)\big)\,,$$

$$g_{s,k} = \mathbf{1}\Big[\,\mathrm{rank}_{p^{\mathcal{T}}_{s,k}}(y_{s,k}) \le K\,\Big]\,. \tag{3}$$

Rank is a competence proxy whose usefulness is tested in Section 9; it does not prove domain mismatch. Rejecting a position removes only KL: CE remains active everywhere for $\alpha < 1$, while rejected positions receive no direct loss at $\alpha = 1$. An empty retained set contributes zero KL. DA-KD uses $\alpha = 0.5$ in the reported comparisons. The gate adds nothing at inference and requires training labels. With an aligned teacher, gating can be neutral or forgo useful transfer (Section 10).

Efficiency. This training-only drop-in reuses standard KD's teacher logits and adds no parameters, model passes, or inference cost beyond the top-K test in Equation (3); matched MAESTRO logs show parity (89.1k vs. 80.5k tokens/s, non-interleaved, so we make no throughput claim). K=10 transfers across three text families, and Table 13 shows a broad K=5–200 plateau.

4 Experimental Setup

Corpora: Audio uses FMA-small (Defferrard et al., 2017) (66.6 h) and MAESTRO v3 (Hawthorne et al., 2019) (199 h solo piano, distant from the teacher). **Text** uses the article field of the PubMed long-document summarization corpus (Cohan et al., 2018). **Vision** uses Oxford Flowers-102 (Nilsback & Zisserman, 2008). Exact splits and track counts are in Section J. **Metrics:** Held-out validation cross-entropy/perplexity, and FAD (CLAP-music) over 200 unconditional 10 s clips per model against held-out real audio; separate primed checks are specified in Section K. **Budget:** 10k steps per cell unless stated. Matched cells share seeds and hyperparameters and train in 2 to 5 hours on one 16 GB GPU (details in Section K).

Table 1: Distillation-weight sweep (headline cells). Best-validation checkpoints, scored on identical held-out tracks per corpus (411 tracks/1.64M tokens for FMA; 58 tracks/7.19M tokens for MAESTRO). FAD: 200 unconditional clips, with finite-sample covariance bias (Section L). The scarce-data (FMA 5 h) and parameter-matched Transformer cells are in Section F and Table 5.

Corpus	Backbone	α	Steps	Val CE ↓	Val PPL ↓	FAD ↓
FMA (66.6 h)	xLSTM	0	10k	5.86	352	0.083
	xLSTM	0.5	10k	5.67	291	0.072
	xLSTM	0.75	10k	5.49	242	0.105
	xLSTM	1	10k	5.52	248	0.117
	xLSTM	1	30k	**5.37**	**216**	0.087
	xLSTM	1 (rev.)	10k	5.85	348	0.155
MAESTRO (199 h)	xLSTM	0	10k	3.12	22.7	**0.003**
	xLSTM	0.5	10k	**2.98**	**19.8**	0.008
	xLSTM	1	10k	3.09	22.1	0.016
Teacher (297M, zero-shot): FMA / MAESTRO				4.31 / 2.88	74 / 17.8	–
Codec floor (EnCodec recon.): FMA / MAESTRO				–	–	0.012 / 0.001

5 Results

§5–6 map alignment regimes; §7 tests alignment causally by adapting one teacher. §8 supplies aligned (Flowers-102), null (EuroSAT), and extreme (UniRef50) controls. §9 validates DA-KD across three text families (Figure 2); audio replicates the divergence, and a wider audio gate significantly improves on dense KD. §10 tests an aligned modern-scale pair.

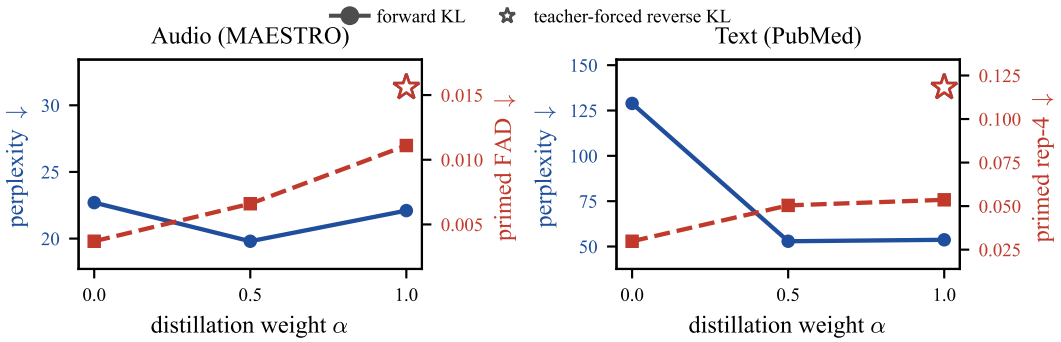

Figure 3: **The prediction–generation trade-off, in two modalities.** Blue (left axis): held-out perplexity. Red (right axis): primed FAD for audio and 4-gram repetition for text, plotted because the powered MiniLM-Fréchet comparison is statistically flat. As the forward-KL weight α rises (0→1), prediction improves while the generation diagnostic worsens relative to scratch, without monotonic prediction gains. Audio uses the separate primed protocol of Table 6. Teacher-forced reverse KL (hollow stars) worsens both axes in these comparisons.

Likelihood: heavy distillation dominates when data is scarce. At the 10k-step budget on 66.6 h, validation PPL falls $352 \to 291 \to 242 \to 248$ for $\alpha = 0, 0.5, 0.75, 1$; the pure-KD run alone is still improving at the budget's end, reaching 216 at 3× budget. On 5 h the contrast is stark: label-only training memorizes the corpus (train CE 0.015 nats) and ends *worse than uniform* on held-out audio (15.8 nats vs. $\ln 2048 = 7.62$), while α=1 holds PPL 526. Soft targets resist memorization, but not unconditionally: with only 5 h of *inputs* even the pure-KD student eventually overfits (validation peak at step 1,500), so regularization comes from input diversity × target softness (mechanism in Section D).

Fidelity degrades across scales. FAD rises with distillation weight (Fig. 3): on MAESTRO, $0.003 \to 0.008 \to 0.016$ for $\alpha = 0, 0.5, 1$. Thus label-only training has the best fidelity al-

though $\alpha = 0.5$ predicts better. The divergence persists across data scales, backbones, and sampling controls (Table 5 and Section E), implicating the objective.

Significance. A powered MAESTRO replication (n=1000 paired continuations) confirms that dense KD worsens FAD relative to scratch (0.00303 vs. 0.00068; $\Delta = +0.00235$, 95% CI $[+0.00161, +0.00309]$). DA-KD at K=64 significantly improves FAD over dense KD (-0.00082, CI $[-0.00143, -0.00020]$) at no perplexity cost, and beats random and anti-confidence controls trained at its own retention count (-0.00112 and -0.00081, both intervals excluding zero), neither of which differs from dense KD. These matched controls support the benefit of informed position selection. The narrower K=16 width is not resolved (Table 12).

Data scale moves the likelihood optimum only. On 199 h of single-domain piano, where even label-only training never overfits within budget, the *likelihood* optimum shifts to α^*=0.5 (PPL 19.8 vs. 22.7 label-only, 22.1 pure-KD) and the benefit compresses (Fig. 6), while the *fidelity* optimum stays pinned at α=0 regardless of scale. Distillation's two components thus separate along orthogonal axes: its regularization value fades with data, while its distributional pull toward the teacher persists.

5.1 Teacher-forced reverse KL does not recover fidelity

Forward KL is mode-covering and reverse KL mode-seeking (Minka, 2005). We test whether changing direction recovers fidelity under the same teacher-forced prefixes. An otherwise identical FMA α=1 cell is worse on *both* axes (PPL 348 vs. 248; FAD 0.155 [95% CI 0.126–0.198] vs. 0.117), landing near from-scratch prediction. Thus direction alone does not repair this mismatch. This is a token-level, off-policy ablation; it does not evaluate on-policy sequence distillation such as MiniLLM.

6 Cross-modal replication: text

We replicate the sweep on 2M PubMed tokens with a randomly initialized GPT-2 small (124M) student and frozen GPT-2 medium (355M) teacher, preserving the $\approx 3\times$ ratio and shared BPE vocabulary. We score held-out perplexity and primed-continuation Fréchet distance in an independent MiniLM space (Wang et al., 2020).

Table 2: Text replication (PubMed, GPT-2 teacher). Prediction replicates strongly: a large perplexity gain with overfitting prevented. Fidelity is Fréchet distance in an independent MiniLM space on *primed continuations*; ↓ better. †The forward-distillation Fréchet cost is within noise at n=1000 (paired bootstrap); the generation cost registers instead as increased repetition. The matching audio cells are in Table 6. Distinct-2 uses the archived run-level primed evaluation, separately from the powered Fréchet evaluation.

α (dir.)	PPL ↓	overfits?	primed-Fréchet ↓	distinct-2 ↑
0 (scratch)	128.9	yes (ep 10)	0.196	0.445
0.5 forward	**52.9**	no	0.193†	0.479
1 forward	53.7	no	0.197†	0.519
1 *reverse*	137.8	no	0.359	0.348

Prediction replicates. Forward distillation cuts perplexity from 128.9 to 52.9 (2.4×) and prevents the overfitting that stops the α=0 baseline at epoch 10; distilled students still improve at the 20-epoch cap.

The generation cost is degeneration. At n=1000 (paired bootstrap) the primed-Fréchet cost of forward distillation is *within noise* (0.196 scratch vs. 0.193 forward, n.s.): the embedding-space drift seen at smaller n is a rank-deficiency artifact and does not survive. The real cost is *degeneration*. Forward distillation significantly raises 4-gram repetition (rep-4 0.030→0.050, Δ=0.021, CI $[0.014, 0.027]$; independent per-run bootstrap) despite higher distinct-2, and teacher-forced reverse KL is worse on *both* axes (PPL 137.8, near from-scratch; rep-4 0.118, by far the worst), exactly as in audio. Repetition, not embedding distance, cleanly registers the cost at scale, and it is precisely this repetition that DA-KD removes (§9; verbatim continuations at each α in Figure 9). (We use

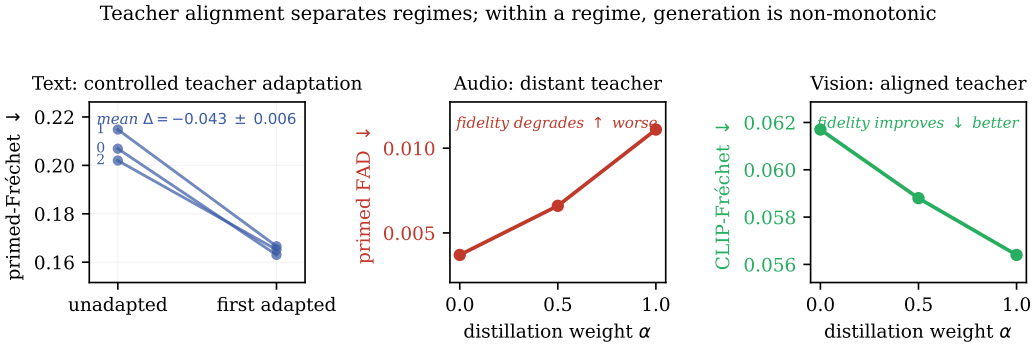

Figure 4: **Controlled alignment and boundary cases.** Left: with corpus and student fixed, adapting one teacher improves all three matched seeds (labels at left endpoints). Center/right: audio and vision show distant and aligned endpoints. Generation within the helpful regime is non-monotonic, supporting a regime change rather than a graded distance law.

primed continuations throughout; App. K explains why unconditional generation collapses at these scales.)

7 Controlled alignment intervention: adapt one teacher

Can the same teacher–student pair cross the harmful/helpful boundary? We fix the PubMed corpus, GPT-2 student, tokenizer, optimization, and generation protocol, using a separate matched run series (Section L), changing only the teacher checkpoint after $0, 50, 125, 250, 1000, 4000,$ or 16000 PubMed adaptation updates. On this fixed corpus, $H(p_D, p_T) = H(p_D) + D_{\mathrm{KL}}(p_D \| p_T)$, so teacher target NLL orders forward-KL mismatch up to the constant data entropy; unlike raw NLL across corpora, this is a controlled distance axis.

Teacher NLL falls monotonically from 3.182 to 2.712 nats/token as adaptation increases, with student perplexity ($53.1 \rightarrow 41.1$). The unadapted teacher harms Fréchet relative to scratch ($\Delta = +0.017$, CI $[+0.010, +0.032]$), while every adapted checkpoint significantly improves it ($\Delta = -0.019$ to -0.034). The adapted checkpoints are not monotonically ordered. Repeating the decisive endpoints at seeds $\{0, 1, 2\}$ gives adapted-minus-unadapted Fréchet effects $-0.044, -0.048, -0.037$ (mean -0.043 ± 0.006, sample SD across training seeds). Within each seed, an n=1000 paired bootstrap gives 95% CIs $[-0.057, -0.034]$, $[-0.060, -0.038]$, and $[-0.051, -0.028]$; all exclude zero. The SD measures training variation; the intervals condition on each trained pair. We make no seed-level significance test. Thus adaptation crosses a harmful/helpful boundary, but target NLL does not calibrate quality within the helpful regime. The intervention establishes that adapting this teacher changes the fidelity effect of distillation; it does not identify target NLL as the sole mediator. Repetition also need not improve in parallel with embedding fidelity, reinforcing the need to evaluate both.

8 Vision: the aligned-teacher control

We test two vision regimes with ImageGPT-small (76M), a frozen ImageGPT-medium teacher (Chen et al., 2020), and their shared 512-cluster ImageNet vocabulary. On the aligned natural-image target Flowers-102 (Nilsback & Zisserman, 2008), increasing $\alpha : 0 \rightarrow 0.5 \rightarrow 1$ improves both perplexity ($13.9 \rightarrow 12.5 \rightarrow 12.5$) and primed CLIP-Fréchet ($0.0617 \rightarrow 0.0588 \rightarrow 0.0564$; Figure 4). On semantically distant but pixel-benign EuroSAT (Helber et al., 2019), KD improves perplexity ($4.2 \rightarrow 4.0$) without a detectable fidelity loss (CLIP-Fréchet 0.0311 vs. 0.0314; paired CI $[-0.0022, 0.0028]$). Together these controls show that transferable low-level token priors, rather than semantic labels alone, determine whether the vision teacher is benign; protocol and samples are in Section G and Figure 12.

Table 3: DA-KD across three teacher families on PubMed (n=1000; ↓ better). Pythia KD/DA-KD rows: mean ± training-seed SD over $\{0, 1, 2\}$. Its random row uses a separate matched evaluation; Δ uses that evaluation's KD reference (Section L), not the seed mean. SmolLM2 rows share one paired bootstrap; both links are significant. GPT-2/Pythia scratch: Tables 2 and 4.

family	teacher→student	objective	primed-Fréchet ↓	vs. KD (Δ)
GPT-2	355M→124M	standard KD	0.192	ref
		uniform KD (α_{eff}=0.36)	0.224	+0.032 worse
		random mask (27%)	0.165	−0.027
		DA-KD (informed)	**0.166**	**−0.027**
Pythia	410M→160M	standard KD	0.204±0.006	ref
		random mask (27%)	0.215	+0.015 *worse*
		DA-KD (informed)	**0.181±0.002**	**−0.025±0.008**
SmolLM2	1.7B→360M	from scratch	0.174	+0.007 *worse*
		standard KD	0.165	ref
		DA-KD (informed)	**0.154**	**−0.011**

The opposite extreme: protein. Distilling a web GPT-2 teacher onto UniRef50 sequences (Suzek et al., 2007) yields α^*=0 (paired-bootstrap 95% CI $[0, 0]$ on the tested grid, n=1000); the positive weights tested worsen ESM-2 Fréchet (Lin et al., 2023) by ∼2.9× (0.12 → 0.34), validating the out-of-domain endpoint.

9 Domain-Anchored KD: gate by domain competence

DA-KD (Section 3.3 and Equation (3)) replaces uniform teacher matching with KL where the true token is in the teacher's top-K, otherwise using ground truth alone. It improves fidelity over standard KD across three teacher families.

