PREDICTIVE ANCHORING SHAPES CONTEXTUAL DEPENDENCE IN LANGUAGE-MODEL PRETRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

We use knowledge distillation to study learned contextual dependence during language-model pretraining by varying teacher context regimes and composition, while holding the student's architecture and input window fixed. A frozen short-context teacher supplies a *predictive anchoring* pressure to match locally conditioned predictions, alongside ordinary next-token learning. This short-context distillation (sKD) improves baseline predictions across WikiText-103 and SlimPajama-100M, while attenuating average contextual effects both where distant text helps baseline prediction and where it hurts. An equal-context distillation (eqKD) control, whose teacher receives the same causal input as the student, achieves lower overall loss than sKD but retains more of both helpful and harmful influences. We observe a tradeoff between helpful-context use and harmful-context attenuation in these contrasting distillations. Mixed-teacher distillation (mixKD) trains a single student using a fixed mixture of the two teacher distributions. mixKD achieves overall prediction loss close to eqKD, with lower loss in most, but not all, comparisons. It demonstrates substantially improved prediction on helpful positions compared with sKD, while retaining a prediction advantage over eqKD on harmful positions. It also shows greater average contextual benefit on helpful positions than sKD and a smaller average contextual penalty on harmful positions than eqKD. These findings hold across three seeds on validation and test data, at two evaluated short-context horizons on WikiText-103 and one on SlimPajama-100M. Equal-weight ensembles of separately trained sKD and eqKD students achieve lower overall prediction loss than mixKD, yet an exploratory post-test analysis finds that no fixed probability weighting of these student pairs simultaneously matches or improves on both mixKD's helpful-context benefit and harmful-context penalty. This study shows that similar overall prediction loss can accompany different learned contextual behavior, which can be shaped by teacher context regimes and composition.

1 INTRODUCTION

Next token prediction can depend on nearby words, on information introduced further back in the text, or on both. A model's overall prediction loss, however, does not reveal how much distant text contributes or whether that contribution is beneficial at every position. In experiments with Transformer language models trained on English Wikipedia, O'Connor & Andreas (2021) found that long contexts supported low perplexity even when their detailed syntactic and propositional content was substantially disrupted. These findings motivate an empirical study of learned contextual dependence during pretraining. We use knowledge distillation (KD) to compare various teacher context regimes while keeping student architecture and input access fixed. A teacher's context regime includes its pretraining context and its input access during distillation. First, we try to answer how a model's reliance on distant text changes when it is trained to match predictions based only on a recent suffix. We call this pressure to match the local predictions *predictive anchoring*.

We implement predictive anchoring through *short-context distillation* (sKD), where a student language model learns from observed next tokens while a KD loss encourages agreement with a frozen short-context teacher. We compare observed-token probabilities with and without text preceding a fixed recent window to characterize the learned response to distant text. Baseline models trained with two other random seeds are used to label positions as *helpful* or *harmful*, when they agree that

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.

distant text raises or lowers the observed token's probability, respectively. To clarify, these labels describe predictive effects rather than semantic relevance. We observe that sKD improves overall prediction compared with ordinary next-token pretraining (baseline) while attenuating the average influence of distant text in both helpful and harmful populations.

We next compare this behavior against *equal-context distillation* (eqKD), whose frozen teacher receives the same causal input as the student model. In our main comparisons on WikiText-103 and SlimPajama-100M, eqKD achieves lower overall loss than sKD. eqKD retains more of both the helpful and harmful effects on average than sKD, as measured by comparing each model's predictions with and without distant text. These comparisons reveal a tradeoff between retaining helpful contextual effects and attenuating harmful ones.

The complementary strengths of these two distillation regimes motivate *mixed-teacher distillation* (mixKD). Using a fixed mixture of the two frozen teacher distributions, mixKD replaces part of eqKD supervision with sKD's local predictive anchoring in a single student. When evaluated on the same full inputs, mixKD predicts better than sKD on helpful positions and better than eqKD on harmful positions. It also shows greater average contextual benefit on helpful positions than sKD and a smaller average contextual penalty on harmful positions than eqKD. This compromise repeats at two short-context horizons on WikiText-103 and one on SlimPajama-100M, each across three training seeds on validation and test data. We also observe lower overall validation and test loss than eqKD in most, but not all, tested combinations of seed, dataset and horizon. The observation that contextual tradeoffs can differ while the overall losses remain close suggests that teacher composition can help shape contextual reliance in ways that aggregate prediction loss alone cannot distinguish.

An ensemble comparison further illustrates this distinction. Equal-weight ensembles of separately trained sKD and eqKD students achieve lower overall prediction loss than mixKD, yet an exploratory post-test analysis finds that no fixed probability weighting of these student pairs simultaneously matches or improves on both mixKD's helpful-context benefit and harmful-context penalty (Section 5.3).

Our empirical contributions are as follows:

A student model utilises distant text differently based on the teacher's context regime. A short-context teacher leads to greater average attenuation of helpful and harmful contextual effects in the student model than an equal-context teacher.

A model can predict better despite attenuation of distant context. sKD lowers overall prediction loss relative to the baseline while reducing both the average benefit at helpful positions and the average penalty at harmful positions. In our main comparisons, eqKD predicts better overall than sKD but retains more of both helpful and harmful contextual effects.

Prediction loss alone does not characterize contextual behavior. mixKD and eqKD achieve close overall prediction losses but show different average benefits and penalties from distant text.

2 Related Work

Distillation and predictive regularization. Knowledge distillation transfers predictive distributions from teachers, including ensembles, to a student (Hinton et al., 2015). In Born Again Neural Networks, Furlanello et al. (2018) show that students with the same architecture as their teachers can outperform those teachers, including in language modeling. Weaker teachers can also benefit larger models during language-model pretraining (Lee et al., 2023), while KD's connection to learned label smoothing provides a regularization perspective (Yuan et al., 2020). MiniPLM uses teacher information to refine the pretraining-data distribution (Gu et al., 2025). Lakha et al. (2026) show that self-distillation with forward KL and next-token cross-entropy can mitigate overfitting under data scarcity. They train a randomly initialized student with the same architecture as a pretrained teacher, providing a close precedent for our eqKD control. Our comparison with sKD examines how the short-context teacher regime changes the student's reliance on distant text.

Supervision across context views. Context distillation internalizes behavior elicited by instructions, examples, or scratchpads so that a model can reproduce it without those context tokens (Snell et al., 2022). Cross-View Training trains restricted-view auxiliary modules to match full-view pre-

dictions through shared representations (Clark et al., 2018). Long–Short Alignment regularizes a single model's predictions across overlapping contexts to improve length generalization (Du et al., 2025). Induced Model Matching compares restricted-model predictions with student predictions averaged over longer contexts sharing the same short context (Muneeb & Ohannessian, 2024). Its idealized analysis shows that direct matching to a restricted predictor can fail to recover the true full-context predictor, even with infinite data. sKD uses direct matching to a frozen short-context teacher; our finite-data experiments examine the resulting predictive and contextual tradeoffs across teacher regimes.

Short-to-long supervision also appears in context extension and post-training. LongReD combines hidden-state restoration and distillation across positional ranges to mitigate short-text degradation during context extension (Dong et al., 2025). LongPO and SoLoPO transfer preferences using short contexts designed to retain task-relevant information (Chen et al., 2025; Sun et al., 2026). OPSDL constructs questions from extracted short passages and trains on generated responses using token-level supervision derived from reverse KL (Zhang et al., 2026). Our short-context teacher receives a fixed recent suffix, which can omit evidence needed to predict the observed next token. This motivates measuring attenuation of helpful as well as harmful contextual effects during pretraining.

Distillation from multiple teachers. Chebotar & Waters (2016) distill weighted averages of acoustic-model predictions into a single student and study how teacher architectures and training objectives affect the resulting gains. You et al. (2017) evaluate distillation from uniformly averaged teacher probabilities, comparing single-teacher supervision with mixtures of three and five teachers. Their full method also transfers relative dissimilarity relationships between intermediate representations. Yuan et al. (2021) compare fixed teacher mixtures with instance-dependent teacher selection for NLP tasks. During language-model domain adaptation, Qin et al. (2022) use multiple teachers, each supervising examples from its own domain. Our mixKD condition uses a fixed probability mixture of short-context and equal-context teacher distributions to investigate how teacher context composition changes helpful and harmful distant-context effects during language-model pretraining.

Context use and training tradeoffs. Context ablations reveal differences in how language models use nearby and distant text (Khandelwal et al., 2018; O'Connor & Andreas, 2021). Vakilian et al. (2026) study how much local context is needed to preserve predictions that are correct and confident under full context. LongPPL focuses evaluation on tokens identified through long–short prediction contrasts and full-context likelihood, while LongCE reweights training toward context-dependent tokens (Fang et al., 2025). In a different setting, He et al. (2026) document a reasoning–recall tradeoff during mid-training and use teacher confidence to route tokens between distillation and ordinary next-token learning. We examine contextual tradeoffs through baseline-defined helpful and harmful populations and comparisons of full- and restricted-context prediction loss.

3 Study Design and Measurements

We use short-context (sKD), equal-context (eqKD), and mixed-teacher (mixKD) distillation as experimental setups for studying learned contextual dependence. For each setup, the student retains its architecture and specified causal input. The teacher regime comprises pretraining context (H), runtime access (h), and teacher composition. We define these regimes, their shared student training objective, and measurements of prediction quality and contextual effects. Figure 1 shows the mixed-teacher setup.

3.1 Student and teacher contexts

A training block $x_{1:L+1}$ provides causal inputs $x_{1:t}$ and next-token targets $y_t = x_{t+1}$ for $t = 1, \ldots, L$. The student model M, with trainable parameters θ, is initialized from scratch and retains its specified context limit L and unchanged architecture. For a short-teacher horizon $1 \leq H < L$, define

$$\begin{aligned} s_t^{(H)} &= x_{\max(1,t-H+1):t}, \\ p_{M,t}(\cdot) &= p_M(\cdot \mid x_{1:t}; \theta), \\ q_{\mathrm{s},t}(\cdot) &= q_{\mathrm{s}}(\cdot \mid s_t^{(H)}), \qquad q_{\mathrm{eq},t}(\cdot) = q_{\mathrm{eq}}(\cdot \mid x_{1:t}). \end{aligned} \tag{1}$$

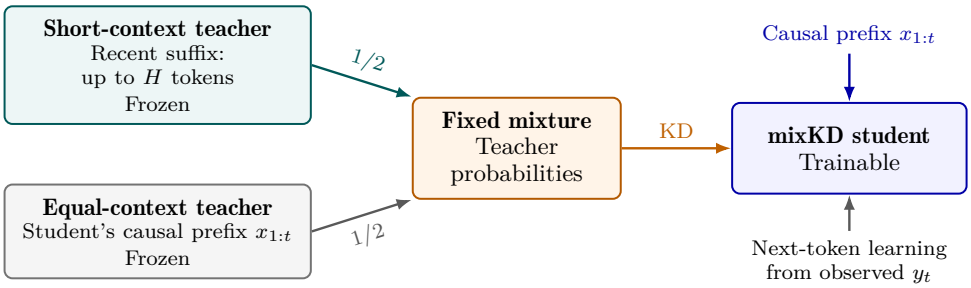

Figure 1: **Composing teacher contexts using mixKD during pretraining.** Two frozen teachers supply an equal probability mixture for distillation, alongside ordinary next-token learning. The student retains its specified causal input; only its parameters are updated, and only the student is deployed.

Here $s_t^{(H)}$ is the recent suffix containing up to H input tokens. The distributions $p_{M,t}$, $q_{\mathrm{s},t}$, and $q_{\mathrm{eq},t}$ are the next-token predictions of the student, short-context teacher, and equal-context teacher, respectively. All three distributions are defined over the same vocabulary and predict the same target $y_t = x_{t+1}$. The short-context teacher is pretrained separately on short blocks and receives only the recent suffix when supplying supervision. The equal-context teacher is the original frozen, same-seed baseline trained only on observed next tokens. Equal context means the same causal prefix as the student, and not an additional source of input. Both the teachers remain frozen during student training, and neither is a subsequently trained sKD or eqKD student. Their context access is illustrated in Figure 1.

3.2 Distillation setups and shared objective

The three setups share a probability-mixture formulation, with teacher probabilities evaluated at a common distillation temperature T:

$$q_{\lambda,t}^{(T)} = \lambda q_{\mathrm{s},t}^{(T)} + (1-\lambda) q_{\mathrm{eq},t}^{(T)}. \tag{2}$$

The superscript (T) denotes dividing logits by $T > 0$ before softmax. The mixture averages these probabilities, not logits. The observed-token cross-entropy (CE) loss uses unsoftened student probabilities ($T = 1$), preserving the baseline's CE formulation. Temperature softening is confined to the distillation terms, where we match softer teacher and student distributions.

For a batch of B examples, observed-token CE covers all L positions, while both teachers supervise the same nonempty set $\mathcal{S} \subseteq \{1, \ldots, L\}$. The implementation averages CE and each teacher-specific forward-KL term separately:

$$\begin{aligned} \mathcal{L}_{\mathrm{CE}} &= \frac{1}{BL} \sum_{b=1}^{B} \sum_{t=1}^{L} -\log p_{M,b,t}(y_{b,t}), \\ D_j &= \frac{T^2}{B|\mathcal{S}|} \sum_{b=1}^{B} \sum_{t \in \mathcal{S}} D_{\mathrm{KL}}\left(q_{j,b,t}^{(T)} \,\|\, p_{M,b,t}^{(T)}\right), \qquad j \in \{\mathrm{s}, \mathrm{eq}\}, \\ \mathcal{L}_{\lambda} &= (1-\alpha)\mathcal{L}_{\mathrm{CE}} + \alpha\left[\lambda D_{\mathrm{s}} + (1-\lambda) D_{\mathrm{eq}}\right]. \end{aligned} \tag{3}$$

Appendix A gives the algebraic identities underlying the teacher-mixture formulation.

The three setups retain the same CE/KD weighting, where sKD uses $\lambda = 1$, eqKD uses $\lambda = 0$, and mixKD uses $\lambda = 1/2$. The CE-only baseline minimizes $\mathcal{L}_{\mathrm{CE}}$ without distillation. All three distillation setups use $\alpha = 0.5$ and $T = 2$.

All three setups use the same target set $\mathcal{S}$, keeping distillation coverage fixed. Positions outside $\mathcal{S}$ still contribute to CE. Section 4 describes the training settings, and Appendix B.2 specifies the target selection and alignment.

Teachers are discarded after training, and mixKD deploys one student, not an inference ensemble of sKD and eqKD students.

3.3 THE ROLE OF PREDICTIVE ANCHORING

The short-context teacher receives only the recent H-token suffix, and therefore, cannot respond to changes in earlier text. The equal-context teacher can respond to such changes because it receives the student's specified causal input. Matching the short-context teacher therefore supplies a locally conditioned target for the student even when it sees distant text. This matching pressure is predictive anchoring. mixKD replaces part of the equal-context teacher's supervision with this predictive-anchoring pressure.

3.4 PREDICTION AND CONTEXTUAL-RELIANCE MEASUREMENTS

We probe the role of distant text in the prediction loss improvement of the distilled students over the CE-only baseline. To do this, we measure the student's prediction improvement over the baseline when both receive the same specified causal input, and the contextual-gain within a model when distant text is added to the same recent suffix. We also examine these measurements separately at positions classified as helpful, harmful, or discordant/zero using baseline contextual effects.

For the same evaluated target i, let $\ell_{M,h,i}$ and $\ell_{M,L,i}$ be model M's observed-token negative log-probabilities with the final h input tokens and the specified longer causal prefix L, respectively. The diagnostic cutoff h is distinct from the teacher's training horizon H. For each target, we define the *contextual-gain* from distant text, omitting the target index i from the notation for readability, as

$$G_M = \ell_{M,h} - \ell_{M,L}. \tag{4}$$

Positive G_M means distant text improves the observed token's prediction, while negative G_M means it worsens it. For each evaluated seed, two CE-only baselines, trained using other seeds, define fixed population labels—helpful requires both donor effects to be positive, harmful both negative, and discordant/zero contains the remaining positions. These labels describe the predictive effects of distant text in the baseline models, not its semantic relevance. They do not use the distilled students' outputs. Both metrics—prediction improvement and contextual-gain—use the same populations, with membership held fixed across methods at a given h and evaluation split.

To relate the two comparisons, we define full-input and restricted-input prediction improvements over the same-seed baseline B as $g_{\text{full},M} = \ell_{B,L} - \ell_{M,L}$ and $g_{h,M} = \ell_{B,h} - \ell_{M,h}$, respectively. Positive values mean lower loss. With $\Delta G_M = G_M - G_B$, per target and over any shared averaging population,

$$g_{\text{full},M} = g_{h,M} + \Delta G_M. \tag{5}$$

This separates restricted-input improvement from the *change* in contextual-gain. So, prediction can improve even when contextual-gain decreases.

Separately, population accounting shows *where* prediction improves or worsens and how these changes combine into the overall result. Let $c \in \{\text{help}, \text{harm}, \text{dz}\}$ denote helpful, harmful, and combined discordant/zero positions. For N evaluated targets, let $w_c = n_c/N$ be each population's share, and let $\text{CE}_{M,c}$ be model M's mean full-input loss on its n_c targets. Overall full-input CE and improvement over a comparator R satisfy

$$\begin{aligned} \text{CE}_M &= w_{\text{help}}\text{CE}_{M,\text{help}} + w_{\text{harm}}\text{CE}_{M,\text{harm}} + w_{\text{dz}}\text{CE}_{M,\text{dz}}, \\ \text{CE}_R - \text{CE}_M &= \sum_{c\in\{\text{help},\text{harm},\text{dz}\}} w_c\big(\text{CE}_{R,c} - \text{CE}_{M,c}\big). \end{aligned} \tag{6}$$

The weights sum to one and do not change across models. Each weighted difference is a *population contribution to prediction improvement*; the unweighted difference is the conditional prediction improvement. Unlike Eq. 5, this partitions targets rather than separating restricted-input prediction from contextual changes. Neither decomposition identifies a causal mechanism.

Average signed effects, mean absolute effects, and tokenwise contraction are distinct measurements. Appendix B.2 specifies the population definitions and inference-ensemble arithmetic; Appendix E reports sensitivity checks and an example distinguishing average attenuation from tokenwise contraction.

4 EXPERIMENTAL SETUP

Experimental conditions and data. We compare baseline, sKD, eqKD, and mixKD models on WikiText-103 (Merity et al., 2016) at $H = h \in \{32, 64\}$ and SlimPajama-100M (Soboleva et al., 2023) at $H = h = 64$, each across three training seeds on validation and test data. Both corpora use GPT-2 BPE tokenization. WikiText-103 training uses a subset of the source-provided training split, with a separate held-out portion of that split for internal checkpoint-selection validation. SlimPajama-100M uses a fixed 100-million-token training sample and source-provided validation for selection. Selected checkpoints are evaluated on source-provided test splits. Additional WikiText-103 evaluation uses the source-provided validation split without checkpoint reselection (Appendix D).

Matched training and checkpoint selection. We use eight-layer decoder-only Transformer models with approximately 204M total parameters (101M non-embedding parameters), where the latter excludes the input embedding and output projection. The student context length is $L = 1024$. Within each corpus, baseline and distilled students share architecture and optimization settings and start from matched fresh initialization for each seed. Baselines, short-context teachers, and distilled students follow the same *most-viable-checkpoint* (MVC) rule, selecting the lowest all-position validation CE without the KD term. Runs stop after five consecutive epochs without improvement. Under the shared MVC rule, distilled students reached their best validation CE at later epochs than the CE-only baselines. Their selected checkpoints therefore reflect greater training exposure. Appendix B documents the full protocol, exposure, and compute accounting.

Evaluation targets. Reported prediction losses use the same target positions across matched methods. Appendix B.2 specifies their selection and alignment with distillation supervision.

Uncertainty and evaluation protocol. Pointwise 95% intervals use block bootstrapping, paired across methods for predictive contrasts. They quantify evaluation uncertainty conditional on trained checkpoints, not training-seed variability. Test data were not used for training or checkpoint selection. Primary test comparisons retained fixed checkpoints, mixture weights, and population rules. Appendix E reports additional post-test checks on WikiText-103, examining sensitivity to how helpful and harmful positions are labeled and to possible dependence between neighboring evaluation blocks.

5 PREDICTION QUALITY AND CONTEXTUAL DEPENDENCE

5.1 BETTER PREDICTION WITH WEAKER CONTEXTUAL EFFECTS

Across WikiText-103 at $H = h \in \{32, 64\}$ and SlimPajama-100M at $H = h = 64$, sKD lowers full-input CE relative to the CE-only baseline in all three seeds on validation and test data, while eqKD achieves lower loss still (Table 1; Appendix C.1). These overall improvements conceal different outcomes across positions. At $H = h = 64$ in both corpora, sKD predicts worse than the baseline on helpful positions, despite improving on harmful and discordant/zero positions. With the same full inputs, eqKD predicts better on helpful positions and worse on harmful positions than sKD (Figure 2). In summary, although eqKD has lower overall CE than sKD, sKD predicts better on harmful positions.

To examine how distant text contributes to these predictions, we compare each model's loss with and without that text. At $H = h = 64$ in both corpora, sKD receives a smaller average benefit from distant text than the baseline on helpful positions. On harmful positions, sKD also incurs a smaller average penalty from distant text. eqKD retains more of both effects than sKD (Figure 3). The student retains the same input access as the baseline. Thus, **sKD improves overall prediction while reducing both the average benefit and the average penalty from distant text**.

5.2 SIMILAR OVERALL LOSSES CAN ACCOMPANY DIFFERENT CONTEXTUAL RESPONSES

Following the sKD and eqKD comparisons, we introduced mixKD using a fixed equal mixture of the two frozen teacher distributions. mixKD and eqKD have similar overall CE on validation and test data, with mixKD achieving lower loss in most comparisons but not all (Table 1; Appendices C.1

Table 1: Full-input test CE and mixKD CE improvement over eqKD, in nats per evaluated target. Bold denotes the lowest point-estimated CE in each row. Improvement, computed before rounding, is $\mathrm{CE}_{\mathrm{eqKD}} - \mathrm{CE}_{\mathrm{mixKD}}$, so positive values favor mixKD. † marks improvements whose pointwise paired-block 95% CI includes zero; full intervals appear in Appendix Table 11. WikiText-103 baseline/eqKD scores and test targets are shared across horizons.

Corpus	H	Seed	Baseline	sKD	eqKD	mixKD	mixKD CE improvement over eqKD
WikiText-103	32	42	3.2299	3.1846	**3.0904**	3.0909	-0.0004†
WikiText-103	32	99	3.2315	3.1748	3.0912	**3.0869**	+0.0042
WikiText-103	32	123	3.2338	3.1853	3.0979	**3.0948**	+0.0031†
WikiText-103	64	42	3.2299	3.1545	3.0904	**3.0815**	+0.0089
WikiText-103	64	99	3.2315	3.1560	3.0912	**3.0813**	+0.0099
WikiText-103	64	123	3.2338	3.1580	3.0979	**3.0881**	+0.0098
SlimPajama-100M	64	42	3.8087	3.7685	**3.6859**	3.6880	-0.0021
SlimPajama-100M	64	99	3.8022	3.7685	3.6845	**3.6763**	+0.0082
SlimPajama-100M	64	123	3.8164	3.7834	3.6918	**3.6815**	+0.0103

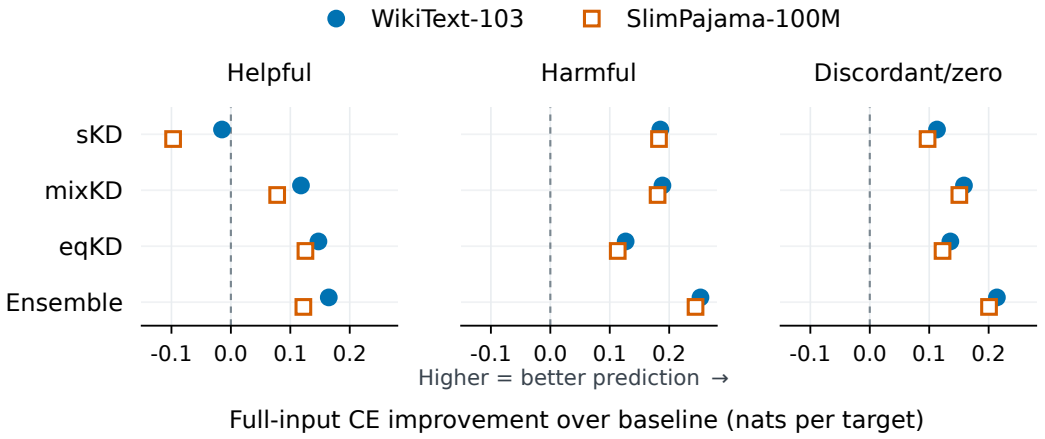

Figure 2: Full-input CE improvement over the same-seed baseline at $H = h = 64$ on source-provided test splits. Ensemble averages the prediction probabilities of same-seed trained sKD and eqKD students with equal weights. Each marker averages the population-specific improvements for seeds 42, 99, and 123 with equal weight; targets are not pooled. Higher values mean better prediction relative to baseline, not higher CE. Colors and shapes distinguish corpora, with slight vertical offsets for visibility. No intervals are plotted. Appendix Table 14 retains per-seed student estimates and paired-block 95% intervals; Appendix C.4 reports ensemble comparisons.

and D). Their overall test losses are approximately 3.1–3.7 nats per target. Their largest difference in test CE is below 0.011 nats per target, which is less than 0.4% of either model's loss.

Despite their similar overall losses, the students differ in prediction performance on helpful and harmful positions. mixKD predicts better than eqKD on harmful positions, while eqKD remains better on helpful positions. Compared with sKD, mixKD predicts better on helpful positions, closing much of the gap to eqKD (Figure 2). These population-level prediction comparisons hold across all three seeds on validation and test data, at all evaluated horizons for each corpus (Appendices C.2 and D).

To examine the contribution of distant text, we next compare each student's loss with and without it. At $H = h = 64$ in both corpora, mixKD receives a greater average benefit from distant text

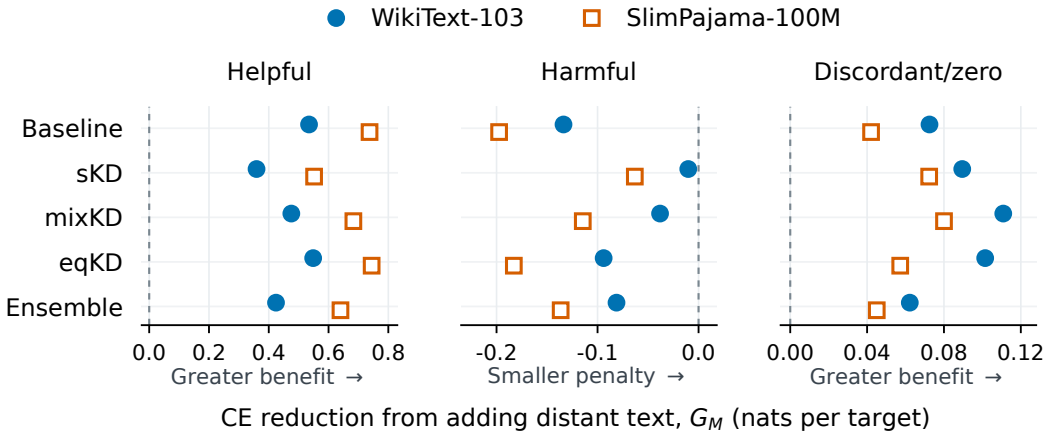

Figure 3: CE reduction from adding distant text, measured by contextual-gain G_M at $H = h = 64$ on the same test populations. Each marker averages the three seed-specific population means with equal weight. Positive values mean benefit and negative values mean a penalty. Farther right means a greater benefit or smaller penalty, not necessarily better prediction. The equal-weight student ensemble, corpus symbols, and offsets follow Figure 2; horizontal scales differ by population. No intervals are plotted. Per-seed individual-model estimates and 95% intervals appear in Appendix Tables 22 and 23, with ensemble comparisons in Table 28.

on helpful positions than sKD. On harmful positions, mixKD incurs a smaller average penalty from distant text than eqKD (Figure 3). Thus, **mixKD and eqKD have similar overall prediction losses but respond differently to distant text**.