DA-KD improves fidelity across three families. (Δ is signed; negative improves a lower-is-better metric.) At n=1000, DA-KD improves primed-Fréchet over standard KD for GPT-2 (Δ=−0.027, CI $[-0.040, -0.019]$), Pythia (−0.025±0.008 across three seeds), and SmolLM2 (Ben Allal et al., 2025) (Δ=−0.011, CI $[-0.017, -0.005]$), and beats from-scratch on GPT-2 (Δ=−0.030, CI $[-0.038, -0.020]$). Each Pythia seed favors DA-KD (−0.028, −0.016, −0.031; all paired CIs exclude zero; Section M). Standard KD also raises rep-4 on GPT-2 (0.030→0.050) and SmolLM2 (0.049→0.055), while DA-KD lowers it to 0.016 and 0.049, respectively. Against scratch, DA-KD is best in every family (GPT-2 0.196→0.166, Pythia 0.194→0.181, SmolLM2 0.174→0.154; for SmolLM2 via two significant paired links, scratch→KD and KD→DA-KD). MAUVE falls under KD (0.92→0.80) and recovers under DA-KD (0.94; Section L). This costs some perplexity (52.9→72.6 on GPT-2; 25.9→26.7 on SmolLM2), controlled by K.

Standard KD falls below scratch on domain-term recall and F1; DA-KD exceeds both (F1 +0.012; Table 14), and MeSH reproduces the ordering without establishing technical-term specificity (Table 16).

The gate and the decoder compose. A tuned decoder (nucleus p=0.95, selected on disjoint data) lets dense KD beat DA-KD-at-default on Fréchet and rep-4, but no decoder lifts its domain-term F1 to DA-KD's (Section Q). With *both* on the tuned decoder, DA-KD still wins in both families: GPT-2 on rep-4 (−0.005) and F1 (+0.010), Pythia on Fréchet (−0.008±0.001) and F1 (+0.006), both Pythia gains holding in all three seeds (Table 18). The choice is not gate *or* decoder: our best cell is DA-KD *with* the tuned decoder (Fréchet 0.152, rep-4 0.009 vs. 0.154, 0.014 for tuned dense KD; Table 17). The absolute deltas are small but the relative effects are not – on GPT-2 the gate cuts repetition by 65% and raises MeSH F1 by 23–31% (Tables 15 and 16) – at no inference cost and no extra training compute.

Blind readers agree. A counterbalanced LLM judge prefers DA-KD over standard KD (61.0%, p=0.004) and scratch (79.3%), and every ordering survives the full 250-prompt sample and a position-aware Bradley–Terry fit (Table 22). A preregistered human panel prefers DA-KD over both baselines, every rater agreeing. It departs from the judge only on KD versus scratch, along reader

competence: the two domain experts reject KD for scratch in 81% of pairs while the lay reader sides with the judge, and the experts prefer DA-KD over both baselines with intervals excluding chance (Table 21).

Confidence is not competence. DA-KD directly tests agreement with the observed token. Random masking and student-proposal verification both match DA-KD on GPT-2 yet are significantly worse on Pythia, the latter on repetition and domain-term F1 (Tables 3 and 4 and Section O). Entropy gating improves fidelity without a detectable repetition reduction, whereas DA-KD improves both; uniform attenuation at $\alpha_{\text{eff}} = 0.36$ likewise fails (GPT-2 Fréchet 0.224 versus 0.192 for dense KD and 0.166 for DA-KD; Table 3). These controls show that confidence alone does not identify teacher–target agreement. Table 4 is seed-0 Pythia; the DA-KD advantage itself replicates across three seeds (Table 3).

Table 4: Pythia selection criteria at matched 27% sparsity on PubMed. Fréchet and rep-4 use n=1000 primed continuations; Δ is paired against standard KD.

objective	ppl ↓	Fréchet ↓	rep-4 ↓	MAUVE ↑	ΔFréchet / Δrep-4
from scratch	88.5	0.194	0.029	0.918	-0.018^* / -0.015^*
standard (dense) KD	**57.5**	0.208	0.043	0.856	(reference)
gates at matched 27% drop					
random (uninformed)	57.5	0.212	0.055	0.868	$+0.004$ / $+0.012^*$ worse
confidence (entropy)	61.4	0.192	0.047	0.886	-0.017^* / $+0.004$ n.s.
competence (DA-KD)	59.3	**0.180**	**0.031**	**0.926**	$\mathbf{-0.030^*}$ / $\mathbf{-0.012^*}$

* paired-bootstrap 95% CI excludes 0.

10 Scaling: alignment, not recency

We test 6× scale with a from-scratch Qwen3-0.6B student and frozen Qwen3-1.7B teacher (Yang et al., 2025) using a *full*-vocabulary KL on the same medical corpus, testing the aligned-teacher limit of Theorem 1(ii).

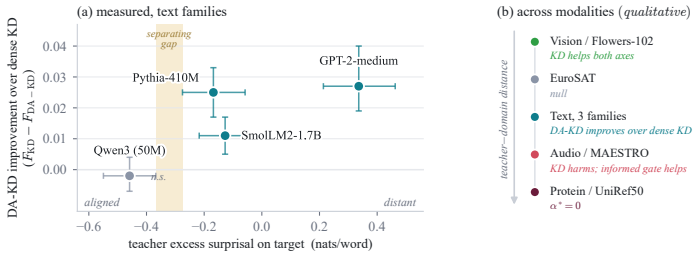

Figure 5: **Teacher–domain alignment separates the observed distillation regimes. (a)** x: teacher excess surprisal over a generic-web reference; y: Fréchet improvement over dense KD. Bars are 95% intervals, Pythia's vertical bar is seed SD (Tables 3 and 11). **(b)** Qualitative cross-modal outcomes.

Two regimes consistent with the framework. Mismatch can limit useful transfer, while alignment favours dense KD and leaves gating only signal to forgo. The pre-registered distance measure placed Qwen3 on the aligned side before either run and predicted the result: at 8M standard KD improves prediction (56.5→32.5) and fidelity (0.225→0.172) while DA-KD is worse (+0.013); at 50M it improves both and DA-KD is neutral (+0.002, n.s.). Qwen3 has the lowest excess surprisal, and same-size SmolLM2-1.7B sits on the trade-off side (Table 11 and Figure 5).

11 Conclusion

Better prediction does not certify better generative transfer. Controlled teacher adaptation reverses fidelity relative to scratch across three seeds. DA-KD improves generation in three text families, composes with decoder tuning, and adds no inference cost, while aligned and extreme controls mark its boundaries. Evaluate prediction, fidelity, repetition and alignment together (Section A).

AI USE STATEMENT

LLMs were used as writing assistant, organization, literature checking, consistency checks. The authors are responsible for reviewing the final text, figures, and claims against the verified code and experiment outputs. Separately, one supplementary experiment uses an LLM as a system-blind judge of generated continuations; its model, prompt, ordering controls, and analysis protocol are disclosed in Section T. The LLM judge is treated as an automated evaluation instrument rather than as human evidence. The authors take responsibility for the final text, claims, code, and experimental conclusions.

REPRODUCIBILITY STATEMENT

Every result fits within a 16 GB GPU budget. Objectives are in Equations (2) and (3), proofs in Sections B and B.1, training and generation settings in Section K, public-corpus splits in Section J, and the $n{=}1000$ paired bootstrap, embedding spaces, and MAUVE check in Section L. The anonymous repository provides code, configs, seeds, and compact result summaries; regenerating samples and checkpoints requires running the released pipeline: `https://anonymous.4open.science/r/domain-aligned-distillation-3894/`.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations (ICLR)*, 2024.

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank. MusicLM: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023. URL https://arxiv.org/abs/2301.11325.

Maximilian Beck, Korbinian Pöppel, Markus Spanring, Andreas Auer, Oleksandra Prudnikova, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. xLSTM: Extended long short-term memory. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL https://arxiv.org/abs/2405.04517.

Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, et al. SmolLM2: When smol goes big – data-centric training of a small language model. *arXiv preprint arXiv:2502.02737*, 2025. URL https://arxiv.org/abs/2502.02737.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning (ICML)*, 2023. URL https://arxiv.org/abs/2304.01373.

Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations (ICLR)*, 2018. URL https://arxiv.org/abs/1801.01401.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. AudioLM: A language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:2523–2533, 2023. doi: 10.1109/TASLP.2023.3288409. URL https://arxiv.org/abs/2209.03143.

Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *International Conference on Machine Learning (ICML)*, 2020. URL https://proceedings.mlr.press/v119/chen20s.html.

Min Jin Chong and David Forsyth. Effectively unbiased FID and inception score and where to find them. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6070–6079, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Chong_Effectively_Unbiased_FID_and_Inception_Score_and_Where_to_Find_CVPR_2020_paper.html.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pp. 615–621. Association for Computational Linguistics, 2018. doi: 10.18653/v1/N18-2097. URL https://aclanthology.org/N18-2097/.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://arxiv.org/abs/2306.05284.

Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. FMA: A dataset for music analysis. In *International Society for Music Information Retrieval Conference (ISMIR)*, 2017. URL https://arxiv.org/abs/1612.01840.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *Transactions on Machine Learning Research (TMLR)*, 2023. URL https://arxiv.org/abs/2210.13438. arXiv:2210.13438.

Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music. *arXiv preprint arXiv:2005.00341*, 2020. URL https://arxiv.org/abs/2005.00341.

Tommaso Furlanello, Zachary C. Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born-again neural networks. In *International Conference on Machine Learning (ICML)*, 2018. URL https://arxiv.org/abs/1805.04770.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *Conference on Language Modeling (COLM)*, 2024. URL https://arxiv.org/abs/2312.00752. arXiv:2312.00752.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2306.08543.

Azalea Gui, Hannes Gamper, Sebastian Braun, and Dimitra Emmanouilidou. Adapting Fréchet Audio Distance for generative music evaluation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024. URL https://arxiv.org/abs/2311.01616.

Zihao Han, Tiangang Zhang, Huaibin Wang, and Yilun Sun. Adaptive teacher exposure for self-distillation in LLM reasoning. *arXiv preprint arXiv:2605.11458*, 2026. URL https://arxiv.org/abs/2605.11458.

Curtis Hawthorne, Andriy Stasyuk, Adam Roberts, Ian Simon, Cheng-Zhi Anna Huang, Sander Dieleman, Erich Elsen, Jesse Engel, and Douglas Eck. Enabling factorized piano music modeling and generation with the MAESTRO dataset. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://arxiv.org/abs/1810.12247.

Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. EuroSAT: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019. doi: 10.1109/JSTARS.2019.2918242. URL https://arxiv.org/abs/1709.00029.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. URL https://arxiv.org/abs/1503.02531. NeurIPS Deep Learning Workshop.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, et al. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL https://arxiv.org/abs/2203.15556.

Haiduo Huang, Jiangcheng Song, Yadong Zhang, and Pengju Ren. SelecTKD: Selective token-weighted knowledge distillation for LLMs. *arXiv preprint arXiv:2510.24021*, 2025. URL https://arxiv.org/abs/2510.24021.

Tao Huang, Yuan Zhang, Shan You, Fei Wang, Chen Qian, Jian Cao, and Chang Xu. Masked distillation with receptive tokens. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2205.14589.

Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi. Fréchet Audio Distance: A reference-free metric for evaluating music enhancement algorithms. In *Interspeech*, 2019. doi: 10.21437/Interspeech.2019-2219.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016. doi: 10.18653/v1/D16-1139. URL https://arxiv.org/abs/1606.07947.

Jahyun Koo, Yerin Hwang, Yongil Kim, Taegwan Kang, Hyunkyung Bae, and Kyomin Jung. SWITCH: Studying with teacher for knowledge distillation of large language models. In *Findings of the Association for Computational Linguistics: NAACL*, 2025. doi: 10.18653/v1/2025.findings-naacl.206. URL https://arxiv.org/abs/2410.19503.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-fidelity audio compression with improved RVQGAN. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL https://arxiv.org/abs/2306.06546.

Hyoje Lee, Yeachan Park, Hyun Seo, and Myungjoo Kang. Self-knowledge distillation via dropout. *Computer Vision and Image Understanding*, 233:103720, 2023. doi: 10.1016/j.cviu.2023.103720. URL https://arxiv.org/abs/2208.05642.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574.

Carolyn E. Lipscomb. Medical subject headings (mesh). *Bulletin of the Medical Library Association*, 88(3):265–266, 2000.

Aditya Krishna Menon, Ankit Singh Rawat, Sashank J. Reddi, Seungyeon Kim, and Sanjiv Kumar. A statistical perspective on distillation. In *International Conference on Machine Learning (ICML)*, 2021. URL https://proceedings.mlr.press/v139/menon21a.html.

Tom Minka. Divergence measures and message passing. Technical Report MSR-TR-2005-173, Microsoft Research, 2005. URL https://www.microsoft.com/en-us/research/publication/divergence-measures-and-message-passing/.

Hossein Mobahi, Mehrdad Farajtabar, and Peter L. Bartlett. Self-distillation amplifies regularization in hilbert space. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. URL https://arxiv.org/abs/2002.05715.

Rafael Müller, Simon Kornblith, and Geoffrey Hinton. When does label smoothing help? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. URL https://arxiv.org/abs/1906.02629.

National Library of Medicine. MeSH 2025 ASCII descriptors. MeSH data archive, 2025. URL https://nlmpubs.nlm.nih.gov/projects/mesh/2025/asciimesh/. Production year 2025, descriptor file d2025.bin.

Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing (ICVGIP)*, 2008. doi: 10.1109/ICVGIP.2008.47.

Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, et al. RWKV: Reinventing RNNs for the transformer era. In *Findings of the Association for Computational Linguistics: EMNLP*, 2023. URL https://arxiv.org/abs/2305.13048.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. MAUVE: Measuring the gap between neural text and human text using divergence frontiers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. URL https://arxiv.org/abs/2102.01454.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI technical report*, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Stéphane Ross, Geoffrey J. Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011. URL https://arxiv.org/abs/1011.0686.

Sentence Transformers. all-MiniLM-L6-v2: Model card. Hugging Face model repository, n.d. URL https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2. Accessed September 23, 2026.

Donghyeok Shin, Yeongmin Kim, Suhyeon Jo, Byeonghu Na, and Il-Chul Moon. AMiD: Knowledge distillation for LLMs with α-mixture assistant distribution. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2510.15982.

Baris E. Suzek, Hongzhan Huang, Peter McGarvey, Raja Mazumder, and Cathy H. Wu. UniRef: Comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, 23(10):1282–1288, 2007. doi: 10.1093/bioinformatics/btm098.

Almog Tavor, Itay Ebenspanger, Neil Cnaan, and Mor Geva. Rethinking selective knowledge distillation. *arXiv preprint arXiv:2602.01395*, 2026. doi: 10.48550/arXiv.2602.01395. URL https://arxiv.org/abs/2602.01395.

Lucas Theis, Aäron van den Oord, and Matthias Bethge. A note on the evaluation of generative models. In *International Conference on Learning Representations (ICLR)*, 2016. URL https://arxiv.org/abs/1511.01844.

Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. URL https://arxiv.org/abs/1711.00937.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Wenshuo Wang. Knowledge distillation must account for what it loses. *arXiv preprint arXiv:2604.25110*, 2026. URL https://arxiv.org/abs/2604.25110.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023. URL https://arxiv.org/abs/2211.06687.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. SoundStream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2021. doi: 10.1109/TASLP.2021.3129994. URL https://arxiv.org/abs/2107.03312.

Anhao Zhao, Haoran Xin, Yingqi Fan, Junlong Tong, Wenjie Li, and Xiaoyu Shen. Decoupling KL and trajectories: A unified perspective for SFT, DAgger, offline RL, and OPD in LLM distillation. *arXiv preprint arXiv:2605.16826*, 2026. URL https://arxiv.org/abs/2605.16826.

A LIMITATIONS

Generalization. The distant-teacher results hold for students up to 360M across three teacher families, with three training seeds behind both the adaptation intervention and the Pythia gate gains. Their boundaries are explicit. The largest student, Qwen3-0.6B, tests a comparatively aligned teacher, so transfer to billion-parameter students and to further domains is untested, and the objective baselines and audio gate controls use one training seed. On audio, $K{=}64$ improves over dense KD and over both matched-retention controls, while $K{=}16$ is unresolved and training from scratch still attains the best FAD (Table 12). The aligned and null controls are included to map where gating helps rather than to promise that it always does.