5.3 Mixing during training versus averaging at inference

mixKD averages the short- and equal-context teachers' predictions with equal weights during training. We compare this with an ensemble that averages the separately trained sKD and eqKD students' predictions with equal weights at inference. The ensemble achieves lower test CE than mixKD overall and within all three populations, but mixKD obtains greater average helpful-context benefit and a smaller harmful-context penalty. Figures 2 and 3 show the population-level contrasts at $H = h = 64$; the findings hold in all nine comparisons (Appendix C.4). However, the equal-weight ensemble evaluates two students per prediction, whereas mixKD evaluates one. To test whether the contextual difference depends on equal weighting, an exploratory post-test analysis varied the averaging weights for these same students' probabilities, allowing each setting its own choice. No fixed weighting in any of the nine settings simultaneously achieved at least mixKD's average helpful-context benefit and at most its average harmful-context penalty (Figure 4; Appendix C.4.1). Thus, **the ensemble predicts better, but averaging these students' predictions with fixed weights does not reproduce mixKD's joint helpful–harmful contextual response**.

The ensemble predicts better than mixKD under both full and restricted input. On helpful and harmful positions, its advantage is larger under restricted input. Its lower contextual-gain therefore reflects the change from an already stronger restricted-input prediction, rather than worse full-input prediction (Appendix C.4).

6 Discussion

The three distillations (sKD, eqKD, and mixKD) differ in prediction quality and their responses to distant text, while the student architecture and input access remain unchanged. All three lower the overall prediction loss relative to baseline. sKD reduces both average helpful and average harmful effects from text outside its teacher's short-context input, whereas eqKD retains more of both these effects. mixKD learns under joint supervision from short-context and equal-context teachers to

predict better than sKD on helpful positions and better than eqKD on harmful positions, while also achieving lower overall loss than eqKD in most, but not all, comparisons.

In a practical use case, an ideal document assistant, while answering a new question, should use relevant customer-specific information supplied earlier without being distracted by unrelated material. This motivates examining the influence of distant text, along the lines of the context-ablation approach discussed in the introduction (O'Connor & Andreas, 2021). Contextual-gain measures this influence by comparing a model's full-input loss with its own restricted-input loss. We observe that mixKD receives a greater average benefit than sKD on helpful positions and a smaller average penalty than eqKD on harmful positions, when using distant text. Together with the prediction results, these measurements describe a compromise between sKD and eqKD, rather than uniformly outperforming both.

Figures 2 and 3 show why contextual-gain is not another performance ranking. The equal-weight student ensemble predicts better than mixKD within all three populations, yet mixKD obtains greater helpful-context benefit and a smaller harmful-context penalty. Contextual-gain compares each model with its own restricted-input prediction, which is also stronger for the ensemble (Appendix C.4). The joint helpful–harmful response therefore distinguishes the models without implying a prediction advantage for mixKD.

Predictive anchoring offers one explanation for the observed contextual effects in sKD and mixKD students. The short-context teacher's predictions cannot respond to distant text, but the student must balance matching this local predictive target against learning from observed tokens using the full-context. **We hypothesize that this raises the burden on distant information to overturn a locally supported prediction.** With the mixKD student held fixed, we compare its output scores with and without distant text. Short-context guidance opposes this change at more target positions than equal-context guidance (Appendix F, Table 41). Restricting the equal-context teacher's input to recent text also increases this opposition, without changing its weights (Table 42). This isolates an immediate input-restriction effect and supports the anchoring interpretation.

The opposition effect observed with both teachers under restricted input raises the question of whether the teacher's pretraining context still matters. In a supplementary WikiText-2 study, we compare teachers pretrained with different context lengths while keeping their input length fixed during distillation (Appendix G.2). Longer contexts during teacher pretraining sometimes improve student prediction, while the students' average contextual-gain remains below that of the baseline. Thus, the option to restrict a teacher's input at runtime does not mean that its pretraining context can be ignored. The teacher's pretraining context length also influences how well the student predicts.

How these behavioral differences develop during training is beyond the scope of this study, as are the individual roles of teacher quality, confidence, and disagreement. In particular, whether interactions between teacher losses explain mixKD's joint contextual response requires a controlled investigation.

A future direction is selective anchoring, where a reliable external signal adjusts the strength of local supervision. Whether this would preserve task-relevant information while resisting distraction remains untested.

7 Conclusion

For a fixed student architecture and input access, teacher context regimes and composition can change how students respond to distant text. In our main comparisons, all three students improve overall prediction over the baseline, with mixKD and eqKD achieving similar losses below sKD's. Under predictive anchoring from its short-context teacher, sKD attenuates both helpful and harmful contextual effects on average. eqKD retains more of both, while mixKD combines greater benefit from helpful context than sKD with a smaller penalty from harmful context than eqKD. The mechanisms leading to these differences and their implications for semantic evidence use remain open questions. Our findings show that similar prediction loss can accompany different learned contextual behavior, giving us a reason to examine contextual response alongside overall loss when studying language-model pretraining.

AI USE STATEMENT

We have used generative AI tools for research ideation and execution, literature retrieval and discovery, and manuscript preparation. These include ChatGPT and Codex, while Gemini and Antigravity were used during earlier development. Research assistance includes refining questions and hypotheses, developing the conceptual framing, formulating and explaining mathematical objectives and diagnostics, providing feedback on experimental design, assisting with implementation and debugging of experimental and analysis code, and interpreting results. Literature assistance includes identifying relevant work and summarizing sources. Writing assistance includes proposing manuscript structure, drafting sections, and revising prose for clarity and readability. We have also used AI assistance to create method schematics and prepare LaTeX tables and plots from recorded experimental outputs.

The main training and evaluation used existing text corpora, not synthetic datasets generated for this study. Exploratory synthetic diagnostic probes were unsuccessful and are excluded from the reported evidence; they provide no support for the paper's conclusions.

The authors have directed the research and guided the review and revision of AI-assisted drafts. Verification includes cross-checking numerical claims and tables against saved experimental artifacts and checking computational work through automated tests and numerical consistency checks. AI-generated suggestions were treated as proposals to evaluate, not as independent evidence for scientific claims. The authors take responsibility for the accuracy, originality, and integrity of the final manuscript and associated research artifacts, including material developed with AI assistance.

REPRODUCIBILITY STATEMENT

Section 3 specifies the training objectives and measurements, with teacher-mixture identities in Appendix A. Section 4 and Appendix B document dataset construction and splits, tokenization, architectures, optimization, random seeds, checkpoint selection, and target alignment, together with realized training exposure and hardware/runtime accounting. Per-seed results and paired comparison intervals for the main experiments are reported in Appendices C and D. Appendix E documents population-definition and resampling checks, and Appendix F specifies the predictive-anchoring diagnostic and its sampling procedure. The supplementary WikiText-2 setup and results appear in Appendix G. A reproducibility kit is included as a separate attachment.

ETHICS STATEMENT

This study uses existing WikiText and SlimPajama-derived corpora, with data sources and preparation documented in Appendix B. These corpora may contain biased, offensive, or personal content; our experiments do not establish that distillation mitigates these risks. The terms "helpful" and "harmful" describe effects on next-token prediction loss, not judgments about content safety or social harm. The observed contextual effects do not establish improvements in fairness, privacy, or safe deployment. The document-assistance example motivates the study but is not an evaluated application.

REFERENCES

Yevgen Chebotar and Austin Waters. Distilling knowledge from ensembles of neural networks for speech recognition. In *Interspeech 2016*, pp. 3439–3443, 2016. doi: 10.21437/Interspeech.2016-1190. URL https://www.isca-archive.org/interspeech_2016/chebotar16_interspeech.html.

Guanzheng Chen, Xin Li, Michael Shieh, and Lidong Bing. LongPO: Long context self-evolution of large language models through short-to-long preference optimization. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=qTrEq31Shm.

Kevin Clark, Minh-Thang Luong, Christopher D. Manning, and Quoc Le. Semi-supervised sequence modeling with cross-view training. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1914–1925, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1217. URL https://aclanthology.org/D18-1217/.

Zican Dong, Junyi Li, Jinhao Jiang, Mingyu Xu, Wayne Xin Zhao, Bingning Wang, and Weipeng Chen. LongReD: Mitigating short-text degradation of long-context large language models via restoration distillation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10687–10707, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.524. URL https://aclanthology.org/2025.acl-long.524/.

Tianqi Du, Haotian Huang, Yifei Wang, and Yisen Wang. Long-short alignment for effective long-context modeling in LLMs. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 14515–14535. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/du25a.html.

Lizhe Fang, Yifei Wang, Zhaoyang Liu, Chenheng Zhang, Stefanie Jegelka, Jinyang Gao, Bolin Ding, and Yisen Wang. What is wrong with perplexity for long-context language modeling? In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=fL4qWkSmtM.

Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 1607–1616. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/furlanello18a.html.

Yuxian Gu, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. MiniPLM: Knowledge distillation for pre-training language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=tJHDw8XfeC.

Jacqueline He, Howard Yen, Shuyue Stella Li, Margaret Li, Hanqing Zeng, Yinglong Xia, Benyu Zhang, Zhuokai Zhao, Qiang Zhang, Pang Wei Koh, Luke Zettlemoyer, and Wen-tau Yih. Knowledge distillation during mid-training favors reasoning over factual recall. arXiv preprint arXiv:2609.01532, 2026. URL https://arxiv.org/abs/2609.01532.

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.

Urvashi Khandelwal, He He, Peng Qi, and Dan Jurafsky. Sharp nearby, fuzzy far away: How neural language models use context. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 284–294, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1027. URL https://aclanthology.org/P18-1027/.

Javid Lakha, Nihal V. Nayak, Bingbin Liu, Sham M. Kakade, and David Alvarez-Melis. Self-distillation for data-scarce language model pretraining. In *High-dimensional Learning Dynamics 2026*, 2026. URL https://openreview.net/forum?id=n6LfAFzq6u.

Hayeon Lee, Rui Hou, Jongpil Kim, Davis Liang, Sung Ju Hwang, and Alexander Min. A study on knowledge distillation from weak teacher for scaling up pre-trained language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 11239–11246, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.714. URL https://aclanthology.org/2023.findings-acl.714/.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. arXiv preprint arXiv:1609.07843, 2016. URL https://arxiv.org/abs/1609.07843.

Usama Muneeb and Mesrob I. Ohannessian. Induced model matching: Restricted models help train full-featured models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 62617–62647. Curran Associates, Inc., 2024. doi: 10.52202/079017-2002. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/7279908471a7dd4898d2715f7c6a7413-Paper-Conference.pdf.

Joe O'Connor and Jacob Andreas. What context features can transformer language models use? In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 851–864, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.70. URL https://aclanthology.org/2021.acl-long.70/.

Yujia Qin, Yankai Lin, Jing Yi, Jiajie Zhang, Xu Han, Zhengyan Zhang, Yusheng Su, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. Knowledge inheritance for pre-trained language models. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3921–3937, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.288. URL https://aclanthology.org/2022.naacl-main.288/.

Charlie Snell, Dan Klein, and Ruiqi Zhong. Learning by distilling context. arXiv preprint arXiv:2209.15189, 2022. URL https://arxiv.org/abs/2209.15189.

Daria Soboleva, Faisal Al-Khateeb, Robert Myers, Jacob R Steeves, Joel Hestness, and Nolan Dey. SlimPajama: A 627B token cleaned and deduplicated version of RedPajama. https://cerebras.ai/blog/slimpajama-a-627b-token-cleaned-and-deduplicated-version-of-redpajama, June 2023. URL https://huggingface.co/datasets/cerebras/SlimPajama-627B.

Huashan Sun, Shengyi Liao, Yansen Han, Yu Bai, Yang Gao, Cheng Fu, Weizhou Shen, Fanqi Wan, Ming Yan, Ji Zhang, and Fei Huang. SoLoPO: Unlocking long-context capabilities in LLMs via short-to-long preference optimization. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=iiBjaiikJG.

Vala Vakilian, Zimeng Wang, Ankit Singh Rawat, and Christos Thrampoulidis. Short-context dominance: How much local context natural language actually needs? arXiv preprint arXiv:2512.08082, 2026. URL https://arxiv.org/abs/2512.08082.

Shan You, Chang Xu, Chao Xu, and Dacheng Tao. Learning from multiple teacher networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '17, pp. 1285–1294, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450348874. doi: 10.1145/3097983.3098135. URL https://doi.org/10.1145/3097983.3098135.

Fei Yuan, Linjun Shou, Jian Pei, Wutao Lin, Ming Gong, Yan Fu, and Daxin Jiang. Reinforced multi-teacher selection for knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 14284–14291, 2021. doi: 10.1609/aaai.v35i16.17680. URL https://ojs.aaai.org/index.php/AAAI/article/view/17680.

Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3903–3911, June 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Yuan_Revisiting_Knowledge_Distillation_via_Label_Smoothing_Regularization_CVPR_2020_paper.html.

Xinsen Zhang, Zhenkai Ding, Tianjun Pan, Run Yang, Chun Kang, Xue Xiong, and Jingnan Gu. OPSDL: On-policy self-distillation for long-context language models. arXiv preprint arXiv:2604.17535, 2026. URL https://arxiv.org/abs/2604.17535.

APPENDIX GUIDE

The appendices follow the study from objective definitions and experimental protocol to detailed results, robustness checks, and supplementary diagnostics. Section titles and page numbers below link to the corresponding material.

A Teacher-mixture identities 15

B Experimental protocol and training exposure 15
B.1 Data, models, and checkpoint selection 15
B.2 Target alignment and measurement conventions 17
B.3 Realized training exposure 18
B.4 Hardware and recorded training time 20

C Primary prediction and contextual-effect results 22
C.1 Overall prediction 22
C.2 Conditional prediction and population contributions 24
C.2.1 Population-weighted prediction contributions 28
C.3 Contextual effects and restricted-input accounting 28
C.3.1 Restricted-input and contextual-change accounting 31
C.4 Prediction-ensemble comparison 31
C.4.1 Exploratory fixed-weight ensemble analysis 35

D Additional source-provided WikiText-103 validation 38
D.1 Prediction and population accounting 38
D.2 Contextual effects and ensemble boundary 40

E Population and dependence robustness 41
E.1 Label rules and donor agreement 41
E.2 Label-independent magnitude and tokenwise contraction 44
E.3 Block dependence and interval interpretation 44

F Predictive-anchoring diagnostic 46

G Supplementary WikiText-2 comparisons 48
G.1 Reversal of the sKD–eqKD ordering 48
G.2 Separating teacher pretraining horizon from runtime access 48

The source-provided validation evaluation in Appendix D uses fixed checkpoints and is distinct from checkpoint-selection validation. For the teacher-regime interpretation, read the final-checkpoint diagnostic in Appendix F alongside the supplementary training comparison in Appendix G.2.