Measurement. Generation is measured with embedding-Fréchet, repetition and lexical overlap, each with paired bootstrap intervals and at least one corroborating instrument (MAUVE, MeSH, an LLM judge, human readers). Those quantities capture selected aspects of generation, not factual accuracy or downstream utility, and two comparisons stay open within them: GPT-2's dense-KD-versus-scratch Fréchet contrast is unresolved, and the MeSH check reproduces the model ordering without establishing that the advantage is specific to technical terminology (Sections L and P). The human panel reproduces the automatic ordering on both DA-KD comparisons, and its domain-expert interval for DA-KD over from-scratch excludes chance, but the panel is small – four raters, eight to sixteen items per comparison – and the pooled interval for that comparison includes chance (Section S). No human audio listening study is included. Primed results concern fixed prompts and horizons and do not establish long-range unconditional quality, and standard KD's lower perplexity can remain preferable for prediction-focused uses.

Comparison scope. DA-KD's gains over dense KD persist when both objectives use a per-family tuned decoder, with the improved metric differing by family (Table 18); contrastive decoding is untested (Section Q). The student-proposal gate is a controlled ablation of the verification criterion rather than a full SelecTKD reproduction. The on-policy comparison covers two mixing weights with fixed rollout settings and no objective tuning, unlike the decoder baseline, which was tuned on our own headline metric; those runs therefore characterize the configurations we trained and do not settle the method comparison (Section O). Reported step timings quantify local costs; we do not provide an aggregate FLOP estimate across the heterogeneous training and evaluation pipeline.

B WHY THE DIAL WORKS: A CLOSED-FORM ACCOUNT

The empirical exchange has a simple explanation at the level of the per-position target distribution. Fix a context; let p_D be the (empirical) data next-token distribution over the vocabulary, p_T the teacher's, and p_S the student's. The per-position objective (Eq. 2) is

$$\mathcal{L}(p_S) = (1-\alpha)\,\mathrm{CE}(p_D, p_S) + \alpha\, D_{\mathrm{KL}}(p_T \,\|\, p_S), \qquad \mathrm{CE}(p_D, p_S) = -\sum_v p_D(v) \log p_S(v). \quad (4)$$

Proposition 1 (The dial is a mixture)**.** *The unique minimizer of $\mathcal{L}$ over the probability simplex is the α-mixture of data and teacher,*

$$p_S^\star \;=\; (1-\alpha)\, p_D \;+\; \alpha\, p_T. \quad (5)$$

Proof. Collecting the p_S-dependent terms, $-(1-\alpha)\sum_v p_D(v)\log p_S(v) - \alpha\sum_v p_T(v)\log p_S(v) = -\sum_v q_\alpha(v)\log p_S(v)$ up to a constant, where $q_\alpha := (1-\alpha)p_D + \alpha p_T$ is a valid distribution (a convex combination). This is $\mathrm{CE}(q_\alpha, p_S)$, minimized over the simplex at $p_S = q_\alpha$ by Gibbs' inequality. □

The proposition assumes temperature one and independent optimization of each conditional distribution. A neural student's shared parameters, finite capacity, and optimization error can prevent it from attaining this target. The following consequences explain the two evaluation axes at the level of that idealized target.

Prediction (Corollary 1). Let p^* be the true next-token distribution. On a small corpus the empirical p_D is a high-variance estimate of p^* (it overfits), while the teacher p_T is a smooth, low-variance, if biased, estimate. $p_S^\star$ is a convex combination, so choosing $\alpha > 0$ trades the variance of p_D

for the bias of p_T. The local quadratic model in Section B.1 makes the conditions for reducing predictive risk explicit. This is the regularization view of distillation and label smoothing (Müller et al., 2019; Mobahi et al., 2020; Menon et al., 2021), consistent with our observations: distillation improves held-out likelihood and suppresses overfitting, and the optimal α *decreases* as data grows (as $p_D \to p^*$, the variance-reduction benefit vanishes), our data-scale result (Fig. 6).

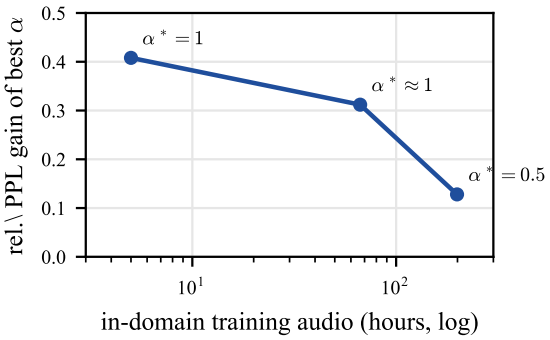

Figure 6: Distillation's likelihood benefit fades with data: the relative perplexity improvement of the best distillation weight over label-only training shrinks from 41% (5 h) to 13% (199 h), and the optimal weight itself decreases. Consistent with distillation acting as prior transfer *plus* a regularizer whose value vanishes when overfitting pressure does.

Generation (Corollary 2). The generator samples from $p_S^\star$. Its distance to the corpus distribution p_D is, in any norm,

$$\|p_S^\star - p_D\| \;=\; \|(1-\alpha)p_D + \alpha p_T - p_D\| \;=\; \alpha\,\|p_T - p_D\|. \tag{6}$$

This *conditional* distance to the empirical target grows linearly in α, with slope $\|p_T - p_D\|$. It identifies the distributional shift introduced by teacher supervision. It does not imply linear or monotone change in population risk, sequence embeddings, or repetition: those depend on sampling error and on the student trajectory distribution. The next subsection separates these effects.

Reverse KL breaks the mixture (Corollary 3). Replacing $\alpha\, D_{\mathrm{KL}}(p_T\|p_S)$ with $\alpha\, D_{\mathrm{KL}}(p_S\|p_T)$ removes the linearity in $\log p_S$: the stationarity condition $-(1-\alpha)p_D(v)/p_S(v) + \alpha[\log p_S(v) - \log p_T(v)] + \lambda = 0$ is transcendental and has no mixture solution. Reverse KL is mode-seeking, it penalizes mass where p_T is small, so p_S concentrates on the teacher's dominant modes. When those modes are far from the corpus (a mismatched teacher), this simultaneously loses the data coverage that aids prediction and concentrates generation on off-corpus modes, worse on *both* axes, as observed in both modalities (§5.1, §6). At $\alpha = 1$, however, both KL directions have the same unrestricted optimum $p_S = p_T$; their empirical differences at pure KD require finite-capacity or optimization effects. The experiments therefore establish that changing direction does not resolve the observed mismatch cost under the tested training conditions.

B.1 From mixture to trade-off: a two-risk theorem

Proposition 1 says where training pulls the student; it does not yet explain why the *same* α helps one metric and hurts the other. An explanatory model applies *different geometries to the same deviation*. Let p^* be a strictly positive next-token distribution on a finite vocabulary at a context, let the empirical distribution of its n training observations be $p_D = p^* + \varepsilon$ (multinomial sampling noise, $\mathbb{E}[\varepsilon] = 0$), and let $b = p_T - p^*$ be the teacher's bias. By Proposition 1 the unconstrained mixture target is

$$p_S^\star(\alpha) \;=\; p^* + (1-\alpha)\,\varepsilon + \alpha\, b\,: \tag{7}$$

a single deviation vector whose *variance* part shrinks with α and whose *bias* part grows with it.

We model the two evaluations with quadratic surrogates in different norms. *Prediction* is held-out log-loss, whose second-order expansion is the χ^2-geometry, $D_{\mathrm{KL}}(p^*\,\|\,q) = \frac{1}{2}\|q - p^*\|_{\chi^2}^2 + O(\|q-p^*\|^3)$ with $\|\delta\|_{\chi^2}^2 = \sum_v \delta(v)^2/p^*(v)$: deviations are amplified by $1/p^*(v)$, so log-loss is *tail-weighted* locally. For a bounded feature map ϕ, the squared error in the embedding mean

is $\|\Phi\delta\|^2$ with $\Phi\delta = \sum_v \delta(v)\,\phi(v)$: deviations are summed under *bounded* weights. We use this quadratic feature error as a generation surrogate. Full Fréchet also contains covariance terms, and its local approximation requires regularity of those covariances; sequence-level Fréchet is not this per-context quadratic. Let $V_j = \mathbb{E}[\varepsilon^\top H_j \varepsilon]$ and $B_j = b^\top H_j b$ for positive semidefinite geometries H_j, with b fixed independently of the sampling noise. Writing R_j for these *surrogate* risks gives

$$R_P(\alpha) \;=\; \tfrac{1}{2}\big[(1-\alpha)^2 V_P + \alpha^2 B_P\big], \qquad R_G(\alpha) \;=\; (1-\alpha)^2 V_G + \alpha^2 B_G \tag{8}$$

(cross terms vanish since $\mathbb{E}[\varepsilon] = 0$).

Theorem 1 (Two-risk trade-off under quadratic surrogates)**.** *Assume $V_j + B_j > 0$ for each geometry and the zero-cross-term condition above. Both risks in Equation* (8) *are strictly convex in α with unique minimizers $\alpha_P^* = V_P/(V_P + B_P)$ and $\alpha_G^* = V_G/(V_G + B_G)$, and:*

(i) ***(Trade-off region.)*** *The interval (α_G^*, α_P^*) is nonempty* if and only if *$V_P B_G > V_G B_P$, and on it R_P is strictly decreasing while R_G is strictly increasing: every α in the interval improves prediction at the expense of generation, and none is dominated.*

(ii) ***(Aligned teacher.)*** *If $b = 0$ and both variances are positive, both risks reduce, up to their fixed scale factors, to $(1-\alpha)^2 V_\bullet$, strictly decreasing on $[0, 1)$: distillation improves* both *axes and no trade-off exists.*

(iii) ***(Data scaling.)*** *If $V_j = c_j/n$ with $c_j > 0$ and each $B_j > 0$ is fixed, both minimizers decrease monotonically to 0 as $n \to \infty$.*

Proof. Each risk is a quadratic $a(1-\alpha)^2 + c\,\alpha^2$ with $a, c \geq 0$, strictly convex when $a + c > 0$, with derivative $-2a(1-\alpha) + 2c\alpha$, zero at $\alpha^* = a/(a+c)$, negative below and positive above. The ordering $\alpha_G^* < \alpha_P^*$ is equivalent to $V_G(V_P + B_P) < V_P(V_G + B_G)$, i.e. $V_G B_P < V_P B_G$, giving (i); on (α_G^*, α_P^*), α lies above R_G's minimizer and below R_P's. (ii) and (iii) follow by substitution. □

This explains a possible mechanism for divergence; it is not an identity for measured rollout Fréchet. The trade-off interval describes marginal changes with α; it does not imply generation is worse than scratch throughout that interval. For the stated quadratic, $R_G(\alpha) > R_G(0)$ requires $\alpha > 2\alpha_G^*$. A positive optimum below the smallest tested weight is consistent with scratch winning a discrete sweep. Rollout drift and covariance effects fall outside this static surrogate.

How mismatch can create the trade-off. The condition compares the teacher's bias with the variance it removes in each geometry. If its feature-space bias is large relative to feature-space variance while its predictive bias is modest relative to predictive variance, then $V_P B_G > V_G B_P$. This provides a mechanism consistent with the observed divergence without assuming that every distant teacher satisfies it. Rare-token sensitivity alone is insufficient: under multinomial sampling on m supported categories, $\mathbb{E}\sum_v \varepsilon(v)^2/p^*(v) = (m-1)/n$. Catastrophic held-out losses also fall outside the local Taylor approximation. The aligned-teacher and data-scaling cases supply complementary qualitative checks (Figures 4 and 6); the experiments do not estimate every quadratic coefficient or identify zero teacher bias.

Adaptation and correlated errors. Adapting a teacher on the student's training sample can make $C_j = \mathbb{E}[\varepsilon^\top H_j b]$ nonzero. The corresponding quadratic includes $2\alpha(1-\alpha)C_j$, so the theorem's minimizers need not apply to that intervention. The controlled experiment establishes the change in fidelity directly. It does not require the quadratic assumptions to hold.

What the available curves can and cannot identify. In principle, three values of $R_G(\alpha)$ plus an independently measured finite-sample floor c_0 determine the quadratic coefficients. In practice this is fragile for Fréchet metrics: their plug-in estimators are sample-size biased, and the bias depends on the evaluated distribution rather than being a universal additive constant (Bińkowski et al., 2018; Chong & Forsyth, 2020; Gui et al., 2024). Our archived n=200 text calculation makes the problem visible: at its measured real–real floor, the fitted cross term violates $X^2 \leq VB$. The previously reported text value B/V=1.14 and implied α_G^*=0.38 require instead an unarchived floor $c_0 \approx 0.160$; no matched-sample calculation in the released logs supports that choice. We therefore withdraw both numbers rather than tune the floor to make the decomposition feasible. The same issue prevents a calibrated cross-modal B/V ranking from the present sample sizes.

The qualitative theorem remains useful, but it is no longer presented as a fitted predictor. Its aligned limit has a genuinely out-of-fit check: the ImageGPT teacher's own primed CLIP-Fréchet is 0.0529, ahead of $\alpha{=}1$ (0.0564), $\alpha{=}0.5$ (0.0588), and from-scratch (0.0617). More importantly, the controlled PubMed intervention in Fig. 4 changes one teacher's adaptation while holding corpus, student, tokenizer, and protocol fixed. It reveals a threshold-like switch, not the monotone graded curve that a scalar-distance prescription would require. A teacher's excess surprisal on the target relative to the web reference (Section 10) remains a useful within-model alignment diagnostic, but it is not equated with B/V or a norm across modalities. Calibrating an unbiased distance-to-risk mapping is future work.

Scope. The theorem concerns quadratic surrogates around the static mixture target. It motivates looking for teacher supervision that reduces variance without introducing excessive bias. DA-KD operationalizes this idea through observed-token rank, whose empirical value is evaluated in Section 9. Rank does not identify the largest distributional bias, and gating can change both bias and variance. Free-running rollout adds a separate context-distribution effect, formalized next in Section B.2 (cf. Zhao et al., 2026; Ross et al., 2011). Our primed protocol fixes the horizon within each comparison, allowing a controlled empirical ordering without assuming that the ordering must persist at every horizon.

B.2 Conditional vs. rollout risk: prediction and generation are different objectives

Theorem 1 explains the trade-off *geometrically*: log-loss and Fréchet weight the same deviation differently (tail vs. bulk). There is a second, *structural* difference between the two axes: they integrate that deviation against different *context distributions*. Prediction is teacher-forced: every context is drawn from the data, whereas generation rolls the student forward on its *own* prefixes. This subsection makes that distinction precise for reverse-KL risks, bounds the possible effect of context drift, and identifies alignment as a sufficient condition for matching trajectory laws. The bound does not impose a positive gap or monotone degradation.

Fix a horizon T and population conditional laws defined on reachable contexts. We analyze the idealized mixture target of Proposition 1, not an identity for a finite trained network: at every context c, let the student be the mixture $q(\cdot \mid c) = (1-\alpha)\, p_D(\cdot \mid c) + \alpha\, p_T(\cdot \mid c)$. Write $q_{1:T}$ and $p_{1:T}$ for the trajectory laws of the student and of the data, $q_{<t}$ and $p_{<t}$ for their prefix marginals, and

$$\bar{R}_{\mathrm{cond}} \;\coloneqq\; \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{x_{<t}\sim p_{<t}}\, D_{\mathrm{KL}}\big(q(\cdot \mid x_{<t}) \,\big\|\, p_D(\cdot \mid x_{<t})\big) \tag{9}$$

for a per-step reverse-KL diagnostic with contexts pinned to the data. Held-out NLL instead measures forward KL up to data entropy; this diagnostic isolates the effect of changing the prefix measure while keeping the integrand fixed. In particular, the forward-KL chain rule uses data prefixes, so likelihood does control forward trajectory KL in the population. It need not determine the ordering of Fréchet or repetition.

Proposition 2 (Conditional and rollout risk under context drift)**.** *Let $\delta \coloneqq \sup_c \mathrm{TV}\big(p_T(\cdot \mid c),\, p_D(\cdot \mid c)\big)$ and $\kappa \coloneqq \sup_c D_{\mathrm{KL}}\big(q(\cdot \mid c) \,\|\, p_D(\cdot \mid c)\big) < \infty$. The laws start from the same prompt distribution. Finite κ requires compatible support; an unsmoothed empirical law with zero-probability tokens need not satisfy it. The total-variation statements do not require finite KL. Then:*

- *(i)* ***(Same integrand, different measure.)*** $D_{\mathrm{KL}}(q_{1:T} \,\|\, p_{1:T}) \;=\; \sum_{t=1}^{T} \mathbb{E}_{x_{<t}\sim q_{<t}}\, D_{\mathrm{KL}}\big(q(\cdot \mid x_{<t}) \,\big\|\, p_D(\cdot \mid x_{<t})\big)$*: the identical per-step divergence as $\bar{R}_{\mathrm{cond}}$, integrated against the student's own prefix law $q_{<t}$ rather than the data's $p_{<t}$.*
- *(ii)* ***(Per-step bias.)*** *For every context c, $\mathrm{TV}\big(q(\cdot \mid c), p_D(\cdot \mid c)\big) = \alpha\, \mathrm{TV}\big(p_T(\cdot \mid c), p_D(\cdot \mid c)\big)$ and $D_{\mathrm{KL}}\big(q(\cdot \mid c) \,\|\, p_D(\cdot \mid c)\big) \le \alpha\, D_{\mathrm{KL}}\big(p_T(\cdot \mid c) \,\|\, p_D(\cdot \mid c)\big)$; for $\alpha > 0$, both vanish at c iff $p_T(\cdot \mid c) = p_D(\cdot \mid c)$.*
- *(iii)* ***(Context drift.)*** *The prefix laws separate at most linearly, $\mathrm{TV}(q_{<t}, p_{<t}) \le \min\{1, (t-1)\,\alpha\delta\}$.*

(iv) ***(Decomposition.)*** $D_{\mathrm{KL}}(q_{1:T} \,\|\, p_{1:T}) \;\le\; T\,\bar{R}_{\mathrm{cond}} \;+\; \frac{1}{2}\,\kappa\,\alpha\delta\,T(T-1)$*: a horizon-linear conditional reverse-KL term plus an upper bound on the effect of context drift. The second term is quadratic in horizon before using the TV cap; it is a worst-case bound, not a guaranteed increase in the realized risk.*

(v) ***(Alignment.)*** *If $p_T = p_D$ then $q_{1:T} = p_{1:T}$ for every α: both risks vanish identically and the two objectives coincide.*

Proof. (i) is the chain rule for the KL divergence of sequential distributions. (ii): the TV identity follows from $q - p_D = \alpha(p_T - p_D)$ pointwise; the KL bound follows from joint convexity of KL in its first argument, $D_{\mathrm{KL}}\big((1{-}\alpha)p_D + \alpha p_T \,\|\, p_D\big) \le (1{-}\alpha)\cdot 0 + \alpha\, D_{\mathrm{KL}}(p_T\|p_D)$. (iii): by the coupling characterization of TV and the standard one-step hybrid argument, $\mathrm{TV}(q_{<t}, p_{<t}) \le \mathrm{TV}(q_{<t-1}, p_{<t-1}) + \mathbb{E}_{q_{<t-1}}\,\mathrm{TV}\big(q(\cdot \mid x_{<t-1}), p_D(\cdot \mid x_{<t-1})\big) \le \mathrm{TV}(q_{<t-1}, p_{<t-1}) + \alpha\delta$ by (ii); induct. The unit cap follows from TV being at most one. (iv): for $g(c) := D_{\mathrm{KL}}(q(\cdot \mid c)\|p_D(\cdot \mid c)) \in [0, \kappa]$, a change of measure gives $\mathbb{E}_{q_{<t}} g \le \mathbb{E}_{p_{<t}} g + \kappa\,\mathrm{TV}(q_{<t}, p_{<t})$. This is the one-sided form: for $0 \le g \le \kappa$, $\mathbb{E}_Q g - \mathbb{E}_P g = \int g\,\mathrm{d}(Q{-}P) \le \kappa \int_{\{\mathrm{d}Q \ge \mathrm{d}P\}} \mathrm{d}(Q{-}P) = \kappa\,\mathrm{TV}(Q, P)$, which saves the factor of two in the uncentered Hölder bound $|\int g\,\mathrm{d}(Q{-}P)| \le \|g\|_\infty \|Q - P\|_1 = 2\|g\|_\infty\,\mathrm{TV}(Q, P)$, valid for g ranging over $[-\kappa, \kappa]$ rather than $[0, \kappa]$. Apply (iii) and sum over t, using $\sum_t (t{-}1) = \frac{1}{2}T(T{-}1)$. (v): the mixture of identical conditionals is the conditional itself, so $q = p_D$ at every context and both trajectory laws agree. □

Interpretation: different measures can matter. The proposition identifies $\alpha\delta$ as a factor in an upper bound on context drift. An aligned teacher, or zero teacher weight, removes this source of drift in the idealized mixture. The converse need not hold for equality of the two scalar risks: if the per-context KL is constant, changing the context measure changes neither expectation, even for a mismatched teacher. The practical distinction is that generation diagnostics score the student's trajectories, and their ordering must be measured. DA-KD uses target-token rank to choose which supervision to retain. A single observed rank is not an estimate of conditional TV and does not prove that the gate removes its largest value.

Two regimes under the surrogate model. The minimizers $\alpha_j^* = V_j/(V_j + B_j)$ distinguish settings where teacher bias limits useful transfer from settings where variance reduction dominates. In the first regime, B_G/V_G exceeds B_P/V_P and the fidelity surrogate favors less teacher weight. In the second, both bias-to-variance ratios are small and both minimizers approach one. These are explanations for the observed orderings, not estimates of the underlying coefficients. Qwen3 illustrates useful dense transfer at both tested token budgets: DA-KD is neutral at 50M tokens and harmful at 8M (Section R). The alignment diagnostic supports interpreting this teacher as comparatively well matched, but does not establish $b = 0$, and the theorem does not by itself predict the effect of gating.

Scope of the measurements. Our primed comparisons fix the prompt distribution and horizon so that differences in generation are evaluated under a common protocol. The proposition allows larger context drift over longer horizons, but does not guarantee that the measured fidelity ordering is preserved or amplified. That is an empirical question. The claim supported here is that likelihood improvements alone do not establish improvement in the reported generation diagnostics.

C Extended related work

Distillation as regularization. Our Proposition 1 places the distillation weight in the same family as label smoothing, which is exactly the α-mixture with a *uniform* teacher (Müller et al., 2019), and as self-distillation, shown to amplify regularization in function space (Mobahi et al., 2020). Menon et al. (2021) give a statistical perspective in which a good teacher supplies lower-variance Bayes-class-probability targets; our Corollary 1 is the sequence-modeling, finite-data instance, and our Corollary 2 adds the generative cost that variance-reduction analyses omit. **KL direction.** The mode-covering/mode-seeking distinction (Minka, 2005) underlies reverse-KL sequence distillation (Gu et al., 2024) and f-divergence generalizations; we contribute the observation that, for a distribution-mismatched teacher, teacher-forced token-level reverse KL is dominated on both axes in the reported comparisons; this does not rule out on-policy distillation. **Likelihood vs. sample quality.** The divergence between held-out likelihood and sample quality is long-noted for

generative models (Theis et al., 2016); we give a controlled, cross-modal, and mechanistically-explained instance in discrete sequence distillation. **Codec and recurrent LMs.** See §3; the length-generalization advantage of recurrent/state-space models over positional-encoding Transformers complements reports for Mamba (Gu & Dao, 2024) and RWKV (Peng et al., 2023) in the language setting. **Full codec lineage.** Jukebox (Dhariwal et al., 2020) models VQ-VAE codes (van den Oord et al., 2017); SoundStream (Zeghidour et al., 2021) and DAC (Kumar et al., 2023) develop neural audio codecs. **Sequence-level and self-distillation.** Soft targets were extended to the sequence level (Kim & Rush, 2016) and to self-distillation via born-again networks (Furlanello et al., 2018); recent work decouples KL direction from the trajectory/prefix choice in autoregressive LLMs (Zhao et al., 2026). Our matched sweeps separately test KL direction and mixture weight (§5.1). A concurrent method, AMiD (Shin et al., 2026), also invokes an "α-mixture," but of a different object (a generalized assistant *distribution* shaping the divergence target), whereas our α is the standard cross-entropy–KL interpolation weight whose optimum we *analyze* (Proposition 1); the shared name denotes different constructions. **Selective and teacher-side masking.** Learned feature-space masking (Huang et al., 2023) and dropout-based self-distillation (Lee et al., 2023) provide distinct forms of regularization; teacher-*side* interventions attack the teacher–student mismatch from the trajectory rather than the position axis, selectively invoking the teacher on student-generated sequences (Koo et al., 2025) or adapting teacher exposure during training (Han et al., 2026), complementary to our position-level masking.

D PREDICTIVE ENTROPY: THE MECHANISM IN THE DATA

Proposition 1 gives the static target $p_S^\star = (1-\alpha)p_D + \alpha p_T$. Entropy concavity implies $H(p_S^\star) \geq (1-\alpha)H(p_D) + \alpha H(p_T)$. This lower bound does not imply monotonicity in α. The measured mean per-token predictive entropy on held-out text rises across the tested forward-KL weights:

α (direction)	0	0.5	1 (forward)	1 (reverse)
mean predictive entropy (nats)	3.87	4.31	4.33	4.05

Forward distillation inflates entropy monotonically ($3.87 \rightarrow 4.33$): the student's predictions become more diffuse and teacher-like, consistent with the generic drift observed in generation (Fig. 9). Reverse KL, by contrast, yields *lower* entropy than forward ($4.05 < 4.33$), mode-seeking concentrates the distribution, but onto teacher-mismatched modes, which is why it is worse on both axes (§5.1). The two KL directions are thus separated not only by their outcomes but by their distributional fingerprint.

E SAMPLING ROBUSTNESS

To confirm the generation-fidelity drift is a property of the learned distribution and not an artifact of the sampling rule, we regenerate the audio α=0 and α=1 corpora with nucleus sampling (top-p=0.9) in place of top-k. The ordering is preserved, primed FAD 0.0091 (α=0) vs. 0.0171 (α=1), the same direction as under top-k (0.0037 vs. 0.0111), so the drift is a property of the learned distribution, not a multinomial-sampling artifact.

F FULL DISTILLATION-WEIGHT SWEEP

Endpoints of the dial. The zero-shot teacher predicts held-out FMA at PPL 74, ahead of every student, and the best student closes 32% of the label-only-to-teacher gap with a third of the parameters. Distillation transfers the prior efficiently but does not manufacture capacity; an α=1 model is the best predictor and the worst imitator of a small corpus.

Table 1 reports the headline cells. Table 5 adds the two blocks moved here for space: the scarce-data FMA 5 h corpus (which drives the memorization result of §5) and the parameter-matched Transformer, whose α-orderings replicate the xLSTM's on both corpora and establish that the trade-off is a property of the objective rather than the backbone (§5, Section I).

Table 5: Additional sweep cells. Protocol identical to Table 1. The Transformer is parameter-matched to the xLSTM (98.5M); FAD is not computed for the 5 h cells, whose purpose is the likelihood/memorization comparison.

Corpus	Backbone	α	Steps	Val CE ↓	Val PPL ↓	FAD ↓
FMA (5 h)	xLSTM	0	10k	6.79	889	–
	xLSTM	0.5	10k	6.37	582	–
	xLSTM	1	10k	6.26	526	–
FMA (66.6 h)	Transf.	0	10k	5.83	340	**0.057**
	Transf.	0.5	10k	5.52	250	0.093
MAESTRO (199 h)	Transf.	0	10k	3.11	22.3	**0.003**
	Transf.	0.5	10k	**2.96**	**19.4**	0.010
	Transf.	1	10k	3.06	21.4	0.014

Table 6 gives the audio counterpart of the text replication table (Table 2), scored under the same primed protocol. The MAESTRO reverse-KL cell is the audio analogue of the text result in §6: reverse KL is worse on *both* axes in *both* modalities.

Table 6: Audio counterpart of Table 2 (single-domain MAESTRO piano, MusicGen teacher). Fidelity is CLAP-Fréchet on 150 primed continuations, separately from the unconditional sweep and powered check; ↓ better.

α (dir.)	PPL ↓	overfits?	primed-Fréchet ↓
0 (scratch)	22.7	no	**0.0037**
0.5 forward	**19.8**	no	0.0066
1 forward	22.1	no	0.0111
1 *reverse*	31.4	no	0.0156

G THE EUROSAT CONTROL: THE DISTANCE IS IN TOKEN SPACE

This appendix records the supporting protocol for the compact comparison in Section 8. We use the same ImageGPT teacher–student pair and generation settings as Flowers-102, evaluate n=1000 primed completions, and compute the paired CLIP-Fréchet interval against scratch. Representative completions are shown in Figure 12, alongside the Flowers-102 completions of Figure 11; the full numerical result is reported in Section 8.

H XLSTM RECURRENCES AND STATE-SIZE DERIVATION

The recurrent backbone (§3.1) is an mLSTM (xLSTM (Beck et al., 2024)), which compresses history into a matrix memory $C_t \in \mathbb{R}^{d_{qk} \times d_v}$ updated by exponential input/forget gating:

$$C_t = f_t\, C_{t-1} + i_t\, k_t v_t^\top, \qquad n_t = f_t\, n_{t-1} + i_t\, k_t, \qquad h_t = o_t \odot \frac{C_t^\top q_t}{\max\big(|n_t^\top q_t|,\ e^{-m_t}\big)}, \tag{10}$$

with q_t, k_t, v_t the projections, n_t a normalizer, m_t the exponential-gate stabilizer from the cited mLSTM implementation, and o_t an output gate. The recurrence is associative and admits a chunkwise-parallel $\Theta(S)$ training algorithm. At generation, the entire state is the per-head triple (C, n, m), whose size is *independent of sequence length*:

$$\text{State}_{\text{xLSTM}} = Nh\,(d_{qk} d_v + d_{qk} + 1) \approx 0.45\text{M floats} \approx 1.8\,\text{MB (fp32)}. \tag{11}$$

Here N is layer count and h is heads per layer. A parameter-matched Transformer keeps a key–value cache that grows linearly with generated length S:

$$\text{State}_{\text{Transf}} = 2NSd \approx 75\,\text{MB (fp32) at } S{=}1024, \tag{12}$$

and unboundedly for long-form audio, the structural advantage that makes the many full sweeps in this paper affordable.

Table 7: Cross-length evaluation on MAESTRO (α=0, equal tokens/update and steps at every length): validation CE (nats/token) of models trained at S_{train}, evaluated at S_{eval}, same held-out tracks. Lengths are actual codec frames; archive names `s512/s1024/s2048` denote runs launched with 500/1000/2000 frames.

		S_{eval} (codec frames)		
Backbone	S_{train}	500	1000	2000
xLSTM	500	3.137	3.124	3.117
xLSTM	1000	3.138	3.123	3.114
xLSTM	2000	3.151	3.135	3.126
Transf.	500	3.114	3.607	4.357
Transf.	1000	3.121	**3.105**	3.637
Transf.	2000	3.131	3.113	**3.101**

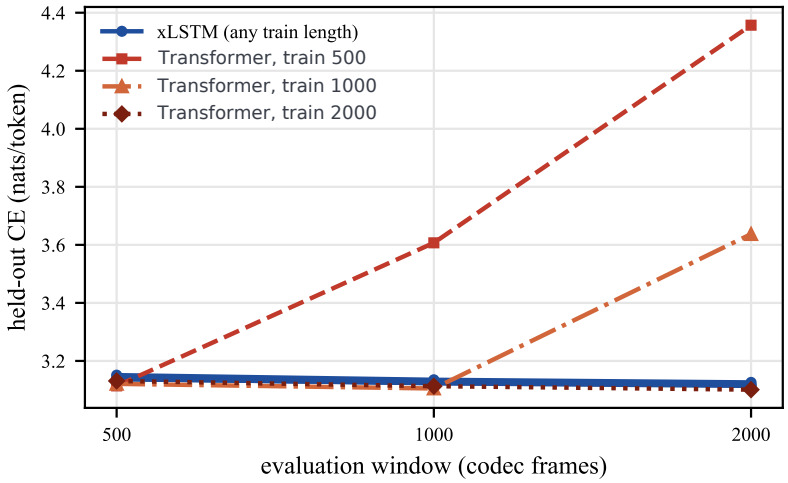

Figure 7: Length generalization (Table 7 visualized). The recurrent model is length-invariant, held-out CE flat and improving as the evaluation window grows past training length, while the matched Transformer degrades sharply once evaluated beyond its training window. With a constant generation state, this is the paper's core architectural payload.