A TEACHER-MIXTURE IDENTITIES

Section 3.2 describes mixKD using a mixture of teacher predictions, while the implementation computes a weighted sum of two teacher-specific KD losses. We first explain why these two formulations provide the same guidance to the student during training. We then examine how the mixture target responds when distant text changes but the recent suffix stays the same.

Why the two loss formulations give the same student gradients. Consider one target position at a fixed distillation temperature T. Let p be the student's next-token distribution, and let q_{s} and q_{eq} be the distributions from the short-context and equal-context teachers. All three distributions use temperature T; we omit the temperature superscript and target index here to keep the notation readable. The mixture assigns weight λ to the short-context teacher and $1-\lambda$ to the equal-context teacher, giving $q_\lambda = \lambda q_{\mathrm{s}} + (1-\lambda) q_{\mathrm{eq}}$.

We can either compare the student with each teacher separately and weight the two KL losses, or compare it directly with the mixture. Expanding the KL terms and collecting those that depend on p gives

$$(1-\lambda) D_{\mathrm{KL}}(q_{\mathrm{eq}} \| p) + \lambda D_{\mathrm{KL}}(q_{\mathrm{s}} \| p) = D_{\mathrm{KL}}(q_\lambda \| p) + C, \tag{7}$$

where $C = (1-\lambda) D_{\mathrm{KL}}(q_{\mathrm{eq}} \| q_\lambda) + \lambda D_{\mathrm{KL}}(q_{\mathrm{s}} \| q_\lambda)$. This additional term depends only on the teacher predictions and their mixture. For a given input, frozen teachers and fixed mixture weights keep C unchanged as the student's parameters change. The numerical losses can therefore differ, but their gradients with respect to the student parameters are identical.

The same reasoning applies when we average over the shared distilled positions and batch, and multiply both formulations by the same T^2 factor. The observed-token CE term is unchanged, so either formulation supplies the same total gradient to the student. This equivalence concerns two ways of computing the training objective. It does not make the resulting mixKD student an inference ensemble of separately trained sKD and eqKD students.

How distant text changes the mixture target. Now consider two input prefixes of the same length whose final H tokens are identical, but whose earlier text differs. We keep the teacher parameters, mixture weights, temperature, and target position fixed. The short-context teacher receives the same recent suffix in both cases, so its predictions do not change. The equal-context teacher receives the earlier text as well, so its predictions can change.

Let $\Delta_{\mathrm{remote}} q$ denote the difference between a teacher's predicted probability distributions for these two inputs. Because the mixture averages teacher probabilities, only the equal-context teacher contributes to this difference,

$$\Delta_{\mathrm{remote}} q_{\mathrm{s}}^{(T)} = 0, \qquad \Delta_{\mathrm{remote}} q_\lambda^{(T)} = (1-\lambda) \Delta_{\mathrm{remote}} q_{\mathrm{eq}}^{(T)}. \tag{8}$$

For mixKD, $\lambda = 1/2$, so the change in each vocabulary token's probability in the mixture is exactly half its change in the equal-context teacher's prediction. This is a property of the supervision supplied to the student, not a prediction that its learned response to distant text will also be halved. The student also learns from observed next tokens, and its response develops through optimization. The identities above therefore describe the training signal; the student's prediction loss, representations, and contextual effects require separate investigation.

B EXPERIMENTAL PROTOCOL AND TRAINING EXPOSURE

B.1 DATA, MODELS, AND CHECKPOINT SELECTION

WikiText-103 partitions. The recorded preparation concatenates the train, validation, and test splits of `Salesforce/wikitext` (Merity et al., 2016), configuration `wikitext-103-raw-v1`, in that order, retaining a newline after each text row. GPT-2 BPE tokenization gives $N = 118{,}511{,}293$ tokens. Training uses the first $\lfloor .8N \rfloor = 94{,}809{,}034$ tokens, internal checkpoint-selection validation the following $\lfloor .1N \rfloor$, and internal held-out evaluation the remainder. Both training and internal validation lie within the source-provided training text. This is not the original benchmark partitioning.

The standalone source-provided test file contains 283,427 tokens and byte-exactly matches the suffix of the concatenated file. The source-provided validation file contains 247,417 tokens and matches

the immediately preceding span. Both lie in the historical internal held-out tail, not in training or checkpoint-selection validation. Evaluating these files without another split therefore gives fixed-checkpoint evaluations on subsets of the historical held-out text, not new disjoint corpora or a claim of never-evaluated text. Source identities and hashes are preserved in the accompanying provenance records.

SlimPajama-100M sample and splits. The fixed training sample contains 100,000,000 GPT-2 BPE tokens from 95,017 documents in `DKYoon/SlimPajama-6B` (Soboleva et al., 2023), revision `b5f90f419b7489cdba26fdbc8c022fcb5562f968`. A seeded document-priority procedure (seed 20260909) hashes shard/row identities, selects a candidate pool, and concatenates documents in ascending priority order. Only the final document is truncated; documents are separated by two newlines (token 628). Actual sample proportions are 49.06% Common Crawl, 25.74% C4, 6.97% GitHub, 5.10% arXiv, 4.73% Wikipedia, 4.23% books, and 4.17% Stack Exchange. Source disjointness from WikiText is not established. Source-provided validation and test contain 9,479,563 and 9,639,564 tokens and are not repartitioned. Both corpora use GPT-2 BPE with vocabulary size 50,257.

Table 2: Main comparison grid. Each row covers baseline, sKD, eqKD, and mixKD at seeds 42, 99, and 123. WikiText-103 baseline/eqKD checkpoints and full-input predictions are shared across horizons. Additional source-provided WikiText-103 validation evaluation is reported in Appendix D.

Corpus	$H = h$	Evaluation splits
WikiText-103	32, 64	Internal validation; source-provided test
SlimPajama-100M	64	Source-provided validation and test

Table 3: Architecture and optimization settings for main students. Non-embedding counts exclude the input embedding and output projection, including its bias. The supplementary WikiText-2 configuration is given in Appendix G.

Setting	Recorded value
Decoder blocks / width / attention heads	8 / 1024 / 16
Total / non-embedding parameters	203,748,433 / 100,771,840
Student input context	1024 tokens
Normalization / positions	Pre-layer norm; final layer norm / RoPE
Feed-forward network / dropout	GELU, width $4d$ / zero
Input embedding and output classifier	Separate matrices, not weight-tied
Optimizer / initial learning rate	AdamW / 3×10^{-4}
Weight decay / learning-rate schedule	0.01 / cosine to 10^{-6}
Maximum epochs / early-stopping patience	300 / 5
Student sequences per rank / accumulation	8 / 8 steps
Training ranks / tokens per full student update	4 / 262,144
Distillation α / temperature T	0.5 / 2
Short-teacher weight in mixKD	$1/2$
KD positions / CE positions per block	897 / 1024

Teacher origins and selection. Short teachers are pretrained on observed next tokens in blocks of their own horizon, using the corpus-specific architecture. eqKD uses the same-seed CE-only baseline teacher; mixKD combines it with the corresponding original short teacher. Teacher weights are frozen before student training. Students start from matched fresh initialization, not teacher checkpoints. Every condition follows the most-viable-checkpoint (MVC) rule, selecting minimum all-position validation CE without the KD term and restoring that checkpoint after five epochs without

improvement. Student schedules use 370 warmup steps on WikiText-103 and 390 on SlimPajama-100M. Teacher-window batching during KD is distinct from the student optimizer batch.

Table 4: Common-tail evaluation populations per seed and method. Each non-overlapping 1024-token block supplies 897 shifted targets. WikiText-103 full-input targets are shared across horizons, while restricted inputs and diagnostic populations change with h.

Corpus	Evaluation split	Blocks	Targets
WikiText-103	Internal validation	11,573	10,380,981
WikiText-103	Source-provided validation	241	216,177
WikiText-103	Source-provided test	276	247,572
SlimPajama-100M	Source-provided validation	9,257	8,303,529
SlimPajama-100M	Source-provided test	9,413	8,443,461

B.2 Target alignment and measurement conventions

For a block $x_{1:L+1}$, prediction position t follows input x_t and predicts x_{t+1}. The student and equal-context teacher receive the causal prefix $x_{1:t}$. The short teacher physically receives the suffix $x_{t-H+1:t}$ at each supervised position. All predictions in a KL term concern the same next-token target. The shared set $\mathcal{S} = \{128, \ldots, 1024\}$ corresponds to zero-based output-logit indices 127–1023 and target offsets 128–1024. It accommodates full windows through $H = 128$ and was retained for $H = 32, 64$. The equal-context teacher uses these same 897 positions; its prefix lengths vary from 128 to 1024, rather than supplying 1024 preceding tokens at every position. CE covers all 1024 shifted targets. Equation 3 averages CE over BL and KD over $B|\mathcal{S}|$ separately. Positions outside the KD tail still receive CE training and can influence later outputs through causal attention.

For evaluated seed s, labels use only the other two baseline seeds. Helpful membership requires both donor G values to be strictly positive, harmful membership both strictly negative, and discordant/zero includes the entire complement. The minimum-consensus strength is the shared sign multiplied by the smaller absolute donor effect. Masks are shared across methods within a seed, horizon, and split, but need not coincide across those settings. Donor agreement does not guarantee the evaluated seed's baseline has the same sign at every selected position. These are reproducibility-based predictive labels, not semantic annotations.

Prediction-ensemble arithmetic. The inference comparator averages the unsoftened probabilities of trained sKD and eqKD students, $p_{\text{ens}} = (p_{\text{sKD}} + p_{\text{eqKD}})/2$. For target losses ℓ_{sKD} and ℓ_{eqKD}, its loss is $-\log[(e^{-\ell_{\text{sKD}}} + e^{-\ell_{\text{eqKD}}})/2]$, evaluated stably using log-sum-exp. This is not the average of their losses and is distinct from mixing the original frozen teachers during mixKD training.

Evaluation history and uncertainty. The trainers reported held-out test metrics after loading validation-selected weights. Primary test comparisons retained those checkpoints, mixture weights, and population rules without test-based checkpoint reselection or tuning. The WikiText-103 comparisons at $H = h = 32$ preceded those at $H = h = 64$; reporting them jointly does not imply prespecification. The label-rule and dependence checks in Appendix E are post-test sensitivities and were not repeated on SlimPajama-100M. The ensemble-weight search in Appendix C.4.1 is a separate exploratory post-test analysis; its descriptive weight selection does not change checkpoints or training weights.

Primary pointwise 95% intervals use 1,000 resamples of non-overlapping 1024-token blocks, paired across methods for differences. They quantify evaluation uncertainty conditional on trained checkpoints, not training-seed variability. Seeds and evaluation splits are not pooled. Prediction-improvement intervals retain their original statistic-specific attribution draws; ensemble differences retain their separately recorded paired loss-difference draws. Individual contextual means retain their own intervals. A paired interval is never obtained by subtracting marginal interval endpoints. These original intervals remain unchanged. The exploratory ensemble-weight analysis uses 2,000

additional paired resamples for approximate simultaneous bands on its evaluated grids, as specified in Appendix C.4.1.

B.3 Realized training exposure

Tables 5 and 6 distinguish exposure at the selected checkpoint from the complete run. The most-viable-checkpoint (MVC) rule selects the minimum all-position validation CE, excluding the KD term. Selected epochs are the minima in the saved logs; stopping epochs are the final logged epochs. Every run continued for five epochs without improvement after its selected checkpoint, and none reached the 300-epoch limit. The records cover all 30 distinct baseline and distilled-student checkpoints in the main comparisons. WikiText-103 baseline and eqKD checkpoints are shared across $H \in \{32, 64\}$ comparisons and appear only once. The nine separately pretrained short teachers are reported below in Table 7.

Counting convention. Token totals count training inputs, equivalently all-position CE targets, across four ranks, including repeated passes over the same training sample. The baseline/student tables exclude short-teacher pretraining, which is counted separately in Table 7. All tables exclude validation/test evaluation and frozen-teacher forward passes during student training. Epochs are read from the saved logs and cross-checked against resolved results where available. Token and optimizer-update counts are reconstructed from those epochs, recorded batch settings, and archived trainer code. Update counts denote optimizer-step attempts; the logs do not distinguish successful parameter updates from steps skipped by mixed-precision loss scaling.

For baselines and distilled students, each epoch processes 94,797,824 tokens on WikiText-103 and 99,975,168 on SlimPajama-100M after incomplete global sequence batches are dropped. With eight sequences per rank and eight accumulation steps, this gives 362 and 382 update attempts per epoch, respectively. Full updates cover 262,144 tokens. The final update in each epoch instead covers 163,840 or 98,304 tokens, so multiplying every update by the full-update batch would overstate exposure.

Table 5: WikiText-103 training exposure. Each pair reports the selected checkpoint / complete stopped run. Tokens are billions of student training inputs, rounded to three decimals. Update attempts include the partial final accumulation of each epoch.

Condition	Seed	Epochs	Update attempts	Tokens (10^9)
Baseline	42	9 / 14	3,258 / 5,068	0.853 / 1.327
Baseline	99	9 / 14	3,258 / 5,068	0.853 / 1.327
Baseline	123	9 / 14	3,258 / 5,068	0.853 / 1.327
sKD $H = 32$	42	17 / 22	6,154 / 7,964	1.612 / 2.086
sKD $H = 32$	99	21 / 26	7,602 / 9,412	1.991 / 2.465
sKD $H = 32$	123	18 / 23	6,516 / 8,326	1.706 / 2.180
sKD $H = 64$	42	17 / 22	6,154 / 7,964	1.612 / 2.086
sKD $H = 64$	99	17 / 22	6,154 / 7,964	1.612 / 2.086
sKD $H = 64$	123	16 / 21	5,792 / 7,602	1.517 / 1.991
eqKD	42	19 / 24	6,878 / 8,688	1.801 / 2.275
eqKD	99	18 / 23	6,516 / 8,326	1.706 / 2.180
eqKD	123	18 / 23	6,516 / 8,326	1.706 / 2.180
mixKD $H = 32$	42	17 / 22	6,154 / 7,964	1.612 / 2.086
mixKD $H = 32$	99	18 / 23	6,516 / 8,326	1.706 / 2.180
mixKD $H = 32$	123	17 / 22	6,154 / 7,964	1.612 / 2.086
mixKD $H = 64$	42	17 / 22	6,154 / 7,964	1.612 / 2.086
mixKD $H = 64$	99	19 / 24	6,878 / 8,688	1.801 / 2.275
mixKD $H = 64$	123	17 / 22	6,154 / 7,964	1.612 / 2.086

Table 6: SlimPajama-100M training exposure, using the same selected / stopped convention. The 100-million-token sample is reused across epochs; it is not the cumulative training-token budget.

Condition	Seed	Epochs	Update attempts	Tokens (10^9)
Baseline	42	8 / 13	3,056 / 4,966	0.800 / 1.300
Baseline	99	8 / 13	3,056 / 4,966	0.800 / 1.300
Baseline	123	8 / 13	3,056 / 4,966	0.800 / 1.300
sKD $H = 64$	42	18 / 23	6,876 / 8,786	1.800 / 2.299
sKD $H = 64$	99	15 / 20	5,730 / 7,640	1.500 / 2.000
sKD $H = 64$	123	15 / 20	5,730 / 7,640	1.500 / 2.000
eqKD	42	15 / 20	5,730 / 7,640	1.500 / 2.000
eqKD	99	15 / 20	5,730 / 7,640	1.500 / 2.000
eqKD	123	15 / 20	5,730 / 7,640	1.500 / 2.000
mixKD $H = 64$	42	15 / 20	5,730 / 7,640	1.500 / 2.000
mixKD $H = 64$	99	17 / 22	6,494 / 8,404	1.700 / 2.199
mixKD $H = 64$	123	15 / 20	5,730 / 7,640	1.500 / 2.000

WikiText-103 baselines were selected at epoch 9, compared with epochs 16–21 for distilled students. SlimPajama-100M baselines were selected at epoch 8, compared with epochs 15–18 for distilled students. Exposure also differs among distilled conditions. These are comparisons under a shared validation-selection and early-stopping procedure, not equal-token-budget comparisons. They do not isolate a regularization mechanism or the effect of additional training exposure.

Short-teacher pretraining. The six WikiText-103 teachers ($H = 32, 64$ at seeds 42/99/123) and three SlimPajama-100M teachers ($H = 64$ at the same seeds) use the same corpus-specific training and validation partitions as the students. Each is an eight-layer, 203,748,433-parameter causal Transformer trained on CE alone in non-overlapping blocks of its own horizon. All positions within each short block contribute to CE. The teachers use AdamW with initial learning rate 3×10^{-4}, weight decay 0.01, and a cosine schedule with minimum learning rate 10^{-6}, with 370 warmup steps on WikiText-103 and 390 on SlimPajama-100M. Selection follows the same MVC rule with five-epoch patience and a maximum of 300 epochs. The selected weights are then frozen for sKD and mixKD.

Teacher pretraining uses 512 sequences per rank at $H = 32$ and 256 at $H = 64$, with four ranks and four accumulation steps in both cases. Validation uses 1024 or 512 sequences per rank, respectively. This gives the same 262,144 tokens per full optimizer-step attempt as student training, but different final batches. Each teacher epoch processes 94,765,056 training tokens on WikiText-103 or 99,942,400 on SlimPajama-100M, giving 362 or 382 update attempts. The partial final update covers 131,072 or 65,536 tokens, respectively. These counts use the teacher's own block length and sequence batching, not the student loader or the frozen-teacher window batching used during KD.

Table 7: Short-teacher pretraining exposure. Each pair reports the selected checkpoint / complete stopped run. Tokens are billions of teacher training inputs, rounded to three decimals. Each teacher is shared by the corresponding sKD and mixKD conditions and is listed once.

Seed	Epochs	Update attempts	Tokens (10^9)
WikiText-103, $H = 32$			
42	8 / 13	2,896 / 4,706	0.758 / 1.232
99	8 / 13	2,896 / 4,706	0.758 / 1.232
123	8 / 13	2,896 / 4,706	0.758 / 1.232
WikiText-103, $H = 64$			
42	8 / 13	2,896 / 4,706	0.758 / 1.232
99	8 / 13	2,896 / 4,706	0.758 / 1.232
123	8 / 13	2,896 / 4,706	0.758 / 1.232
SlimPajama-100M, $H = 64$			
42	7 / 12	2,674 / 4,584	0.700 / 1.199
99	7 / 12	2,674 / 4,584	0.700 / 1.199
123	7 / 12	2,674 / 4,584	0.700 / 1.199

All six WikiText-103 short teachers were selected at epoch 8 and stopped at epoch 13; all three SlimPajama-100M short teachers were selected at epoch 7 and stopped at epoch 12. Their logs and resolved results agree on these epochs. Checkpoint-provenance records confirm that sKD and mixKD use the same short teacher for each corpus, horizon, and seed. The equal-context teacher is the already-counted same-seed baseline checkpoint, so its exposure is given in Tables 5 and 6, not added again as a separate pretraining run. These tables inventory the training exposure of the checkpoints used in the main comparisons, not all development runs. Exact counts and source hashes are retained in the training-exposure provenance record. Recorded runtime and memory, including online teacher computation within student timings, are reported in Appendix B.4.

B.4 Hardware and recorded training time

All WikiText-103 and SlimPajama-100M training runs used four RTX PRO 6000 GPUs with 96 GB VRAM per GPU. Teacher predictions were generated online during KD, so their forward-pass time is included in the student timings; no offline teacher-prediction cache was constructed. The preserved YAML configurations confirm five-epoch patience for all 30 baseline/student and nine short-teacher training phases. Tables 8–10 report their recorded runtimes and memory separately from the token-exposure counts above.

Timing and counting. Elapsed hours sum the rank-zero epoch timers through the selected checkpoint or through stopping, including the five patience epochs in the latter. Mean epoch time averages the complete stopped run. The archived trainer starts each timer before training and stops it after validation and metric synchronization, confirming that validation is included. Initialization, checkpoint handling after the timer, final model export, and subsequent test/diagnostic evaluation are outside this accounting. GPU-hours multiply stopped elapsed hours by four, without implying continuous device utilization. These are implementation-specific timings under validation-based stopping, not a matched-efficiency benchmark.

Memory measurement. The saved YAML field reports the local process's peak CUDA memory *reserved*, divided by 1024^3 and rounded to two decimals, so the table unit is GiB despite the field's GB name. Peak statistics are reset at phase start and read after its training, evaluation, and FLOP profiling. This is neither a training-only peak nor a maximum across all four GPUs, and it is not total device memory usage. The reporting rank is not identified in the saved field. The WikiText-103 eqKD seed-42 entry is estimated rather than measured, as specified in Table 8.

Table 8: WikiText-103 runtime and peak reserved memory. Elapsed hours are selected / stopped; GPU-hours cover the stopped run. Baseline and eqKD checkpoints shared across $H \in \{32, 64\}$ comparisons are counted once. †Estimated from the other two WikiText-103 eqKD runs, both of which recorded 29.53 GiB. The seed-42 run completed training, but a post-run reporting failure prevented recording its peak memory.

Condition	Seed	Mean epoch (min)	Elapsed hours selected / stopped	GPU-hours stopped	Peak VRAM (GiB)
Baseline	42	3.42	0.51 / 0.80	3.19	19.14
Baseline	99	3.43	0.51 / 0.80	3.20	19.14
Baseline	123	3.43	0.51 / 0.80	3.20	19.14
sKD $H = 32$	42	22.41	6.35 / 8.22	32.87	35.00
sKD $H = 32$	99	22.38	7.83 / 9.70	38.79	34.99
sKD $H = 32$	123	22.44	6.74 / 8.60	34.41	34.99
sKD $H = 64$	42	38.13	10.80 / 13.98	55.93	38.44
sKD $H = 64$	99	38.09	10.79 / 13.97	55.86	42.97
sKD $H = 64$	123	38.07	10.15 / 13.33	53.30	42.97
eqKD	42	5.48	1.74 / 2.19	8.77	29.53†
eqKD	99	5.63	1.69 / 2.16	8.63	29.53
eqKD	123	5.62	1.69 / 2.15	8.62	29.53
mixKD $H = 32$	42	18.99	5.38 / 6.96	27.85	32.33
mixKD $H = 32$	99	19.00	5.70 / 7.28	29.14	32.33
mixKD $H = 32$	123	18.99	5.38 / 6.96	27.86	32.33
mixKD $H = 64$	42	33.16	9.40 / 12.16	48.64	32.48
mixKD $H = 64$	99	33.31	10.54 / 13.32	53.29	32.48
mixKD $H = 64$	123	33.36	9.46 / 12.23	48.93	32.48

Table 9: SlimPajama-100M runtime and peak reserved memory, with the same timing and memory conventions as Table 8. Student durations include online teacher computation.

Condition	Seed	Mean epoch (min)	Elapsed hours selected / stopped	GPU-hours stopped	Peak VRAM (GiB)
Baseline	42	3.53	0.47 / 0.77	3.06	19.15
Baseline	99	3.53	0.47 / 0.76	3.06	19.15
Baseline	123	3.54	0.47 / 0.77	3.07	19.15
sKD $H = 64$	42	40.09	12.03 / 15.37	61.46	38.99
sKD $H = 64$	99	40.06	10.01 / 13.35	53.41	39.02
sKD $H = 64$	123	40.07	10.02 / 13.36	53.42	42.75
eqKD	42	5.86	1.47 / 1.95	7.82	29.55
eqKD	99	6.13	1.53 / 2.04	8.17	29.55
eqKD	123	5.94	1.49 / 1.98	7.92	29.55
mixKD $H = 64$	42	35.10	8.77 / 11.70	46.80	32.49
mixKD $H = 64$	99	35.07	9.94 / 12.86	51.44	32.49
mixKD $H = 64$	123	34.88	8.72 / 11.63	46.50	32.49

Table 10: Short-teacher pretraining runtime and peak reserved memory. Each teacher is counted once, although it is reused by both sKD and mixKD. Equal-context teachers reuse the baseline checkpoints and incur no additional pretraining entry.

Condition	Seed	Mean epoch (min)	Elapsed hours selected / stopped	GPU-hours stopped	Peak VRAM (GiB)
WikiText-103, $H = 32$					
Short teacher	42	3.37	0.45 / 0.73	2.92	36.70
Short teacher	99	3.36	0.45 / 0.73	2.91	33.61
Short teacher	123	3.37	0.45 / 0.73	2.92	33.61
WikiText-103, $H = 64$					
Short teacher	42	3.31	0.44 / 0.72	2.87	34.88
Short teacher	99	3.31	0.44 / 0.72	2.87	34.90
Short teacher	123	3.32	0.44 / 0.72	2.87	34.90
SlimPajama-100M, $H = 64$					
Short teacher	42	3.42	0.40 / 0.68	2.73	36.72
Short teacher	99	3.41	0.40 / 0.68	2.73	33.69
Short teacher	123	3.42	0.40 / 0.68	2.74	36.72

Across these 39 distinct training phases, recorded epoch time sums to 228.54 elapsed hours, or 914.16 GPU-hours. WikiText-103 accounts for 559.83 GPU-hours and SlimPajama-100M for 354.33; these totals include 25.56 GPU-hours of short-teacher pretraining. They are cumulative run times, not calendar duration or an end-to-end research-project budget. Development runs, supplementary WikiText-2 experiments, and later analyses are excluded. Deploying one student instead of two does not establish a measured efficiency advantage.

C Primary Prediction and Contextual-Effect Results

These results use the fixed checkpoints and common-tail populations specified in Appendix B. We report each seed separately and distinguish selection-validation results from source-provided test results. Additional source-provided WikiText-103 validation is in Appendix D. Positive prediction improvement means comparator CE minus method CE; contextual-gain G_M instead compares inputs within the same model. All losses and differences are in nats per target. Unless stated otherwise, difference intervals are the original pointwise paired-block intervals, not intervals inferred from separate model means.

C.1 Overall prediction

On test, mixKD has lower overall CE than eqKD in seven of nine configurations, with positive intervals in six. WikiText-103 seed 42 at $H = h = 32$ has a negative point difference, and both its interval and seed 123's positive-point interval include zero. SlimPajama-100M seed 42 favors eqKD with an interval excluding zero. These exceptions remain part of the comparison; zero-crossing intervals do not establish equivalence.

Table 11: Full-input source-provided test CE and mixKD improvement over eqKD, with paired 95% intervals. These are the unabridged intervals for Table 1. WikiText-103 baseline and eqKD full-input results are shared across horizons.

Seed	Baseline	sKD	eqKD	mixKD	mixKD improvement over eqKD [95% CI]
WikiText-103, $H = h = 32$					
42	3.2299	3.1846	3.0904	3.0909	-0.000432 [-0.004834, +0.003872]
99	3.2315	3.1748	3.0912	3.0869	+0.004232 [+0.000449, +0.007840]
123	3.2338	3.1853	3.0979	3.0948	+0.003117 [-0.000667, +0.006652]
WikiText-103, $H = h = 64$					
42	3.2299	3.1545	3.0904	3.0815	+0.008894 [+0.004894, +0.013062]
99	3.2315	3.1560	3.0912	3.0813	+0.009892 [+0.005916, +0.013533]
123	3.2338	3.1580	3.0979	3.0881	+0.009834 [+0.006021, +0.013696]
SlimPajama-100M, $H = h = 64$					
42	3.8087	3.7685	3.6859	3.6880	-0.002071 [-0.002811, -0.001327]
99	3.8022	3.7685	3.6845	3.6763	+0.008203 [+0.007364, +0.009016]
123	3.8164	3.7834	3.6918	3.6815	+0.010258 [+0.009511, +0.011086]

Table 12: Full-input CE on the recorded selection-validation populations and paired mixKD improvement over eqKD. WikiText-103 rows here are the original internal-validation mixture comparison at $H = h = 32$; source-provided validation at $H = h = 64$ is reported separately in Appendix D. Probe losses use 897 targets per block, unlike all-position checkpoint-selection CE.