I ARCHITECTURE: LENGTH GENERALIZATION AND STATE ECONOMICS

At α=0 the matched Transformer and xLSTM sit at likelihood parity (PPL 340 vs. 352, $\sim$1% in nats, identical validation-peak step); under distillation the Transformer leads (250 vs. 291); both α-orderings replicate across backbones. In systems terms the Transformer trains 2.2$\times$ faster here (fused attention kernels vs. pure-PyTorch recurrence, a kernel, not architecture, comparison), while the recurrent model's advantages are structural: $\Theta(S)$ training, and a generation state fixed at 1.8 MB versus a KV cache at 75 MB per 1024 steps and growing, the property that makes unbounded streaming generation and edge deployment plausible. At S=1024, attention's quadratic term is not yet binding; we therefore make no quality claim for recurrence beyond parity, and attribute its value to the state economics, and to length generalization, which the cross-length matrix of Table 7 isolates directly.

The recurrent model is *length-invariant*: all nine cells within 0.037 nats, every row improving monotonically with evaluation context (Fig. 7), a model trained on 10-second windows predicts 40-second windows better than its training length. The Transformer holds a $\sim$2–3% PPL edge at matched train/eval lengths but fails beyond them ($+1.24$ nats at 4$\times$ its training length). The released baseline uses Llama rotary position embeddings, not untrained absolute position embeddings. These results characterize this implementation and window range, not all self-attention models; the recurrent model's state remains constant with length.

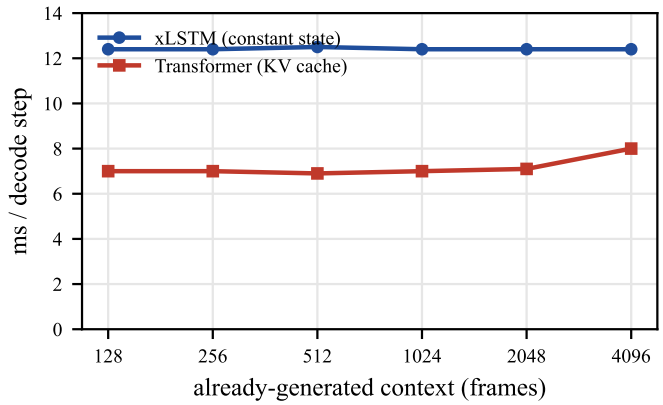

Figure 8: Per-step decode latency vs. context (RTX 5080). The recurrent step is flat across a 32× range; the Transformer is faster below ∼4k frames but rises as its cache grows. The recurrence advantage is in state *memory*, not latency, at these lengths.

Table 8: Corpora, ordered by teacher–domain distance. That ordering is the paper's independent variable: in our endpoint studies it predicts the direction of the generation effect (§8).

Distance	Corpus	Modality	Teacher	Train	Held-out
ALIGNED	Flowers-102	vision	ImageGPT	2,000 img	val split
BENIGN[a]	EuroSAT	vision	ImageGPT	2,000 img	1,200 img
DISTANT	FMA-small	audio	MusicGen	66.6 h (8,000 tr.)	5%
	MAESTRO v3	audio	MusicGen	199 h (1,276 perf.)	5%
	PubMed	text	4 families[b]	2–50 M tok	1,000 art.
EXTREME	UniRef50	protein	GPT-2	≈2 M tok	1,000 seq.

[a] Semantically distant from ImageNet but *pixel*-benign, so the effective token-space distance stays small (§8).
[b] GPT-2, Pythia, SmolLM2, Qwen3; per-family configurations in Table 10.
Tokens: audio, EnCodec-32 kHz (4×2048 codes at 50 Hz); vision, ImageGPT 512-cluster indices (32×32 flattened, 1024/image); text and protein, the teacher family's BPE (49–152k). Splits are by whole track, article, sequence, or image.
Sources: Flowers-102 (Nilsback & Zisserman, 2008), EuroSAT (Helber et al., 2019), FMA (Defferrard et al., 2017), MAESTRO (Hawthorne et al., 2019), PubMed long-document corpus (Cohan et al., 2018), UniRef50 (Suzek et al., 2007).

Compute and latency. Profiled at the ATen level (PyTorch `FlopCounterMode`; module-hook counters like `fvcore` undercount the recurrence), a forward at $S{=}1024$ costs 0.187 GFLOP/frame (xLSTM) and 0.221 (Transformer), 0.047 GFLOP per codec token given the $Q{=}4$ delay packing, matching the analytic $2P$ to 5%. We also measure single-step decode latency (batch 1, bf16, Fig. 8): the xLSTM step is context-invariant (12.4 ms, flat), confirming the constant-state property, but the Transformer decodes *faster* at every tested length (7–8 ms, its KV attention cheap below ∼4k frames) and the xLSTM constant is inflated by an unfused native step kernel. The recurrence's realized generation advantage is thus *memory* (fixed 1.8 MB vs. an unbounded KV cache), not latency at music-relevant lengths; both stream faster than real time. This is the honest scope of the architectural claim.

J DATASETS

Table 8 gives exact counts for all four corpora. The three specialized corpora (piano, medical, flowers) are the narrow domains; "teacher–domain" names whether the pretrained teacher's distribution covers the corpus (aligned) or not (distant), which §8 associates with the direction of the generation trade-off.

Table 9: Shared training recipe by modality. Text covers four teacher families (Table 10); the recipe below is common to all.

	Audio (MUSIFY)	Text (LM)	Vision (ImageGPT)	Protein (GPT-2)
Student / teacher	98.5M / 297M	Table 10	76M / 380M	124M / 355M
Optimizer	fused AdamW	fused AdamW	fused AdamW	fused AdamW
Base LR	3×10^{-4}	3×10^{-4}	3×10^{-4}	3×10^{-4}
(β_1, β_2)	$(0.9, 0.95)$	$(0.9, 0.95)$	$(0.9, 0.95)$	$(0.9, 0.95)$
Weight decay	0.05	0.05	0.05	0.05
LR schedule	cosine, $0.1\times$ floor	cosine	cosine, $0.1\times$ floor	cosine
Grad clip	1.0	1.0	1.0	1.0
Batch	8 seq ($\approx$1k frames)	8 seq $\times$ 512	8 img $\times$ 1024	8 seq $\times$ 512
Precision	bf16 autocast	bf16 autocast	bf16 autocast	bf16 autocast
Memory	none	none	grad. checkpointing	none

Table 10: Text/protein distillation families. All use the shared recipe (Table 9); students are random-initialized and distilled with full-vocabulary KL against the frozen teacher on tokenized PubMed (protein: UniRef50). "tok/param" is the token budget over student parameters, the fit criterion: below $\approx$0.05 the student is undertrained. The Qwen pair is listed at both budgets it was run at: starved at 8M, but well-fit at 50M, where the absence of a trade-off is explained by teacher *alignment* rather than capacity (Section 10).

Family	Student (init)	Teacher (frozen)	Tokens	tok/param	Role
GPT-2 (Radford et al., 2019)	124M	355M (medium)	2 M	0.016	primary text
Pythia (Biderman et al., 2023)	160M	410M	8 M	0.050	2nd family (well-fit)
SmolLM2 (Ben Allal et al., 2025)	360M	1.7B	20 M	0.055	modern family (well-fit)
Qwen3 (Yang et al., 2025)	0.6B	1.7B	8 M	0.013	scale probe (starved)
larger budget	0.6B	1.7B	50 M	0.083	well-fit; teacher aligned
Protein (GPT-2)	124M	355M (medium)	$\approx$2 M	0.016	extreme-distance point

K FULL REPRODUCIBILITY: HYPERPARAMETERS AND GENERATION

Table 9 lists the training recipe, deliberately shared across modalities to keep the comparison clean; Table 10 gives the per-family text/protein distillation configurations (four teacher families across a 124M–1.7B range). All runs use fused AdamW, bf16 autocast without gradient scaling, gradient clipping at 1.0, and identical seeds across cells. Distillation temperature is τ=1 throughout, and DA-KD uses $\alpha = 0.5$, with K=10 for text unless swept; the powered audio control uses K=16. The audio setting is a separate control, not evidence of a universal cutoff. The released audio configuration specifies hidden size 768, 12 layers, and 8 mLSTM heads. Its Llama baseline uses 12 attention heads and MLP width 2048; complete implementation settings are in `musify/model.py`.

Step budget and epoch counts (§4). Audio cells use a fixed effective batch of 16×1000 codec frames/step (two accumulated batches of 8 sequences), so a 20k-step run is $\approx$1.3B target tokens. Substituting that token count into the $6PD$ proxy gives 7.7×10^{17} (Hoffmann et al., 2022), but overcounts the shared backbone: all codebooks share each frame-level forward pass. This is not a measured training-FLOP total; the per-frame profiling in Section I is the relevant compute unit. Because the step is a fixed frame count, 10k steps is $\approx$14 epochs of FMA-66.6 h, $\approx$187 of the 5 h subset, and $\approx$5 of MAESTRO, which is why the overfitting reported in §5 tracks epochs rather than steps.

Generation protocols. The headline audio sweep (Tables 1 and 5) uses 200 unconditional clips. The separate audio replication (Table 6) primes with 4.0 s of real audio, then samples at temperature 1.0, top-k=250; 150 continuations per model. The powered audio check (Table 12) instead uses n=1000 matched primed continuations. These are distinct evaluation sets. *Text*, prime with the first 24 words of a real passage, then sample 150 new tokens at temperature 0.9, top-k=50; all reported fidelity comparisons use the powered n=1000 bootstrap (App. L). *Vision*, prime with the first 33% (338 of 1024) of a real image's pixels, then complete to 1024 at temperature 1.0, top-k=100; scored at n=1000 (CLIP is 512-d, so the earlier n=200 was rank-deficient). *Protein*, prime with the first 24 residues, then sample 150 tokens at temperature 0.9, top-k=50; scored at n=1000 in ESM-2 space.

Prompt (real): *"reversible protein phosphorylation is a major mechanism in the regulation of fundamental signaling events in all living organisms. in"*

α=0**:** "the case of pdd1d, the pd in vitro and then showed that pkd3-mediated expression was important in human patients. these data suggest that pd-1d-mediated signaling in the cell line…" [on-domain, rough]

α=0.5**:** "vivo, cell growth is dependent on the ability of different receptors to act on multiple protein interactions in the blood. In, we can explore their efficacy in different cells or transgenic models…" [fluent, drifts generic]

α=1**:** "hibition of the p21 phosphorylation process. [1213],] We, [4] These, [4] In…[2,8],], [3,3,4-" [over-committed, degenerates]

Figure 9: Primed continuations at increasing α (identical prompt). Distillation trades on-domain specificity (α=0) for general fluency (α=0.5) and, at the extreme, coherence (α=1), the qualitative face of the prediction/generation exchange.

PROMPT: *"reversible protein phosphorylation is a major mechanism in the regulation of fundamental signaling events in all living organisms. in bacteria, four types…"*

Scratch (α=0)**:** "…of inflammation: phagocytic (phag), myocytosis, pla2, and the p2 gene…the phosphorylation of the in response of proinflammatory cytokines such as p38 mapk and p38 mapk." [on-domain, rough]

Standard KD (α=0.5)**:** "…of autophagy. A. 2. p. (p. (1.p.3%)0 (p.0.03)1.0 (p.0.03) = 0.004-1.0…(p.067)1.0.1 (p.0.08)0.0 (p.02) 0.020.1 (p.0.06)1.7 (p.0001) 1.2 (p.01)2.3…" [degenerates into notation]

DA-KD (informed, K=10)**:** "…of transcription factors, including protein kinase and p53, are involved in the control of apoptosis…inhibition of p38 mapk activation was shown to be dependent on the p44/p7 binding, which is inactivation of the p38 mapk pathway." [coherent, on-domain]

Figure 10: **The generation cost, and DA-KD's fix, made concrete** (verbatim primed continuations, identical prompt, GPT-2→PubMed). Standard KD collapses into repetitive statistical notation, the qualitative face of the rep-4 increase (Section 9); the informed gate keeps the student coherent and in-domain while retaining the teacher's prediction benefit.

Why primed, not unconditional. Unconditional generation from a bare start token collapses for these small models (the $\alpha \geq 0.5$ students emit near-degenerate text), inflating the unconditional Fréchet to 0.57→1.16, an artifact rather than a 2× fidelity cliff. The primed protocol (Table 2, Fig. 9) removes this confound; the powered GPT-2 comparison resolves increased repetition, not a fidelity loss; the prediction results carry no such caveat.

Fitting distillation on a 16 GB GPU. Two memory patterns recur and are worth flagging for reproduction on consumer hardware. First, models with eager (non-fused) attention retain the full attention maps for the backward pass: ImageGPT at sequence length 1024 peaks at ≈21 GB and silently spills to host memory over PCIe, collapsing throughput (a 20× slowdown that presents, misleadingly, as 99% GPU utilization at low power). *Gradient checkpointing*, recomputing activations in the backward rather than storing them, cuts this to ≈5 GB and restores full speed with no change to the result (checkpointing is exact). Second, materializing the cross-entropy logits in fp32 over a large vocabulary is a needless memory spike, since autocast already evaluates the loss in fp32 internally; leaving them in bf16 and keeping the training batch below the memory ceiling is what makes large-vocabulary students trainable in 16 GB. Every run in this paper trains within this budget.

L Statistical methodology and metrics

Evaluation provenance. GPT-2 replication, adaptation (`dose_response.json`), and gate-width sweep (`bootstrap_frechet_n1000.json`) are separate evaluations with their own baselines. Pythia's random-mask, joint-gate, and seed results are archived respectively as `pythia8m_variants_result.json`, `gate_comparison.json`, and `pythia_seeds_result.json`. Each interval uses its own matched reference; the random row's delta is not against the seed-mean row. Archived bootstrap-mean deltas can differ from subtraction of the full-sample estimates.

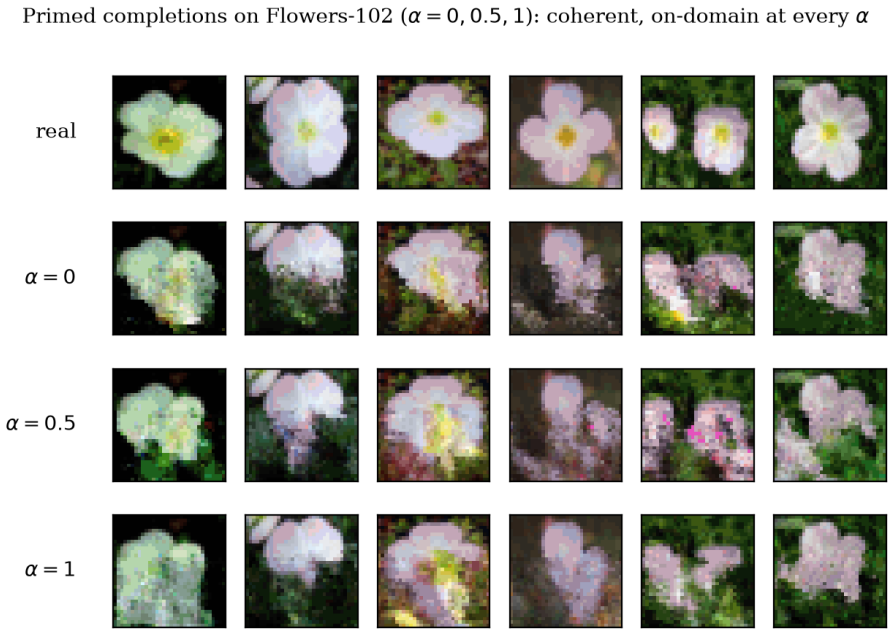

Figure 11: Primed completions on Flowers-102 (each column shares a real 33% prefix). Rows: real, then α=0, 0.5, 1. Completions remain coherent, on-domain flowers at every α, consistent with the monotone fidelity gain of Fig. 4.