Seed	Baseline	sKD	eqKD	mixKD	mixKD improvement over eqKD [95% CI]
WikiText-103 internal validation, $H = h = 32$					
42	3.2928	3.2480	3.1545	3.1562	-0.001713 [-0.002286, -0.001075]
99	3.2953	3.2364	3.1579	3.1514	+0.006417 [+0.005838, +0.006953]
123	3.2981	3.2474	3.1634	3.1567	+0.006630 [+0.005971, +0.007254]
SlimPajama-100M source-provided validation, $H = h = 64$					
42	3.7744	3.7388	3.6536	3.6569	-0.003270 [-0.004076, -0.002525]
99	3.7688	3.7399	3.6520	3.6459	+0.006114 [+0.005261, +0.007030]
123	3.7841	3.7556	3.6596	3.6505	+0.009145 [+0.008349, +0.010044]

Table 13: WikiText-103 internal-validation pure-teacher comparisons. Positive improvement favors eqKD over sKD. Helpful/harmful entries are conditional improvements with paired 95% intervals, not population-weighted contributions.

Seed	sKD CE	Overall improvement [95% CI]	Helpful improvement [95% CI]	Harmful improvement [95% CI]
WikiText-103 internal validation, $H = h = 32$				
42	3.2480	+0.0935 [+0.0925, +0.0945]	+0.2215 [+0.2198, +0.2231]	-0.0962 [-0.0978, -0.0947]
99	3.2364	+0.0785 [+0.0776, +0.0795]	+0.2033 [+0.2015, +0.2049]	-0.1081 [-0.1096, -0.1066]
123	3.2474	+0.0840 [+0.0830, +0.0850]	+0.2116 [+0.2099, +0.2132]	-0.1040 [-0.1055, -0.1025]
WikiText-103 internal validation, $H = h = 64$				
42	3.2164	+0.0619 [+0.0610, +0.0629]	+0.1679 [+0.1663, +0.1694]	-0.0654 [-0.0668, -0.0640]
99	3.2203	+0.0625 [+0.0616, +0.0634]	+0.1627 [+0.1611, +0.1643]	-0.0571 [-0.0585, -0.0558]
123	3.2205	+0.0571 [+0.0562, +0.0581]	+0.1593 [+0.1577, +0.1609]	-0.0647 [-0.0662, -0.0632]

C.2 Conditional prediction and population contributions

The tables retain all three populations and their counts. Within a row, the two contrasts share the same target mask. Both focal advantages recur, with mixKD improving helpful-position prediction over sKD and harmful-position prediction over eqKD. mixKD nevertheless remains worse than eqKD on helpful positions. Conditional differences are not added directly; their population-weighted contributions follow in Appendix C.2.1.

Table 14 retains the baseline-relative prediction improvements underlying the seed means in Figure 2. Each plotted mean gives equal weight to the three seed-specific estimates, whose donor-defined populations can differ across seeds. The table retains the original within-seed paired intervals; these are not intervals for the displayed seed means.

Table 14: Seed-level full-input CE improvement over the same-seed baseline at $H = h = 64$ on source-provided test splits. Positive values favor the distilled student. Each cell gives the original estimate and paired-block 95% interval, in nats per target. All three methods use the same population within each seed. Counts are retained in Tables 16 and 17.

Seed	Population	sKD	mixKD	eqKD
WikiText-103, $H = h = 64$				
42	Helpful	-0.0145 [-0.0246, -0.0050]	0.1198 [0.1122, 0.1277]	0.1504 [0.1438, 0.1569]
42	Harmful	0.1880 [0.1777, 0.1990]	0.1845 [0.1756, 0.1935]	0.1253 [0.1175, 0.1332]
42	Discordant/zero	0.1107 [0.1039, 0.1179]	0.1593 [0.1524, 0.1657]	0.1355 [0.1299, 0.1410]
99	Helpful	-0.0103 [-0.0211, -0.0006]	0.1220 [0.1146, 0.1292]	0.1476 [0.1405, 0.1544]
99	Harmful	0.1736 [0.1637, 0.1848]	0.1878 [0.1788, 0.1973]	0.1275 [0.1193, 0.1355]
99	Discordant/zero	0.1139 [0.1069, 0.1213]	0.1601 [0.1541, 0.1667]	0.1395 [0.1340, 0.1450]
123	Helpful	-0.0201 [-0.0300, -0.0108]	0.1120 [0.1055, 0.1194]	0.1440 [0.1375, 0.1506]
123	Harmful	0.1936 [0.1819, 0.2044]	0.1937 [0.1841, 0.2033]	0.1274 [0.1193, 0.1344]
123	Discordant/zero	0.1148 [0.1075, 0.1226]	0.1560 [0.1493, 0.1630]	0.1317 [0.1254, 0.1376]
SlimPajama-100M, $H = h = 64$				
42	Helpful	-0.0821 [-0.0850, -0.0792]	0.0628 [0.0611, 0.0645]	0.1205 [0.1193, 0.1218]
42	Harmful	0.1787 [0.1769, 0.1805]	0.1857 [0.1842, 0.1874]	0.1224 [0.1210, 0.1237]
42	Discordant/zero	0.0949 [0.0934, 0.0963]	0.1468 [0.1455, 0.1480]	0.1255 [0.1244, 0.1266]
99	Helpful	-0.0992 [-0.1022, -0.0963]	0.0728 [0.0711, 0.0745]	0.1136 [0.1126, 0.1147]
99	Harmful	0.1815 [0.1797, 0.1833]	0.1828 [0.1812, 0.1843]	0.1190 [0.1177, 0.1203]
99	Discordant/zero	0.0953 [0.0939, 0.0968]	0.1517 [0.1505, 0.1530]	0.1215 [0.1204, 0.1225]
123	Helpful	-0.1104 [-0.1138, -0.1071]	0.0980 [0.0965, 0.0997]	0.1422 [0.1411, 0.1435]
123	Harmful	0.1884 [0.1864, 0.1902]	0.1728 [0.1712, 0.1745]	0.0988 [0.0975, 0.1001]
123	Discordant/zero	0.1015 [0.1001, 0.1029]	0.1536 [0.1524, 0.1548]	0.1203 [0.1193, 0.1213]

Table 15: WikiText-103 source-provided test conditional full-input prediction improvements at $H = h = 32$. Both columns favor mixKD when positive and retain paired 95% intervals. Counts include the entire discordant/zero population.

Seed	Population	n	Over sKD [95% CI]	Over eqKD [95% CI]
42	Helpful	115,534	0.1646 [0.1565, 0.1730]	-0.0531 [-0.0585, -0.0471]
42	Harmful	44,212	-0.0091 [-0.0164, -0.0021]	0.0835 [0.0752, 0.0916]
42	Discordant/zero	87,826	0.0524 [0.0471, 0.0573]	0.0265 [0.0204, 0.0325]
99	Helpful	115,735	0.1565 [0.1468, 0.1663]	-0.0459 [-0.0511, -0.0405]
99	Harmful	44,619	-0.0098 [-0.0188, -0.0010]	0.0844 [0.0763, 0.0920]
99	Discordant/zero	87,218	0.0467 [0.0401, 0.0535]	0.0297 [0.0247, 0.0350]
123	Helpful	115,670	0.1659 [0.1581, 0.1740]	-0.0492 [-0.0545, -0.0436]
123	Harmful	44,278	-0.0146 [-0.0224, -0.0067]	0.0848 [0.0773, 0.0919]
123	Discordant/zero	87,624	0.0440 [0.0383, 0.0503]	0.0310 [0.0252, 0.0365]

Table 16: WikiText-103 source-provided test conditional full-input prediction improvements at $H = h = 64$. Both columns favor mixKD when positive and retain paired 95% intervals. Counts include the entire discordant/zero population.

Seed	Population	n	Over sKD [95% CI]	Over eqKD [95% CI]
42	Helpful	100,973	0.1344 [0.1261, 0.1419]	-0.0305 [-0.0363, -0.0244]
42	Harmful	50,662	-0.0035 [-0.0101, 0.0032]	0.0592 [0.0514, 0.0670]
42	Discordant/zero	95,937	0.0487 [0.0435, 0.0536]	0.0238 [0.0179, 0.0291]
99	Helpful	100,827	0.1323 [0.1245, 0.1412]	-0.0256 [-0.0317, -0.0197]
99	Harmful	50,602	0.0142 [0.0066, 0.0211]	0.0603 [0.0529, 0.0680]
99	Discordant/zero	96,143	0.0462 [0.0409, 0.0514]	0.0206 [0.0153, 0.0258]
123	Helpful	101,144	0.1320 [0.1240, 0.1411]	-0.0320 [-0.0378, -0.0261]
123	Harmful	50,417	0.0000 [-0.0079, 0.0077]	0.0663 [0.0589, 0.0737]
123	Discordant/zero	96,011	0.0412 [0.0361, 0.0463]	0.0243 [0.0191, 0.0296]

Table 17: SlimPajama-100M source-provided test conditional full-input prediction improvements at $H = h = 64$. Both columns favor mixKD when positive and retain paired 95% intervals. Counts include the entire discordant/zero population.

Seed	Population	n	Over sKD [95% CI]	Over eqKD [95% CI]
42	Helpful	3,501,520	0.1449 [0.1429, 0.1468]	-0.0577 [-0.0590, -0.0563]
42	Harmful	1,882,132	0.0071 [0.0058, 0.0084]	0.0634 [0.0622, 0.0645]
42	Discordant/zero	3,059,809	0.0519 [0.0509, 0.0529]	0.0213 [0.0205, 0.0221]
99	Helpful	3,509,661	0.1720 [0.1702, 0.1740]	-0.0408 [-0.0424, -0.0393]
99	Harmful	1,878,321	0.0014 [0.0001, 0.0028]	0.0638 [0.0624, 0.0651]
99	Discordant/zero	3,055,479	0.0564 [0.0554, 0.0575]	0.0303 [0.0292, 0.0313]
123	Helpful	3,505,271	0.2084 [0.2060, 0.2109]	-0.0442 [-0.0457, -0.0429]
123	Harmful	1,894,178	-0.0155 [-0.0168, -0.0143]	0.0740 [0.0728, 0.0752]
123	Discordant/zero	3,044,012	0.0521 [0.0511, 0.0531]	0.0333 [0.0324, 0.0342]

Table 18: WikiText-103 internal validation conditional full-input prediction improvements at $H = h = 32$, with the same sign and interval conventions as the test tables.

Seed	Population	n	Over sKD [95% CI]	Over eqKD [95% CI]
42	Helpful	4,835,634	0.1644 [0.1631, 0.1657]	-0.0571 [-0.0580, -0.0562]
42	Harmful	1,878,408	-0.0120 [-0.0131, -0.0108]	0.0843 [0.0831, 0.0855]
42	Discordant/zero	3,666,939	0.0493 [0.0485, 0.0501]	0.0273 [0.0265, 0.0281]
99	Helpful	4,841,601	0.1565 [0.1551, 0.1579]	-0.0468 [-0.0476, -0.0460]
99	Harmful	1,869,065	-0.0150 [-0.0162, -0.0138]	0.0931 [0.0919, 0.0943]
99	Discordant/zero	3,670,315	0.0414 [0.0406, 0.0423]	0.0325 [0.0317, 0.0332]
123	Helpful	4,828,370	0.1677 [0.1663, 0.1690]	-0.0439 [-0.0448, -0.0429]
123	Harmful	1,877,499	-0.0151 [-0.0164, -0.0139]	0.0889 [0.0876, 0.0901]
123	Discordant/zero	3,675,112	0.0434 [0.0425, 0.0443]	0.0310 [0.0302, 0.0318]

Table 19: SlimPajama-100M source-provided validation conditional full-input prediction improvements at $H = h = 64$, with the same sign and interval conventions as the test tables.

Seed	Population	n	Over sKD [95% CI]	Over eqKD [95% CI]
42	Helpful	3,480,362	0.1489 [0.1470, 0.1509]	-0.0579 [-0.0592, -0.0566]
42	Harmful	1,838,975	0.0066 [0.0054, 0.0080]	0.0615 [0.0603, 0.0627]
42	Discordant/zero	2,984,192	0.0503 [0.0492, 0.0514]	0.0205 [0.0197, 0.0214]
99	Helpful	3,489,693	0.1739 [0.1718, 0.1758]	-0.0435 [-0.0450, -0.0420]
99	Harmful	1,836,173	0.0016 [0.0001, 0.0029]	0.0617 [0.0604, 0.0630]
99	Discordant/zero	2,977,663	0.0573 [0.0562, 0.0584]	0.0300 [0.0290, 0.0310]
123	Helpful	3,486,028	0.2123 [0.2098, 0.2147]	-0.0468 [-0.0482, -0.0454]
123	Harmful	1,843,797	-0.0136 [-0.0149, -0.0124]	0.0755 [0.0743, 0.0767]
123	Discordant/zero	2,973,704	0.0530 [0.0520, 0.0539]	0.0336 [0.0327, 0.0345]

C.2.1 POPULATION-WEIGHTED PREDICTION CONTRIBUTIONS

Table 20 weights the conditional mixKD–eqKD prediction differences by population size, as in Eq. 6. On SlimPajama-100M test, the helpful-position cost exceeds the harmful-position improvement in all three seeds. Discordant/zero improvements turn the overall result positive in seeds 99 and 123. These positions record donor disagreement or zero effects, not semantic neutrality.

Table 20: Population contributions to mixKD improvement over eqKD, in nats per all evaluated targets. Each row's three contributions sum to its overall improvement before rounding. Splits and horizons are identified explicitly; seeds are not pooled.

Seed	Helpful	Harmful	Discordant/zero	Overall
WikiText-103, $H = h = 32$*, source-provided test*				
42	-0.024760	+0.014915	+0.009413	-0.000432
99	-0.021457	+0.015209	+0.010480	+0.004232
123	-0.023008	+0.015158	+0.010967	+0.003117
WikiText-103, $H = h = 64$*, source-provided test*				
42	-0.012453	+0.012119	+0.009229	+0.008894
99	-0.010435	+0.012317	+0.008010	+0.009892
123	-0.013085	+0.013502	+0.009417	+0.009834
SlimPajama-100M, $H = h = 64$*, source-provided test*				
42	-0.023912	+0.014126	+0.007715	-0.002071
99	-0.016951	+0.014202	+0.010952	+0.008203
123	-0.018348	+0.016597	+0.012010	+0.010258
WikiText-103 internal validation, $H = h = 32$				
42	-0.026611	+0.015249	+0.009650	-0.001713
99	-0.021827	+0.016763	+0.011480	+0.006417
123	-0.020416	+0.016077	+0.010969	+0.006630
SlimPajama-100M source-provided validation, $H = h = 64$				
42	-0.024277	+0.013625	+0.007383	-0.003270
99	-0.018293	+0.013646	+0.010760	+0.006114
123	-0.019645	+0.016766	+0.012023	+0.009145

C.3 CONTEXTUAL EFFECTS AND RESTRICTED-INPUT ACCOUNTING

The test tables report mean G_M within each fixed population for every method, with the original intervals on those individual means. They measure the benefit or penalty of adding distant input within a model, not prediction improvement over another model. The validation table reports paired changes in contextual-gain between mixKD and each pure-teacher student. Positive and negative changes have different implications on populations with positive and negative contextual effects. No between-method interval is inferred from marginal bounds.

Table 21: WikiText-103 source-provided test mean contextual-gain G_M at $H = h = 32$. Positive means contextual benefit and negative means a penalty. Intervals concern individual means. Population counts are in Table 15.

Seed	Method	Helpful [95% CI]	Harmful [95% CI]	Discordant/zero [95% CI]
42	Baseline	0.8587 [0.8345, 0.8884]	-0.1769 [-0.1866, -0.1683]	0.1295 [0.1230, 0.1360]
42	sKD	0.6189 [0.6009, 0.6403]	0.0326 [0.0246, 0.0410]	0.1892 [0.1827, 0.1955]
42	eqKD	0.8691 [0.8443, 0.8984]	-0.1192 [-0.1288, -0.1106]	0.1769 [0.1701, 0.1832]
42	mixKD	0.7964 [0.7723, 0.8238]	-0.0098 [-0.0191, -0.0006]	0.2130 [0.2058, 0.2200]
99	Baseline	0.8452 [0.8215, 0.8732]	-0.1613 [-0.1705, -0.1520]	0.1308 [0.1244, 0.1373]
99	sKD	0.6442 [0.6248, 0.6665]	0.0686 [0.0598, 0.0772]	0.2155 [0.2085, 0.2226]
99	eqKD	0.8914 [0.8664, 0.9203]	-0.0855 [-0.0953, -0.0767]	0.2030 [0.1965, 0.2095]
99	mixKD	0.7438 [0.7226, 0.7695]	-0.0454 [-0.0539, -0.0366]	0.1724 [0.1667, 0.1783]
123	Baseline	0.8438 [0.8208, 0.8703]	-0.1704 [-0.1796, -0.1620]	0.1288 [0.1231, 0.1349]
123	sKD	0.6511 [0.6325, 0.6742]	0.0745 [0.0648, 0.0834]	0.2189 [0.2126, 0.2256]
123	eqKD	0.8990 [0.8747, 0.9271]	-0.0953 [-0.1048, -0.0863]	0.1958 [0.1895, 0.2022]
123	mixKD	0.7901 [0.7679, 0.8153]	-0.0164 [-0.0260, -0.0070]	0.2036 [0.1969, 0.2102]

Table 22: WikiText-103 source-provided test mean contextual-gain G_M at $H = h = 64$. Positive means contextual benefit and negative means a penalty. Intervals concern individual means. Population counts are in Table 16.

Seed	Method	Helpful [95% CI]	Harmful [95% CI]	Discordant/zero [95% CI]
42	Baseline	0.5372 [0.5177, 0.5608]	-0.1369 [-0.1439, -0.1304]	0.0743 [0.0699, 0.0790]
42	sKD	0.3628 [0.3486, 0.3788]	-0.0051 [-0.0104, 0.0005]	0.0937 [0.0893, 0.0984]
42	eqKD	0.5386 [0.5183, 0.5613]	-0.1060 [-0.1123, -0.1000]	0.0930 [0.0885, 0.0980]
42	mixKD	0.4811 [0.4623, 0.5025]	-0.0399 [-0.0461, -0.0337]	0.1129 [0.1075, 0.1180]
99	Baseline	0.5356 [0.5178, 0.5570]	-0.1273 [-0.1342, -0.1210]	0.0733 [0.0688, 0.0775]
99	sKD	0.3622 [0.3478, 0.3781]	-0.0098 [-0.0153, -0.0042]	0.0913 [0.0869, 0.0960]
99	eqKD	0.5506 [0.5313, 0.5737]	-0.0841 [-0.0912, -0.0779]	0.1095 [0.1047, 0.1142]
99	mixKD	0.4648 [0.4468, 0.4849]	-0.0419 [-0.0482, -0.0361]	0.1072 [0.1021, 0.1121]
123	Baseline	0.5315 [0.5119, 0.5544]	-0.1369 [-0.1444, -0.1298]	0.0697 [0.0652, 0.0744]
123	sKD	0.3524 [0.3385, 0.3684]	-0.0157 [-0.0208, -0.0108]	0.0836 [0.0796, 0.0876]
123	eqKD	0.5565 [0.5358, 0.5797]	-0.0918 [-0.0992, -0.0848]	0.1019 [0.0970, 0.1071]
123	mixKD	0.4806 [0.4623, 0.5021]	-0.0326 [-0.0390, -0.0256]	0.1125 [0.1078, 0.1174]

Table 23: SlimPajama-100M source-provided test mean contextual-gain G_M at $H = h = 64$. Positive means contextual benefit and negative means a penalty. Intervals concern individual means. Population counts are in Table 17.

Seed	Method	Helpful [95% CI]	Harmful [95% CI]	Discordant/zero [95% CI]
42	Baseline	0.7328 [0.7238, 0.7421]	-0.1941 [-0.1955, -0.1927]	0.0435 [0.0428, 0.0443]
42	sKD	0.5953 [0.5883, 0.6029]	-0.0322 [-0.0334, -0.0311]	0.1048 [0.1039, 0.1057]
42	eqKD	0.7359 [0.7272, 0.7452]	-0.1731 [-0.1745, -0.1718]	0.0616 [0.0609, 0.0624]
42	mixKD	0.6625 [0.6542, 0.6711]	-0.1102 [-0.1114, -0.1091]	0.0756 [0.0748, 0.0764]
99	Baseline	0.7478 [0.7390, 0.7572]	-0.2043 [-0.2056, -0.2029]	0.0408 [0.0401, 0.0416]
99	sKD	0.5455 [0.5386, 0.5526]	-0.0826 [-0.0837, -0.0816]	0.0562 [0.0556, 0.0569]
99	eqKD	0.7440 [0.7348, 0.7532]	-0.1805 [-0.1820, -0.1792]	0.0572 [0.0563, 0.0579]
99	mixKD	0.7066 [0.6984, 0.7153]	-0.0979 [-0.0991, -0.0967]	0.0991 [0.0981, 0.1000]
123	Baseline	0.7311 [0.7220, 0.7402]	-0.1940 [-0.1955, -0.1926]	0.0419 [0.0411, 0.0426]
123	sKD	0.5143 [0.5080, 0.5210]	-0.0745 [-0.0755, -0.0735]	0.0559 [0.0553, 0.0566]
123	eqKD	0.7527 [0.7439, 0.7622]	-0.1951 [-0.1966, -0.1936]	0.0529 [0.0522, 0.0537]
123	mixKD	0.6801 [0.6720, 0.6884]	-0.1358 [-0.1371, -0.1346]	0.0654 [0.0647, 0.0662]

Table 24: Selection-validation changes in contextual-gain, $G_{\mathrm{mixKD}} - G_R$, with original paired 95% intervals. These are contextual changes, not CE improvements; counts are in the corresponding conditional-prediction tables.

Seed	Population	$G_{\mathrm{mixKD}} - G_{\mathrm{sKD}}$ [95% CI]	$G_{\mathrm{mixKD}} - G_{\mathrm{eqKD}}$ [95% CI]
WikiText-103 internal validation, $H = h = 32$			
42	Helpful	0.1722 [0.1709, 0.1735]	-0.0747 [-0.0759, -0.0737]
42	Harmful	-0.0460 [-0.0475, -0.0447]	0.1113 [0.1099, 0.1128]
42	Discordant/zero	0.0283 [0.0273, 0.0293]	0.0417 [0.0408, 0.0427]
99	Helpful	0.0966 [0.0955, 0.0979]	-0.1488 [-0.1499, -0.1476]
99	Harmful	-0.1194 [-0.1209, -0.1180]	0.0486 [0.0473, 0.0500]
99	Discordant/zero	-0.0453 [-0.0463, -0.0443]	-0.0275 [-0.0284, -0.0264]
123	Helpful	0.1350 [0.1338, 0.1364]	-0.1117 [-0.1129, -0.1105]
123	Harmful	-0.0892 [-0.0907, -0.0878]	0.0786 [0.0772, 0.0799]
123	Discordant/zero	-0.0144 [-0.0154, -0.0133]	0.0066 [0.0056, 0.0076]
SlimPajama-100M source-provided validation, $H = h = 64$			
42	Helpful	0.0701 [0.0687, 0.0715]	-0.0729 [-0.0742, -0.0718]
42	Harmful	-0.0777 [-0.0788, -0.0766]	0.0614 [0.0605, 0.0624]
42	Discordant/zero	-0.0287 [-0.0294, -0.0279]	0.0138 [0.0132, 0.0145]
99	Helpful	0.1608 [0.1589, 0.1627]	-0.0390 [-0.0405, -0.0376]
99	Harmful	-0.0161 [-0.0171, -0.0151]	0.0806 [0.0796, 0.0817]
99	Discordant/zero	0.0419 [0.0411, 0.0426]	0.0408 [0.0400, 0.0415]
123	Helpful	0.1667 [0.1645, 0.1689]	-0.0760 [-0.0774, -0.0746]
123	Harmful	-0.0611 [-0.0621, -0.0602]	0.0598 [0.0588, 0.0608]
123	Discordant/zero	0.0098 [0.0091, 0.0105]	0.0126 [0.0118, 0.0133]

C.3.1 RESTRICTED-INPUT AND CONTEXTUAL-CHANGE ACCOUNTING

Table 25 gives the separate identity in Eq. 5, comparing mixKD with eqKD on the same targets. Restricted-input improvement plus change in contextual-gain equals full-input improvement before rounding. This accounting does not identify a universal local-prediction mechanism or a causal mediation effect.

Table 25: Paired mixKD–eqKD accounting on all targets, with original 95% intervals. Restricted-input CE improvement plus contextual-gain change equals full-input CE improvement. This partitions the input comparison, not the target populations.

Seed	Restricted-input improvement [95% CI]	Contextual-gain change [95% CI]	Full-input improvement [95% CI]
WikiText-103, $H = h = 32$, source-provided test			
42	+0.001168 [-0.004221, +0.006442]	-0.001600 [-0.006362, +0.003077]	-0.000432 [-0.004834, +0.003872]
99	+0.076729 [+0.071762, +0.081681]	-0.072497 [-0.076904, -0.068255]	+0.004232 [+0.000449, +0.007840]
123	+0.037168 [+0.032035, +0.042314]	-0.034051 [-0.039332, -0.028925]	+0.003117 [-0.000667, +0.006652]
WikiText-103, $H = h = 64$, source-provided test			
42	+0.011111 [+0.006507, +0.015997]	-0.002216 [-0.005626, +0.001115]	+0.008894 [+0.004894, +0.013062]
99	+0.037098 [+0.032346, +0.041697]	-0.027206 [-0.031157, -0.023637]	+0.009892 [+0.005916, +0.013533]
123	+0.024652 [+0.020480, +0.029044]	-0.014818 [-0.018347, -0.011279]	+0.009834 [+0.006021, +0.013696]
SlimPajama-100M, $H = h = 64$, source-provided test			
42	+0.009297 [+0.008632, +0.009959]	-0.011368 [-0.012033, -0.010704]	-0.002071 [-0.002811, -0.001327]
99	-0.009821 [-0.010645, -0.008883]	+0.018024 [+0.017335, +0.018712]	+0.008203 [+0.007364, +0.009016]
123	+0.022616 [+0.021958, +0.023346]	-0.012358 [-0.013087, -0.011669]	+0.010258 [+0.009511, +0.011086]
WikiText-103 internal validation, $H = h = 32$			
42	-0.001791 [-0.002637, -0.000955]	+0.000079 [-0.000635, +0.000825]	-0.001713 [-0.002286, -0.001075]
99	+0.076760 [+0.076048, +0.077461]	-0.070343 [-0.071049, -0.069643]	+0.006417 [+0.005838, +0.006953]
123	+0.042043 [+0.041164, +0.042863]	-0.035413 [-0.036182, -0.034658]	+0.006630 [+0.005971, +0.007254]
SlimPajama-100M source-provided validation, $H = h = 64$			
42	+0.008721 [+0.008036, +0.009416]	-0.011990 [-0.012676, -0.011345]	-0.003270 [-0.004076, -0.002525]
99	-0.009929 [-0.010827, -0.009115]	+0.016043 [+0.015292, +0.016816]	+0.006114 [+0.005261, +0.007030]
123	+0.023271 [+0.022510, +0.024037]	-0.014126 [-0.014876, -0.013378]	+0.009145 [+0.008349, +0.010044]

C.4 PREDICTION-ENSEMBLE COMPARISON

The equal-probability ensemble combines the same-seed sKD and eqKD students, not their teachers. It averages their unsoftened probabilities separately under full and restricted inputs before computing

losses and contextual-gain (Appendix B.2). The comparison retains the main analysis's fixed donor-defined populations and test targets, without an additional effect-magnitude threshold.

The ensemble has lower overall CE in all nine corpus–horizon–seed comparisons (Table 26) and lower conditional CE in all 27 population comparisons (Table 27). Nevertheless, mixKD obtains greater average contextual benefit on helpful positions and incurs a smaller average contextual penalty on harmful positions in all nine settings (Table 28). The paired 95% intervals support each of these differences. The discordant/zero population is retained in both tables and is not interpreted as semantically neutral.