GPT-2 rep-4 uses the frozen `primed_rep4_n1000.npy` caches for `a0/a05/a1/a1_rev/dakd_k10`. We recompute its percentile intervals with B=10,000, seed 0, sampling each run independently; these are not prompt-paired intervals. The lexical table is recomputed from the saved `gens_n1000.json` files, with prompt-paired resampling (B=2000, seed 0). The scratch generation JSON and rep-4 array are different cached draws, so the lexical table does not combine their scores. The released `metric_reconciliation.json` records input hashes and exact values. Earlier Fréchet, MAUVE, and reader evaluations retain their archived evaluation snapshots; they are corroborating evaluations, not a single shared generation draw.

Table 11: Excess: target minus common generic-web surprisal (descending); Δcov: top-10 coverage change. Brackets: document-clustered bootstrap 95% CIs (Section L); †per character.

teacher	target	excess [95% CI]	Δcov [95% CI]	trade-off?
GPT-2-med	UniRef50†	$+2.409$ $[+2.358, +2.452]$	-0.399 $[-0.412, -0.385]$	yes
GPT-2-med	PubMed	$+0.336$ $[+0.214, +0.463]$	-0.052 $[-0.061, -0.044]$	yes
SmolLM2-1.7B	PubMed	-0.127 $[-0.217, -0.034]$	-0.007 $[-0.015, +0.001]$	yes
Pythia-410M	PubMed	-0.168 $[-0.275, -0.058]$	-0.019 $[-0.027, -0.010]$	yes
Qwen3-1.7B	PubMed	-0.459 $[-0.549, -0.368]$	$+0.040$ $[+0.032, +0.049]$	no

Web reference. Released alignment scripts load `Skylion007/openwebtext` and held-out PubMed articles. Token-, word-, and character-normalized diagnostics must retain their stated units; they are not a common cross-modality norm.

Audio sweep CIs (§5). We bound the 200-clip sampling variance with percentile bootstrap 95% CIs (1000 resamples). The FAD optimum's *location* sits entirely within that variance: on FMA, α=0.5 $[0.067, 0.091]$ overlaps α=0 $[0.079, 0.092]$, so we make no claim about which low-α weight is best. The low-vs-heavy gap, by contrast, is highly significant. The FMA α=1 degradation is completely isolated ($[0.102, 0.148]$, overlapping neither), and on MAESTRO the α=1 degradation separates cleanly from label-only ($[0.015, 0.017]$ vs. $[0.003, 0.004]$). A paired bootstrap on the independent

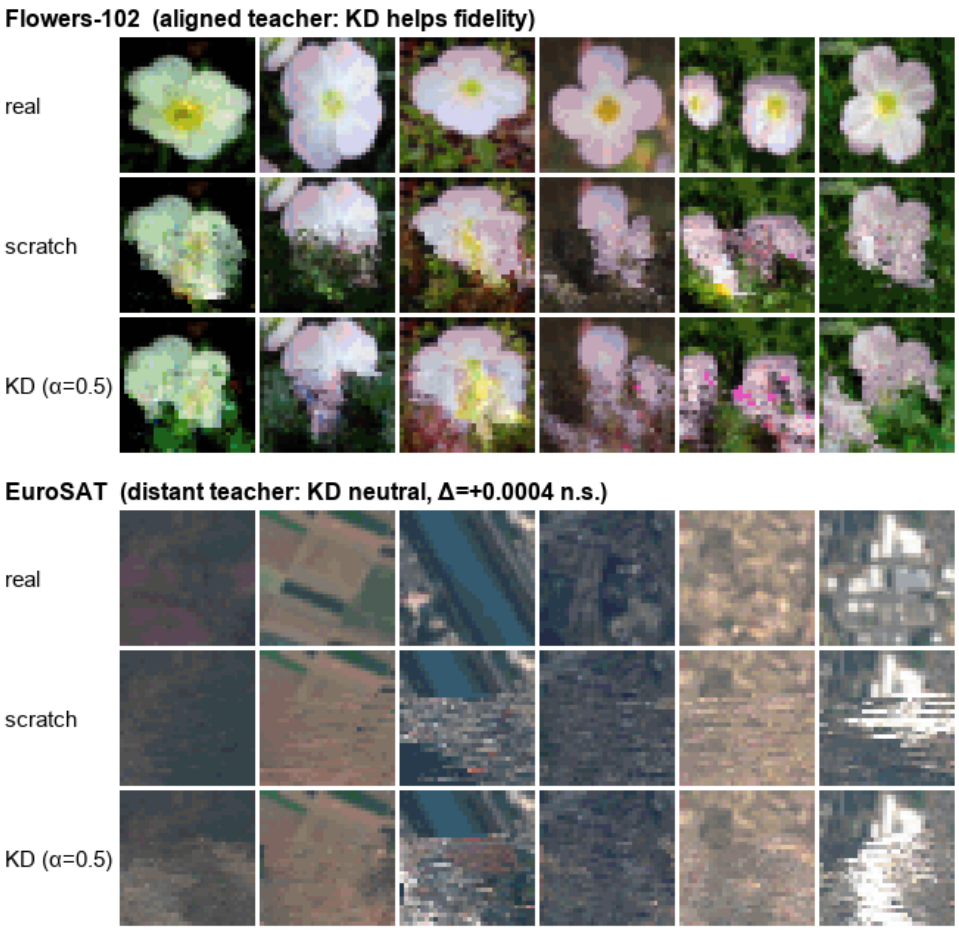

Figure 12: **Primed vision completions across the distance axis** (each column shares a real 33% prefix; rows: real source, scratch α=0, standard KD α=0.5). Completions are plausible and on-domain in *both* settings, and by eye scratch and KD are close, so the teacher's effect is quantitative, not a gross visual artifact. On aligned Flowers the teacher gives a small but significant fidelity gain (CLIP-Fréchet $\Delta=-0.0039$, $n=1000$); on distant EuroSAT scratch and KD are indistinguishable, matching the null ($\Delta=+0.0004$, n.s.). The coherent satellite textures under an ImageNet teacher show why the distant-vision case is *pixel-benign* rather than harmful (§8): the teacher's low-level priors transfer even off-domain.

primed protocol confirms the same direction: α=1 is significantly worse than α=0, while the α=0.5 midpoint sits within sampling variance.

The dense-KD fidelity cost is significant despite its better likelihood (26.1 vs. 30.6). Every gate leaves perplexity at dense KD's level (26.0–26.2), so on audio the gate carries no prediction cost. **At K=64 the informed gate significantly improves fidelity over dense KD** ($\Delta=-0.00082$, CI $[-0.00143, -0.00020]$), closing about a third of the gap between dense KD and scratch, while scratch remains best.

On audio the informed selection is what helps, at matched retention. We trained both uninformed controls at the K=64 retention count, so that the three rules differ only in *which* positions keep the teacher KL. Neither control shows a resolved fidelity improvement over dense KD: the random gate has $\Delta=+0.00030$ (CI $[-0.00031, +0.00095]$), and the anti-confidence gate has $\Delta=-0.00001$ (CI $[-0.00062, +0.00073]$), while the informed gate is significantly better than both (random $\Delta=-0.00112$, CI $[-0.00183, -0.00045]$; anti $\Delta=-0.00081$, CI $[-0.00152, -0.00006]$, the latter only marginally excluding zero). The direct comparisons support informed posi-

Table 12: Powered non-text robustness check on MAESTRO. Each model completes the same n=1000 deterministic held-out windows after a 4 s real prime; the 10 s tails are embedded with CLAP-music. The informed gate applies KD where the true token lies in the teacher's top K, for $K \in \{16, 64\}$; each gate width has its own random and anti-confidence controls, retaining the same dynamically determined number of positions per batch as that width, so rules within a block differ only in *which* positions are kept. Intervals are paired percentile-bootstrap intervals (B=500). All rows are one matched run series, so perplexities here differ from the separate sweep of Table 1.

Training rule	Val PPL ↓	FAD ↓	ΔFAD vs. dense KD [95% CI]
Scratch (α=0)	30.6	**0.00068**	−0.00235 [−0.00309, −0.00161]
Dense KD (α=0.5)	26.1	0.00303	reference
DA-KD, K=64	26.0	**0.00220**	**−0.00082** [−0.00143, −0.00020]
random gate, K=64 count	26.1	0.00332	+0.00030 [−0.00031, +0.00095]
anti-confidence, K=64 count	26.1	0.00302	−0.00001 [−0.00062, +0.00073]
DA-KD, K=16	26.2	0.00272	−0.00031 [−0.00096, +0.00036]
random gate, K=16 count	26.1	0.00305	+0.00003 [−0.00057, +0.00065]
anti-confidence, K=16 count	26.1	0.00271	−0.00031 [−0.00093, +0.00028]

tion selection over these retention-matched alternatives in this audio run, complementing the Pythia controls (Table 4). At the narrower K=16 width the informed gate's recovery is not significant, and its differences from the K=16-matched controls are unresolved (−0.00034, CI [−0.00097, +0.00032]; 0.00000, CI [−0.00058, +0.00069]). Both widths and all controls are reported; pairwise intervals are in `runs/audio_dakd_k64_pairwise.json` and `runs/audio_k64_controls_vs_{dense,dakd}.json`.

Embedding spaces. Every fidelity number is a Fréchet distance between generated and real sets in an *independent* embedding space (never the teacher's), to avoid circularity: CLAP-music (Wu et al., 2023) for audio (FAD), all-MiniLM-L6-v2 (Sentence Transformers, n.d.), a contrastively fine-tuned MiniLM sentence encoder, for text; CLIP ViT-B/32 (Radford et al., 2021) for vision; and ESM-2 (Lin et al., 2023) (`esm2_t6_8M_UR50D`, 320-d) for protein. Using a domain-appropriate embedder matters: a natural-image/text embedder collapses on protein strings (a degenerate 0.0000), the cross-modal analogue of the rank-deficiency artifact above. We compute the Fréchet distance $\|\mu_1 - \mu_2\|^2 + \mathrm{Tr}\big(\Sigma_1 + \Sigma_2 - 2(\Sigma_1\Sigma_2)^{1/2}\big)$ via its eigenvalue-trace form: since $\mathrm{Tr}((\Sigma_1\Sigma_2)^{1/2}) = \sum_i \sqrt{\lambda_i(\Sigma_1\Sigma_2)}$, the interaction term needs only the eigenvalues of $\Sigma_1\Sigma_2$, not the matrix square root. This is numerically identical to the standard `sqrtm` form (agreeing to four decimals) but ∼50× faster, which matters because each bootstrap replicate recomputes the distance thousands of times (App. L, next). We also drop the dependence on `scipy.linalg.sqrtm` for this calculation; that function remains available in SciPy.

Corroboration with MAUVE. Because primed-Fréchet is our own construction, we cross-check the GPT-2/PubMed text result against MAUVE (Pillutla et al., 2021), a standard divergence between model and human text distributions (higher is better; opposite direction to Fréchet). In its archived n=1000 primed-continuation evaluation, MAUVE is 0.920 from-scratch, **drops to** 0.797 **under standard KD**, and **recovers to** 0.938 **under DA-KD**, exceeding both baseline point estimates and supporting, on a metric outside our embedding space, that standard KD can impose a generation cost that the gate removes. Notably MAUVE tracks the *repetition* finding (§9), not primed-Fréchet, which was within noise for standard KD: two independent metrics (rep-4, MAUVE) thus agree the cost is real, so embedding-Fréchet if anything *understates* it. Two caveats: we report the point estimate over the full n=1000 set without a bootstrap CI, so this comparison does not establish statistical significance, and we featurize with the locally cached `gpt2-medium` rather than MAUVE's default `gpt2-large`; robustness of the ranking to that featurizer change is untested.

The $n = 1000$ paired bootstrap. To test whether DA-KD's fidelity recovery is significant, we regenerate 1000 primed continuations per run against a fresh 1000-passage held-out set (disjoint from training; §9), embed them, and bootstrap (B=2000). We report the *paired* difference CI, not marginal per-run CIs, to target the run-to-run comparison. Fréchet/FID has finite-sample bias (Bińkowski et al., 2018); in our resamples, marginal intervals lie above the point estimates. Scoring every run against the *same* resampled reference set preserves dependence from the shared

reference and estimates uncertainty in the paired difference. Pairing does not guarantee cancellation of estimator bias or nominal interval coverage. At n=200, per-generation noise alone swung baselines by $\sim$0.014; n=1000 provides greater precision for the comparison. The resulting significant improvements (for K=10: Δ=-0.028 Fréchet vs. standard KD, CI $[-0.041, -0.020]$) refer to the gate-width evaluation. The separately reconciled rep-4 comparison gives Δ=-0.034, CI $[-0.041, -0.028]$ with independent per-run resampling. The extended sweep in Table 13 finds significant fidelity gains for every tested $K \geq 5$; the extreme K=1 gate is not significant.

M CONTROL EXPERIMENTS: WHAT MAKES DA-KD WORK

DA-KD gates the KL to the positions where the teacher is domain-competent. Three questions determine why it helps; we settle each with a powered n=1000 paired bootstrap (text, PubMed; Section 9, Table 3).

The informed selection is what generalizes. We first asked whether the benefit is the informed *selection* or merely sparsity. On GPT-2 the two are hard to separate: at matched sparsity (27% dropped), a rank-informed gate (drop the surprised tail), a random mask, and an *anti*-informed mask (drop the teacher's most-confident tokens) are statistically indistinguishable on primed-Fréchet (all pairwise paired-bootstrap CIs include 0). Read on GPT-2 alone, that suggests any sparse mask suffices. A second teacher family separates them: on Pythia-410M→160M the random mask is significantly *worse* than standard KD (Δ=+0.015, CI $[0.006, 0.025]$) while the informed gate is significantly *better* (Δ=-0.019, CI $[-0.027, -0.010]$). Audio shows the same ordering outside text: at matched retention neither the random nor the anti-confidence gate differs from dense KD, while the informed gate beats both (Table 12). Where the comparison resolves, it favours dropping the teacher-surprised positions; GPT-2 leaves it unresolved rather than contradicting it. On this evidence the informed selection is load-bearing in two of the three settings tested, which is weaker than universal transfer.

Protocol for the confidence gate (Table 4). The competing informed gate of Section 9 keeps the KL at the 73% of positions with the lowest teacher predictive entropy $H(p^{\mathcal{T}}_{s,k})$, thresholded per batch by the empirical quantile so the realized drop rate matches the random and rank gates exactly; unlike the rank gate it never consults the observed token, which is the property under test. It runs in the most discriminative cell (Pythia-410M→160M, PubMed 8M tokens) under the setup of Table 3: same frozen teacher, random-init student, KD weight, temperature, optimizer, seed, prompts, decoding and n=1000 paired bootstrap. Only the selection rule differs across the three gates.

Not a lower effective α. Because the objective renormalizes over kept positions, a uniform random 27% drop preserves the expected KL gradient at α=0.5, conditional on retaining a nonempty subset. It concentrates KL on retained tokens rather than multiplying its mean by the retention rate. The heuristic attenuation α_{eff}=$0.5 \times 0.73 \approx 0.364$ is a separate control, not an equivalent objective: uniform KD at α=0.364 (perplexity-matched to the random-drop-27% run) scores primed-Fréchet 0.2237, *worse* than standard KD and significantly worse than random-drop-27% (Δ=+0.058, CI $[0.049, 0.073]$). Lowering the weight uniformly hurts fidelity.

Alignment with prediction-axis selection. Student-entropy selection can outperform random selection on predictive metrics (Tavor et al., 2026); this does not establish an advantage for every informed selector. Our Pythia controls separately show a generation-fidelity advantage for the observed-target gate, consistent with the teacher-bias mechanism of Theorem 1.

Seed replication. Retraining Pythia-410M→160M at seeds $\{0, 1, 2\}$ preserves the ordering in every matched pair (Figure 13). Across the three training seeds, primed-Fréchet is 0.2039 ± 0.0065 for standard KD and 0.1805 ± 0.0015 for DA-KD (mean $\pm$ sample SD); these SDs describe *training-seed variation*. Separately, within each trained seed, an n=1000 paired evaluation bootstrap estimates the DA-KD effect as Δ=-0.028 (95% CI $[-0.038, -0.019]$), -0.016 ($[-0.024, -0.008]$), and -0.031 ($[-0.041, -0.022]$), matching the seed-level values quoted in §9. These intervals quantify uncertainty from the paired generated evaluation samples, not variation over training seeds. With only three seeds we do not perform or imply a seed-level significance test; the evidence is the consistent direction and the descriptive across-seed summaries. The accompanying perplexity ranges are DA-KD 58.7–59.8 and standard KD 56.0–57.6.

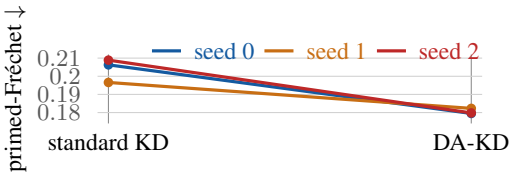

Figure 13: **Matched-seed Pythia replication.** Each line connects point estimates from standard KD and DA-KD trained with the same seed; all three slopes favor DA-KD. This plot shows training-seed variation. The paired evaluation-bootstrap intervals reported in the text condition on each trained pair and answer a different uncertainty question.

Robust to the gate width K. K is the lone hyperparameter, and DA-KD is not brittle to it (Table 13). Perplexity varies *monotonically* with K: tighter gates drop more positions and give up prediction ($52.9 \to 82.9$ as K falls from V to 1). Fidelity has a wide operating range: for every $K \in \{5, 10, 20, 50, 100, 200\}$, primed-Fréchet falls from standard KD's 0.194 to 0.160–0.172, and the paired-bootstrap 95% CI for the difference excludes zero. The extreme $K=1$ gate retains only 46% of KD positions and is not significantly better than KD ($\Delta = -0.006$, CI $[-0.021, 0.002]$). Thus K behaves as a prediction dial with a broad fidelity sweet spot, not a knife-edge threshold; we use $K=10$ for text as a moderate point in that range.

Table 13: Gate-width sweep (GPT-2→PubMed, primed-Fréchet at $n=1000$, paired bootstrap vs. standard KD). Perplexity is monotone in K; every $K \geq 5$ significantly improves fidelity over standard KD.

variant	drop rate	val ppl ↓	primed-Fréchet ↓	Δ vs. KD
from scratch	n/a	128.9	0.196	n/a
standard KD ($K \to V$)	0%	**52.9**	0.194	(ref)
DA-KD, $K=200$	9%	67.3	0.166	-0.029^*
DA-KD, $K=100$	12%	69.3	0.165	-0.030^*
DA-KD, $K=50$	16%	69.6	0.172	-0.022^*
DA-KD, $K=20$	22%	70.7	0.166	-0.028^*
DA-KD, $K=10$	27%	72.6	0.166	-0.028^*
DA-KD, $K=5$	34%	73.5	**0.160**	-0.034^*
DA-KD, $K=1$	54%	82.9	0.189	-0.006 (n.s.)

* paired-bootstrap 95% CI excludes 0.

Fidelity is not monotone in perplexity across the K sweep, which rules out the reading that DA-KD is simply an undertrained student.

N LEXICAL DOMAIN-TERM PRESERVATION

Table 14: Domain-term preservation on 996 eligible articles from 1000 GPT-2/PubMed continuations (articles must have at least 300 words). *Paired bootstrap 95% CI vs. standard KD excludes 0.

model	domain-term recall ↑	precision ↑	F1 ↑
from scratch	0.033^*	0.0566	0.041^*
standard KD	0.028	0.0546	0.035
DA-KD	$\mathbf{0.039}^*$	$\mathbf{0.0601}^*$	$\mathbf{0.046}^*$

We evaluate lexical preservation on frozen GPT-2/PubMed generation JSONs ($n=1000$; Section L): does each model reproduce domain vocabulary from the article's true continuation? This reference-overlap diagnostic requires no new training or human annotation. It does not measure factual comprehension or task accuracy.

Protocol. Each model receives the first 24 words of a held-out article and generates 150 tokens; the prompt is then discarded. We compare the domain terms in that continuation against those in the

true continuation (words 24:174), excluding stopwords and every prompt token, so nothing can be won by echoing the prompt. Lacking an offline biomedical NER model, we use a lexical proxy for entity-like tokens: alphabetic terms of length $\geq$7, or alpha–digit mixes such as `p38` and `il6`. This is a proxy, not true NER, but it is computed identically for every model, and Section P repeats the measurement against an external controlled vocabulary so the criterion does not rest on a rule of our own choosing.

Result. DA-KD recovers significantly more of the article's domain vocabulary than standard KD (recall $+0.012$, precision $+0.006$, F1 $+0.012$; all paired-bootstrap CIs exclude 0), and standard KD falls *below from-scratch* training: its collapse into statistical notation (Figure 10) contributes almost no biomedical terms. The ordering DA-KD $>$ scratch $>$ KD agrees with the model ordering in the separate rep-4 evaluation: higher repetition accompanies lower domain-term overlap. This association does not separate lexical loss from repetition. Section Q tests the decoding-side question directly and finds that no decoder we tried lifts dense KD's overlap to DA-KD's, including one that removes repeated 4-grams outright; Section P repeats the comparison against MeSH with the same outcome.

Scope. Lexical overlap probes the generation axis, not every downstream use. Prediction and generation can prefer different students (Theorem 1 and Proposition 2); standard KD has the lowest perplexity here, while DA-KD preserves more reference vocabulary.

O OBJECTIVE BASELINES: STUDENT-PROPOSAL GATING AND ON-POLICY DISTILLATION

DA-KD verifies the *observed* next token against the teacher's top-K. Two families of prior work change that choice: selective distillation verifies the *student's* proposed token instead (Huang et al., 2025), and on-policy distillation introduces training on the student's own samples (Agarwal et al., 2024). This section compares both against DA-KD under one seeded generation session per family, on the n=1000 articles and primed protocol used throughout, so all rows here are mutually comparable. Perplexity is each cell's own best validation value.

The proposal gate is a controlled greedy student-top-1 gate ablation inspired by SelecTKD, not a reproduction of its full training recipe. Our GKD implementation uses forward KL on student continuations with the prompt masked; remaining steps use the standard $0.5\,\mathrm{CE} + 0.5\,\mathrm{KL}$ objective on data windows. Training uses 9760 optimizer steps, batch size 8, 512-token data windows, and 24-token rollout prompts with temperature 1 and top-$k = 50$ sampling. Intervals use $B = 2000$ percentile resamples: Fréchet shares the reference resample but resamples each generated set independently; repetition and lexical scores are paired by article. These intervals condition on the trained checkpoints.

The selection criterion separates on Pythia. On GPT-2 the proposal gate has no resolved difference from DA-KD on any metric (Fréchet -0.0015, rep-4 -0.0005, F1 -0.0014, MeSH -0.0033; all intervals contain zero), so this does not establish equivalence. On Pythia the proposal gate is significantly worse than DA-KD on rep-4 ($+0.0103$, CI $[+0.0049, +0.0159]$) and on domain-term F1 under our extractor (-0.0075, CI $[-0.0111, -0.0041]$), though the MeSH version of that difference is not resolved (-0.0043, n.s.), and it fails to improve on dense KD's repetition at all ($+0.0067$, CI $[+0.0006, +0.0130]$, worse). This repeats the structure of the random-mask control in Table 4: GPT-2 does not separate the selection rules, and Pythia supports the observed-token criterion's benefit there. Two families are not universal transfer, and we claim only that the criterion matters in the one family where the comparison resolves.

On-policy distillation removes the repetition defect but not at better fidelity. GKD reaches DA-KD's repetition at λ=0.5 (Δ=-0.0000 vs. DA-KD) and beats it at λ=0.25 (-0.0052, CI $[-0.0078, -0.0023]$), which is consistent with our account of the defect, though on-policy training changes several things at once and does not isolate the rollout mechanism. Embedding fidelity worsens across the tested λ values: GKD is worse than dense KD at λ=0.25 ($+0.0191$, CI $[+0.0033, +0.0251]$) and worse still at λ=0.5 ($+0.0429$, CI $[+0.0256, +0.0484]$), in both cases worse than DA-KD and, at λ=0.5, worse than training from scratch. Lexical overlap is between dense KD and DA-KD and does not separate from DA-KD at λ=0.25.

Table 15: Objective comparison on PubMed, one evaluation snapshot per family (n=1000; ↓ better for Fréchet and rep-4, ↑ for F1). The proposal gate applies the same top-K verification as DA-KD to the student's top-1 prediction instead of the observed token, at an approximately matched retention rate (0.757 vs. 0.728 on GPT-2). GKD samples 128 tokens from the student on a fraction λ of steps and matches the teacher there; an on-policy step costs ≈4.0× a supervised step. Measured step times imply estimated training-time ratios of 1.75× at λ=0.25 and 2.53× at λ=0.5 relative to all-supervised steps, excluding evaluation and checkpointing ($1 - \lambda + \lambda r$, with step-time ratio r; `text/runs/gkd*/result.json`).

family	objective	PPL ↓	Fréchet ↓	rep-4 ↓	F1 (ours) ↑	F1 (MeSH) ↑
GPT-2	from scratch	128.9	0.1957	0.0321	0.0419	0.0510
	dense KD	**52.9**	0.1881	0.0506	0.0334	0.0442
	DA-KD	72.6	0.1655	0.0176	**0.0478**	**0.0571**
	proposal gate	70.7	**0.1640**	0.0171	0.0464	0.0537
	GKD (λ=0.25)	58.2	0.2072	**0.0124**	0.0454	0.0558
	GKD (λ=0.5)	63.6	0.2309	0.0176	0.0444	0.0535
Pythia	from scratch	88.5	0.1937	**0.0285**	**0.0494**	**0.0596**
	dense KD	57.5	0.1992	0.0379	0.0410	0.0504
	DA-KD	59.3	**0.1830**	0.0343	0.0492	0.0576
	proposal gate	**57.0**	0.1919	0.0446	0.0416	0.0533

Scope of the on-policy comparison. We ran two λ values at one seed with a fixed sampling length (128 tokens), temperature and top-k, and did not tune the objective; the decoder baseline of Section Q, by contrast, was tuned on our own headline metric. These runs characterize the two configurations we trained; they do not bound what GKD can achieve, and a tuned on-policy baseline may do better. What they show is that on-policy training is not a free substitute in this setting: its estimated training time is 1.75–2.53× the all-supervised baseline, it does not reach DA-KD's fidelity at either mixing weight, and its resolved generation advantage is confined to repetition. It also achieves lower validation perplexity than DA-KD at both weights.

P External-vocabulary check: MeSH

The extractor in Section N is ours, so a reader cannot check it against anything outside this paper. We therefore repeat the measurement with membership in MeSH (Lipscomb, 2000) as the term criterion. The vocabulary is NLM's 2025 ASCII descriptor release (`d2025.bin`; National Library of Medicine, 2025): headings, ENTRY and PRINT ENTRY terms, plus un-inverted variants of comma-stored forms ("Neoplasms, Hepatic" → "hepatic neoplasms"), giving 234,655 surface forms. Terms under four characters and single-token stopwords are dropped, matching is longest-first over 4- down to 1-grams without overlap, and prompt terms are excluded on both sides. The rule was fixed before any result was computed and is applied identically to every row.

Table 16: Domain-term overlap under MeSH, on the single seeded generation session of Section Q (n=996 articles with at least one MeSH term in the reference continuation; ↑ better). ΔF1 is against DA-KD, paired bootstrap (B=2000, seed 0); * marks intervals excluding zero.

objective	decoder	recall ↑	F1 ↑	ΔF1 vs. DA-KD
DA-KD	default	**0.0495**	**0.0564**	—
standard KD	rep. penalty 1.10	0.0462	0.0507	−0.0056*
from scratch	default	0.0427	0.0491	−0.0074*
standard KD	no repeat 4-gram	0.0411	0.0481	−0.0083*
standard KD	default	0.0389	0.0460	−0.0105*
standard KD	nucleus 0.95	0.0375	0.0412	−0.0153*
DA-KD	nucleus 0.95	0.0454	0.0493	−0.0071*

The ordering holds under an external vocabulary. DA-KD recovers more of the reference continuation's MeSH vocabulary than every other row, including all four dense-KD decoder settings of Section Q and the decoder that beats it on primed-Fréchet. The separate frozen generations be-

hind Table 14 reproduce this independently (DA-KD 0.0563, scratch 0.0488, standard KD 0.0431; DA-KD over standard KD $\Delta{=}+0.0132$, CI $[+0.0078, +0.0185]$), so the effect is not a property of one generation draw, and the MeSH criterion agrees with the heuristic of Section N on the model ordering.

What the match consists of. Of the MeSH terms matched in reference continuations, 73% are single tokens, and the most frequent include general-purpose words that are also MeSH headings (*patients*, *treatment*, *time*, *role*, *history*, *form*). Restricting the criterion to multi-word MeSH terms, a stricter terminology subset, leaves no resolved difference between DA-KD and the dense-KD decoders (vs. standard KD $\Delta{=}-0.0013$, CI $[-0.0068, +0.0038]$; vs. the repetition penalty and the 4-gram block likewise n.s.), and the two generation snapshots disagree in sign there. These results support an advantage in reference lexical overlap; they do not establish that the advantage is specific to biomedical terminology.

Q DECODING-SIDE CONTROLS

DA-KD changes the training objective to address a defect that appears at generation time. A decoder change addresses generation directly, costs nothing at training, and leaves perplexity untouched. This section asks whether the decoder is sufficient.

Protocol. All rows use the GPT-2/PubMed cells of §9 and the primed protocol of Section K: the first 24 words of a held-out article, 150 new tokens, MiniLM primed-Fréchet against the real passages, per-text rep-4, and the domain-term F1 of Section N. Decoder settings are chosen on `val_real.json` (290 articles) and reported on `val_real_1k.json` (1000 articles); the two files are disjoint, and the reporting set is the one used throughout the paper. The selection criterion is lowest validation primed-Fréchet, the metric on which we would otherwise claim DA-KD's advantage, so the baseline is tuned on our own headline instrument; nucleus $p{=}0.95$ wins it. Every row below is generated in one seeded session so the table is a single evaluation snapshot rather than a mix of draws. Intervals are paired percentile bootstraps ($B{=}2000$, seed 0) sharing a real resample per replicate, as in Section L. Contrastive decoding is absent: in our pinned `transformers` version contrastive search is no longer in the library and requires executing remote code at generation time, which we do not do in the evaluation pipeline.

Table 17: Decoding-side controls on GPT-2/PubMed ($n{=}1000$ held-out articles; ↓ better for Fréchet and rep-4, ↑ better for F1). Decoders are selected on a disjoint 290-article set. Perplexity is a property of the trained model and is unchanged by decoding. Daggers mark the two rows whose decoder is the validation-selected one.

objective	decoder	PPL ↓	primed-Fréchet ↓	rep-4 ↓	domain-term F1 ↑
from scratch	default	128.9	0.1963	0.0312	0.0412
standard KD	default	**52.9**	0.1880	0.0509	0.0350
standard KD	rep. penalty 1.10	**52.9**	0.1733	0.0131	0.0408
standard KD	no repeat 4-gram	**52.9**	0.1761	**0.0002**	0.0371
standard KD	nucleus $0.95^{\dagger}$	**52.9**	0.1541	0.0141	0.0299
DA-KD	default	72.6	0.1679	0.0180	**0.0479**
DA-KD	nucleus $0.95^{\dagger}$	72.6	**0.1517**	0.0094	0.0397

The decoder recovers the repetition and the embedding distance. Every decoder setting significantly lowers rep-4 against dense KD at the default decoder, and the validation-selected nucleus setting also lowers primed-Fréchet ($\Delta{=}-0.034$, CI $[-0.043, -0.025]$). Against DA-KD at the shared default decoder it is better on both ($\Delta_{\text{Fréchet}}{=}-0.012$, CI $[-0.021, -0.002]$; $\Delta_{\text{rep-4}}{=}-0.004$, CI $[-0.007, -0.000]$), and it does so at dense KD's perplexity of 52.9 rather than DA-KD's 72.6. On these two axes the training-time gate is not the only route to the improvement, and not the cheapest one.

Lexical overlap orders the other way. Dense KD stays below DA-KD's domain-term F1 under every decoder tested: nucleus $\Delta{=}-0.018$, CI $[-0.021, -0.015]$; the hard 4-gram block $\Delta{=}-0.011$, CI $[-0.014, -0.008]$; repetition penalty 1.10 $\Delta{=}-0.007$, CI $[-0.010, -0.004]$. The decoder that wins on Fréchet is the worst of the four on overlap, falling below dense KD's own default

($\Delta{=}-0.005$, CI $[-0.008, -0.003]$). Applying the same decoder to DA-KD reproduces the pattern within that objective: Fréchet and rep-4 improve, overlap falls from 0.048 to 0.040.

Repetition and fidelity are not the same quantity. The hard 4-gram block drives rep-4 to 0.0002, an order of magnitude below real text, and still does not reach DA-KD's Fréchet ($\Delta{=}+0.010$, CI $[+0.000, +0.019]$). Removing the loops is therefore not sufficient for the fidelity difference, which bears on the reading of Section N in which DA-KD's lexical advantage would be an artifact of its lower repetition.

DA-KD's gains survive decoder tuning, in two families. The fair test of the objective holds the decoder fixed. Table 18 compares DA-KD with dense KD when both use the default decoder and when both use a tuned decoder, for GPT-2 and for Pythia-410M→160M at three training seeds. Each family's decoder is selected separately, on the same disjoint validation set and criterion; both independently select nucleus $p{=}0.95$. With both objectives on the tuned decoder, DA-KD still improves over dense KD in both families. In GPT-2 the gain is in repetition and reference overlap, under our extractor and under MeSH. In Pythia it is in embedding fidelity, which improves in all three seeds with little spread ($-0.008{\pm}0.001$), and in reference overlap under our extractor, also three of three seeds. The domain-term gain is the one common to both families. The training-time gate therefore contributes something the best decoder we found does not supply, although which generation axis carries the gain depends on the family.

Table 18: DA-KD minus dense KD at a matched decoder on PubMed ($n{=}1000$). GPT-2: one seed, 95% interval, * if it excludes zero. Pythia: mean $\pm$ sample SD over training seeds $\{0, 1, 2\}$, with the number of seeds whose own 95% interval excludes zero; we make no seed-level significance test. Negative Fréchet and rep-4, and positive F1, favour DA-KD. Fréchet intervals share the real resample per replicate; the others are paired over prompts ($B{=}2000$, seed 0). Tuned decoder: nucleus $p{=}0.95$, selected per family.

	GPT-2 355M→124M		Pythia 410M→160M (3 seeds)	
metric	default	tuned	default	tuned
primed-Fréchet ↓	-0.020^*	-0.002	$-0.020{\pm}0.006$ (3/3)	$\mathbf{-0.008{\pm}0.001}$ **(3/3)**
rep-4 ↓	-0.033^*	$\mathbf{-0.005}^*$	$-0.016{\pm}0.012$ (2/3)	$-0.002{\pm}0.003$ (0/3)
domain-term F1 (ours) ↑	$+0.013^*$	$\mathbf{+0.010}^*$	$+0.010{\pm}0.001$ (3/3)	$\mathbf{+0.006{\pm}0.003}$ **(3/3)**
domain-term F1 (MeSH) ↑	$+0.010^*$	$\mathbf{+0.008}^*$	$+0.009{\pm}0.003$ (3/3)	$+0.005{\pm}0.003$ (1/3)
MeSH multi-word F1 ↑	$+0.001$	$+0.001$	$+0.001{\pm}0.005$ (1/3)	$+0.001{\pm}0.001$ (0/3)

Scope. Matched comparisons cover two decoders in two families on one corpus. Dense KD keeps its perplexity advantage under any decoder, and specificity of the overlap gain to multi-word technical terms is not established (Section P); neither overlap measure is a factual-accuracy or task result.

R QWEN AT TWO DATA BUDGETS: THE ALIGNED SIGNATURE AND ITS DATA DEPENDENCE

Section 10 argues from the aligned-teacher limit; this section gives the full three-objective measurements behind it (Table 19), at the two budgets for which we have complete informed-gate runs. Both show the aligned signature: standard KD improves *both* axes over from-scratch training, which is what Theorem 1(ii) predicts when $b \to 0$ and is the opposite of the distant-teacher pattern of Section 6. What changes with the data budget is the size of the variance term, and with it the cost of gating.

S HUMAN EVALUATION

Section T evaluates the GPT-2/PubMed continuations with an LLM judge. This appendix replaces that proxy with human readers on the same stimuli (Table 20). The study is preregistered and the plan is released with the code.

Protocol. The design mirrors Section T so the two instruments are comparable: pairwise, blind, forced choice with ties permitted, the rubric wording copied verbatim, the same 24-word real pre-

Table 19: Qwen3-1.7B→0.6B on PubMed at two token budgets (primed-Fréchet and rep-4 at n=1000; Δ is the paired bootstrap against standard KD). At both budgets standard KD improves perplexity *and* fidelity, the aligned-teacher signature. As the budget grows the teacher's fidelity advantage over scratch shrinks ($-0.051 \to -0.007$) and DA-KD moves from significantly harmful to indistinguishable: with an aligned teacher there is no imported bias to remove in these comparisons; gating can forgo useful signal, and costs less as the student's own estimate improves.

budget	objective	ppl ↓	primed-Fréchet ↓	rep-4 ↓	Δ vs. KD
8 M (0.013 tok/param)	from scratch	56.5	0.225	0.055	n/a
	standard KD	**32.5**	**0.172**	**0.053**	(ref)
	DA-KD	34.5	0.186	0.056	+0.013 worse*
50 M (0.083 tok/param)	from scratch	35.7	0.159	0.067	n/a
	standard KD	**22.5**	**0.151**	0.068	(ref)
	DA-KD	23.6	0.153	**0.066**	+0.002 (n.s.)

* paired-bootstrap 95% CI excludes 0.

fixes, the archived n=1000 generation pool, and the same 1500-character truncation. Each rater judges 32 items, disjoint across comparisons so no prompt is seen twice: 16 for DA-KD vs standard KD and 8 each for the other two. All raters see the same items in independently shuffled order. The LLM judge sees every item in both presentation orders; a person would remember the item, so we counterbalance *across* raters instead, showing each item with the first system in slot A to half the panel. Position is balanced by construction and reported: slot A was chosen in 51.7% of non-tie trials.

Estimator. Items, not raters, are the generalization target, so intervals are 95% cluster bootstraps resampling items (B=10,000, seed 0), with ties dropped. The console report and released `human_eval/results.json` use the same bootstrap draw. Ties counted as one half and leave-one-rater-out recomputation are reported as preregistered sensitivity checks; neither changes any direction.

Table 20: Human preferences on GPT-2/PubMed primed continuations, n=4 raters, 128 judgements. Rates are for the preferred system with ties dropped; brackets are 95% cluster-bootstrap intervals over items. The LLM-judge column quotes the stable-only rate of Table 22, the comparable ties-dropped estimand.

comparison	humans prefer	rate [95% CI]	leave-one-out	LLM judge
DA-KD vs. standard KD	DA-KD	0.667 [0.528, 0.789]	0.634–0.698	DA-KD 0.610
DA-KD vs. from-scratch	DA-KD	0.688 [0.469, 0.875]	0.667–0.708	DA-KD 0.793
standard KD vs. scratch	**scratch**	0.645 [0.531, 0.781]	0.565–0.783	**KD** 0.607

DA-KD wins with every reader. Every rater, of every background, prefers DA-KD over standard KD and over from-scratch training (Table 21), and so does the LLM judge. On the two comparisons that involve DA-KD, humans and the judge agree.

Preferences by reader background. Humans and the judge part ways only on standard KD versus from-scratch. The descriptive background split shows differing preferences. Amendment 1 of the preregistration, filed before six of the seven recruited raters collected, records each rater's background before analysis and commits to reporting results split by it, naming the domain subgroup as the closest human analogue of a judge prompted as "a careful evaluator of biomedical scientific writing". The two domain experts, a practising physician and a biochemistry PhD, prefer from-scratch in 81% of non-tie pairs, with an item interval that excludes chance (KD preferred 0.19, [0.06, 0.38]). The computer scientist agrees (0.71 for scratch). The only reader who prefers standard KD is the lay rater, a literature student, at 0.75, which agrees in direction with the judge (0.61). These observations do not establish why readers differ or what cues determine their preferences. The physician completed pairs at a median of 2 s; timing alone does not identify the basis for a judgment. Pooled over the two domain experts, preference for DA-KD over scratch is 0.69 (CI [0.56, 0.88]); over all four raters, the estimate is also 0.69, but its interval includes chance ([0.47, 0.88]). The lay rater's

preference is 0.75, so the wider pooled interval does not reflect a lower preference from that reader. These are descriptive estimates from a small panel, not a test of differences between reader groups.

Table 21: Preference rates split by rater background, as preregistered (Amendment 1). Rates for the first-named system, ties dropped; brackets are 95% item-cluster bootstraps (B=10,000, seed 0) within each group. The split is descriptive: the preregistration fixes that no subgroup comparison is tested. Two domain raters, one technical, one lay.

readers	DA-KD > KD	DA-KD > scratch	KD > scratch
domain experts (2)	**0.71** [0.55, 0.87]	**0.69** [0.56, 0.88]	**0.19** [0.06, 0.38]
technical (1)	0.67	0.63	0.29
lay (1)	0.57	0.75	0.75
all humans (4)	0.67 [0.53, 0.79]	0.69 [0.47, 0.88]	0.35 [0.23, 0.47]
LLM judge (stable items)	0.61	0.79	0.61

Panel composition. Per rater, the preference for scratch over standard KD is 0.88 (physician), 0.75 (biochemistry PhD) and 0.71 (computer scientist), against 0.25 for the literature student. With two domain raters and eight to sixteen items per comparison these are estimates, not a subgroup test, as the preregistration specifies.

Panel size and exclusions. The preregistration planned seven fully crossed raters; four contributed analysed data, and every exclusion is recorded in its amendments. One exclusion is a researcher degree of freedom and is reported as such: the first submission for rater slot 4 passed every data-quality check but was withdrawn after scoring, when we chose to re-run that slot with a rater of a different professional background. As the preregistration requires, we report the panel both ways. Including that submission gives DA-KD over standard KD 0.685 ([0.551, 0.803]), DA-KD over from-scratch 0.650 ([0.425, 0.875]), and scratch over standard KD 0.692 ([0.575, 0.821]), against 0.667, 0.688 and 0.645 without it. Every estimate moves by less than 0.05, no direction or interval conclusion changes, and the withdrawn submission agreed with the panel majority on all three comparisons, preferring scratch over standard KD as the domain experts do. The background split of Table 21 covers the four analysed raters.

Power and scope. The study was preregistered as an estimation and convergent-validation design, with the power analysis fixed before collection: at the judge's 0.610 margin a panel of this size reaches 0.35 to 0.51, and adding raters to a fixed item set cannot escape item-level variance. We therefore planned for, and report, interval estimates rather than significance verdicts. Within that scope the panel is informative: every interval is consistent with the automatic instruments, the ordering is stable under leave-one-rater-out, and the one divergence between readers and the judge falls exactly where the judge already disagreed with primed-Fréchet and rep-4.

T LLM-JUDGE PROTOCOL

Table 22: Blind, counterbalanced LLM-judge preferences on GPT-2/PubMed (250 prompts per comparison), under two estimands. *Stable only*: verdicts that agree across both presentation orders (Wilson intervals; 29.2/20.8/28.8% of items excluded). *Full sample*: all 250 items, with order-inconsistent and TIE items scored 0.5 to each system (item percentile bootstrap, B=10,000, seed 0). Every ordering survives the full-sample treatment. Protocol below.

		stable only		full sample (ties = 0.5)	
comparison	preferred	rate	95% CI	rate	95% CI
DA-KD vs. standard KD	**DA-KD**	0.610	[0.537, 0.679]	0.578	[0.526, 0.628]
DA-KD vs. from-scratch	**DA-KD**	0.793	[0.731, 0.844]	0.732	[0.684, 0.778]
standard KD vs. from-scratch	standard KD	0.607	[0.533, 0.676]	0.576	[0.524, 0.628]

The judge of Table 22 is a reader-style check on the GPT-2/PubMed primed continuations of Section 9. It is a supplementary metric: unlike primed-Fréchet, rep-4 and MAUVE it calls a third-party API, so it cannot be reproduced from our released code alone. We report it because it probes a

different notion of quality from the distributional metrics, and we design it around the documented failure modes of LLM judges.

Design. Judgements are *pairwise*, since judges are far more reliable at A-vs-B than at absolute scoring; *blind*, with no indication of which system produced which continuation; and *fully counterbalanced*, i.e. each of the 250 prompts per comparison is judged twice, once in each presentation order, rather than relying on randomisation. Ties are permitted so that a forced choice does not manufacture signal, and the rubric explicitly instructs the judge to disregard length, which blunts verbosity bias.

Why counterbalancing is necessary. The judge exhibits a strong second-position preference: in a single-order pass it selects the second-shown continuation in 62% of cases. Under that pass DA-KD appears to beat standard KD 59.6% of the time, but the effect is carried entirely by order: DA-KD wins 70.5% when shown second and only 47.5% (n.s.) when shown first. Judging every item in both orders exposes order sensitivity. Our primary analysis then restricts to items whose verdict is *stable* across the two orders, so that estimand concerns order-stable items rather than the full prompt sample. Flip rates are 29.2% (DA-KD vs. standard KD), 20.8% (DA-KD vs. from-scratch) and 28.8% (standard KD vs. from-scratch); the flip rate falls as the quality gap widens, as expected if position decides only when the content signal is weak. Stable-item rates are exact binomial proportions with Wilson 95% intervals.

The full sample, with flips as ties. Because flips concentrate on the ambiguous comparisons, conditioning on stable items could by itself inflate the rates. We therefore also report the complementary estimand over all 250 items per comparison, scoring an order-inconsistent or TIE item 0.5 to each system (Table 22, right). Every ordering survives, with smaller margins: DA-KD over standard KD 0.578 (95% CI $[0.526, 0.628]$, bootstrap $p{=}0.003$), DA-KD over from-scratch 0.732 ($[0.684, 0.778]$, $p{<}0.001$), and standard KD over from-scratch 0.576 ($[0.524, 0.628]$, $p{=}0.003$); intervals are item percentile bootstraps ($B{=}10{,}000$, seed 0).

As a third check that does not condition on stability at all, we fit a Bradley–Terry model with an explicit position parameter over all 500 individual judgements per comparison, $\mathrm{logit}\ \Pr(x \text{ wins}) = \beta_0 + \beta_{\mathrm{pos}} z$ with $z{=}\pm\frac{1}{2}$ for x shown first or second, so that $\sigma(\beta_0)$ is the position-neutral preference. It reproduces the same picture: DA-KD over standard KD 0.582 ($[0.537, 0.626]$), DA-KD over from-scratch 0.740 ($[0.699, 0.777]$), and standard KD over from-scratch 0.582 ($[0.537, 0.626]$). The fitted position terms are large and negative in all three comparisons (β_{pos} from -0.68 to -1.07, SE ≈ 0.19), confirming the second-position preference the counterbalancing is designed to absorb. The full-sample and Bradley–Terry numbers are produced by `scripts/judge_full_sample.py` and released in `generated/judge_full_sample.json`.

Model and decoding. `deepseek-v4-flash` with extended reasoning enabled (`reasoning_effort=high`), 250 prompts per comparison, both orders, 1,500-character truncation per continuation. Prompts are the same 24-word real prefixes used throughout Section 9, and the continuations are the same cached $n{=}1000$ generations scored by the other metrics, so all four instruments judge identical text.

System message.

```
You are a careful evaluator of biomedical scientific
writing.  You compare two candidate continuations of a
real PubMed article and judge which reads like the genuine
continuation.
```

User message (rubric). Placeholders are filled with the shared prompt and the two continuations, in the order dictated by the counterbalancing schedule.

```
Below is the opening of a real PubMed article, followed by
two candidate continuations.
Which continuation is more plausible as the real
continuation of this article?
Judge on:
- domain-appropriate biomedical terminology and specificity
- internal coherence and absence of degenerate repetition or
filler
```

```
- consistency with the topic established by the opening
Explicitly IGNORE differences in length, and ignore that
both may be imperfect.
--- OPENING ---
{prompt}
--- CONTINUATION A ---
{a}
--- CONTINUATION B ---
{b}
Reply with exactly one token:  A, B, or TIE.
```