Intervals retain the original 1,000 paired-block resamples and quantify within-seed evaluation uncertainty conditional on the selected checkpoints, not training-seed uncertainty or simultaneous coverage. The two WikiText-103 horizons share test targets and baseline/eqKD checkpoints, so the nine comparisons are not nine independent datasets. The ensemble evaluates two students per prediction, whereas mixKD evaluates one; no measured latency or training-cost ratio is established.

Table 26: Source-provided test comparison with the equal probability prediction ensemble. Excess loss is mixKD CE minus ensemble CE, so positive values favor the ensemble. Intervals are the recorded paired direct loss-difference intervals.

Seed	Ensemble CE	mixKD CE	mixKD excess loss [95% CI]
WikiText-103, $H = h = 32$			
42	3.037490	3.090852	+0.053362 [+0.050471, +0.056255]
99	3.032024	3.086950	+0.054926 [+0.051945, +0.057976]
123	3.044601	3.094806	+0.050205 [+0.046977, +0.053266]
WikiText-103, $H = h = 64$			
42	3.028537	3.081525	+0.052988 [+0.050275, +0.056004]
99	3.029195	3.081290	+0.052095 [+0.048980, +0.055077]
123	3.032212	3.088089	+0.055877 [+0.052503, +0.059106]
SlimPajama-100M, $H = h = 64$			
42	3.624513	3.688012	+0.063499 [+0.062935, +0.064026]
99	3.631755	3.676291	+0.044536 [+0.043898, +0.045216]
123	3.637754	3.681532	+0.043778 [+0.043215, +0.044344]

Table 27: Conditional full-input excess loss on source-provided test data, $D_{\mathrm{CE}} = \mathrm{CE}_{\mathrm{mixKD}} - \mathrm{CE}_{\mathrm{ensemble}}$, in nats per target. Positive values favor the equal-probability sKD/eqKD prediction ensemble. Counts refer to fixed donor-defined populations. Intervals are the original pointwise paired-block 95% intervals.

Seed	Population	Count	D_{CE} [95% CI]
WikiText-103, $H = h = 32$			
42	Helpful	115,534	+0.0489 [+0.0445, +0.0531]
42	Harmful	44,212	+0.0643 [+0.0586, +0.0707]
42	Discordant/zero	87,826	+0.0538 [+0.0495, +0.0581]
99	Helpful	115,735	+0.0507 [+0.0463, +0.0554]
99	Harmful	44,619	+0.0651 [+0.0576, +0.0727]
99	Discordant/zero	87,218	+0.0553 [+0.0508, +0.0598]
123	Helpful	115,670	+0.0446 [+0.0399, +0.0491]
123	Harmful	44,278	+0.0614 [+0.0554, +0.0672]
123	Discordant/zero	87,624	+0.0520 [+0.0479, +0.0562]
WikiText-103, $H = h = 64$			
42	Helpful	100,973	+0.0459 [+0.0417, +0.0502]
42	Harmful	50,662	+0.0685 [+0.0625, +0.0740]
42	Discordant/zero	95,937	+0.0523 [+0.0483, +0.0569]
99	Helpful	100,827	+0.0448 [+0.0402, +0.0495]
99	Harmful	50,602	+0.0593 [+0.0532, +0.0652]
99	Discordant/zero	96,143	+0.0560 [+0.0519, +0.0602]
123	Helpful	101,144	+0.0497 [+0.0449, +0.0546]
123	Harmful	50,417	+0.0642 [+0.0582, +0.0706]
123	Discordant/zero	96,011	+0.0580 [+0.0538, +0.0622]
SlimPajama-100M, $H = h = 64$			
42	Helpful	3,501,520	+0.0656 [+0.0648, +0.0665]
42	Harmful	1,882,132	+0.0659 [+0.0650, +0.0669]
42	Discordant/zero	3,059,809	+0.0596 [+0.0588, +0.0603]
99	Helpful	3,509,661	+0.0367 [+0.0357, +0.0378]
99	Harmful	1,878,321	+0.0601 [+0.0590, +0.0613]
99	Discordant/zero	3,055,479	+0.0439 [+0.0430, +0.0449]
123	Helpful	3,505,271	+0.0299 [+0.0289, +0.0309]
123	Harmful	1,894,178	+0.0655 [+0.0646, +0.0665]
123	Discordant/zero	3,044,012	+0.0462 [+0.0455, +0.0469]

Table 28: Mean contextual-gains and their paired difference, $D_G = G_{\text{mixKD}} - G_{\text{ensemble}}$, in nats per target on the same test populations as Table 27. Positive D_G indicates greater helpful benefit or a smaller harmful penalty when both means have the corresponding sign. Intervals are the original pointwise paired-block 95% intervals for D_G, not intervals for the individual means.

Seed	Population	G_{ensemble}	G_{mixKD}	D_G [95% CI]
WikiText-103, $H = h = 32$				
42	Helpful	+0.6921	+0.7964	+0.1043 [+0.0983, +0.1109]
42	Harmful	-0.0882	-0.0098	+0.0784 [+0.0706, +0.0861]
42	Discordant/zero	+0.1280	+0.2130	+0.0851 [+0.0791, +0.0906]
99	Helpful	+0.7044	+0.7438	+0.0395 [+0.0340, +0.0446]
99	Harmful	-0.0692	-0.0454	+0.0238 [+0.0165, +0.0307]
99	Discordant/zero	+0.1440	+0.1724	+0.0284 [+0.0237, +0.0331]
123	Helpful	+0.7145	+0.7901	+0.0756 [+0.0696, +0.0816]
123	Harmful	-0.0686	-0.0164	+0.0522 [+0.0446, +0.0602]
123	Discordant/zero	+0.1413	+0.2036	+0.0623 [+0.0568, +0.0681]
WikiText-103, $H = h = 64$				
42	Helpful	+0.4210	+0.4811	+0.0601 [+0.0558, +0.0646]
42	Harmful	-0.0831	-0.0399	+0.0433 [+0.0385, +0.0485]
42	Discordant/zero	+0.0600	+0.1129	+0.0529 [+0.0487, +0.0571]
99	Helpful	+0.4259	+0.4648	+0.0389 [+0.0346, +0.0435]
99	Harmful	-0.0765	-0.0419	+0.0346 [+0.0291, +0.0401]
99	Discordant/zero	+0.0664	+0.1072	+0.0408 [+0.0370, +0.0442]
123	Helpful	+0.4258	+0.4806	+0.0548 [+0.0501, +0.0593]
123	Harmful	-0.0841	-0.0326	+0.0515 [+0.0462, +0.0572]
123	Discordant/zero	+0.0603	+0.1125	+0.0522 [+0.0480, +0.0564]
SlimPajama-100M, $H = h = 64$				
42	Helpful	+0.6438	+0.6625	+0.0187 [+0.0179, +0.0196]
42	Harmful	-0.1276	-0.1102	+0.0174 [+0.0167, +0.0182]
42	Discordant/zero	+0.0529	+0.0756	+0.0227 [+0.0221, +0.0232]
99	Helpful	+0.6393	+0.7066	+0.0673 [+0.0663, +0.0683]
99	Harmful	-0.1405	-0.0979	+0.0426 [+0.0418, +0.0435]
99	Discordant/zero	+0.0417	+0.0991	+0.0574 [+0.0567, +0.0580]
123	Helpful	+0.6367	+0.6801	+0.0434 [+0.0425, +0.0444]
123	Harmful	-0.1415	-0.1358	+0.0056 [+0.0048, +0.0065]
123	Discordant/zero	+0.0405	+0.0654	+0.0250 [+0.0244, +0.0256]

The joint helpful–harmful response. Helpful-context benefit and harmful-context penalty describe two aspects of a model's response to distant text. Let E denote the equal-weight ensemble of the sKD and eqKD students. In each of the nine corpus–horizon–seed test settings, the population means have the following orderings, from smaller to larger,

$$\begin{aligned} \text{Helpful benefit} \quad & \text{sKD} < E < \text{mixKD} < \text{eqKD}, \\ \text{Harmful penalty} \quad & \text{sKD} < \text{mixKD} < E < \text{eqKD}. \end{aligned}$$

Helpful benefit is the mean contextual-gain on helpful positions. Harmful penalty is the negative of mean contextual-gain on harmful positions, so a negative penalty means that the evaluated model benefits from distant text even within this donor-defined population. These are point-estimate orderings, not simultaneous confidence statements for the full rankings. The underlying student means appear in Appendix C.3, and the paired mixKD–ensemble comparisons appear in Table 28.

sKD has the smallest helpful benefit and harmful penalty, whereas eqKD has the largest of both. mixKD and the ensemble lie between these endpoints, but in different orders on the two dimensions. mixKD receives more helpful benefit and incurs less harmful penalty than the ensemble, even though the ensemble predicts better. The distinction concerns their joint response to distant text, not an accuracy advantage for mixKD. The analysis below tests whether changing the ensemble weight can recover that joint response.

How this joint response develops during mixed-teacher training remains an open question. In particular, whether interactions between the two teacher losses explain it requires controlled investigation.

C.4.1 Exploratory fixed-weight ensemble analysis

We next ask whether the equal-weight result is explained by the choice of ensemble weight. This exploratory post-test analysis keeps the same selected checkpoints, source-provided test targets, and donor-defined populations in all nine corpus–horizon–seed settings. It uses cached observed-token losses, without new training or model inference. Each setting may choose its own weight, but one weight must satisfy both contextual requirements.

For the same-seed sKD and eqKD students, the ensemble family is

$$p_{E_w} = w p_{\text{sKD}} + (1-w) p_{\text{eqKD}}, \qquad w \in [0,1]. \tag{9}$$

The short-student weight w is constant across targets and populations and is the same under full and restricted input. It is distinct from the training mixture weight λ. We average unsoftened probabilities, not losses or contextual-gains. Let $B_X = \overline{G}_{X,\text{help}}$ be method X's mean helpful contextual benefit and $P_X = -\overline{G}_{X,\text{harm}}$ its signed harmful contextual penalty. A negative P_X means that distant text is beneficial on average even within that donor-defined harmful population; values are not clipped. Joint attainment requires $B_{E_w} \geq B_{\text{mixKD}}$ and $P_{E_w} \leq P_{\text{mixKD}}$. We summarize the two shortfalls by

$$\begin{aligned} d(w) &= \max\{B_{\text{mixKD}} - B_{E_w},\ P_{E_w} - P_{\text{mixKD}}\}, \\ d_* &= \min_{w \in [0,1]} d(w). \end{aligned} \tag{10}$$

Positive d_* means every weight misses at least one requirement. It does not mean that mixKD outperforms every ensemble on both contextual measures, or on prediction loss.

Numerical coverage. The search starts at $w = 0, 0.01, \ldots, 1$ and the previously evaluated $w = 0.375$, then refines intervals that may contain a feasibility boundary or a better joint compromise. Each setting reaches the budget of 512 evaluated weights. Individual observed-token losses are monotone in w, so their endpoint minima and maxima conservatively bound contextual-gains throughout each interval. The resulting bounds exclude joint attainment over all of $[0,1]$, conditional on the stored losses. This conclusion does not assume monotonic contextual-gains or interpolate a proof from the plotted curve. The numerical padding is 10^{-10} nats. Although only two settings attain the requested 10^{-5}-nat precision for the minimum, every returned minimum enclosure is narrower than 4.5×10^{-5} nats and has a strictly positive lower bound. Thus, the remaining optimization precision does not affect exclusion. Float64 mixture arithmetic does not restore precision absent from the float32 cached losses.

Uncertainty and scope. For each setting, 2,000 new paired resamples of the original 1024-token blocks are shared across weights, methods, and input conditions, including mixKD. Donor masks remain fixed; conditional means use the sampled population target counts. The 95th percentile of the maximum absolute centered bootstrap deviation across both shortfall curves and all evaluated weights defines an approximate simultaneous band. It yields the finite-grid minimum interval in Table 29; every lower endpoint is positive. These bands concern within-seed uncertainty conditional on the checkpoints and masks. The grids were frozen before resampling but refined using the observed test scores, so the bands do not adjust for grid-selection effects. They are neither confidence statements over the full weight continuum nor simultaneous coverage across all nine settings. The numerical full-interval exclusion and the grid-based uncertainty are separate results.

Findings. No constant convex mixture jointly matches mixKD's contextual benefit and penalty in any setting (Figure 4). Both numerical exclusion and grid-based support persist when each shortfall is allowed a margin of 0.0001, 0.001, or 0.005 nats. These predeclared margins are descriptive scale checks, not validated equivalence thresholds. The best evaluated joint shortfalls range from 0.0181 to 0.0875 nats. Nevertheless, every corresponding ensemble has lower overall CE than mixKD, with improvements of 0.0474–0.0638 nats. Its weight is selected on the test split for this descriptive comparison, not validation-selected for an unbiased test estimate. The findings distinguish contextual response from predictive accuracy; they do not establish a training mechanism or exclude input-dependent weights, other ensemble members, logit mixtures, or different training regimes.

Table 29: Exploratory constant-weight student ensembles on source-provided test data. w_* is the best evaluated joint-compromise weight and $d_{*,\mathrm{grid}}$ its shortfall. L_{cont} is a conservative numerical lower bound for d_* over $[0, 1]$, rounded down. The 95% interval instead concerns the evaluated-grid minimum, from the two-curve simultaneous band. $I_{\mathrm{CE}} = \mathrm{CE}_{\mathrm{mixKD}} - \mathrm{CE}_{E_{w_*}}$ is the ensemble's overall CE improvement over mixKD. All losses and differences are in nats per target; positive lower bounds exclude joint attainment, while positive I_{CE} favors ensemble prediction.

Seed	w_*	$d_{*,\mathrm{grid}}$	L_{cont}	Grid minimum 95% interval	I_{CE}
WikiText-103, $H = h = 32$					
42	0.4134	0.0875	0.08744	[0.0764, 0.0985]	+0.0574
99	0.4462	0.0293	0.02931	[0.0192, 0.0395]	+0.0573
123	0.4242	0.0609	0.06085	[0.0497, 0.0720]	+0.0535
WikiText-103, $H = h = 64$					
42	0.4194	0.0491	0.04904	[0.0412, 0.0569]	+0.0552
99	0.4780	0.0358	0.03577	[0.0277, 0.0439]	+0.0530
123	0.4841	0.0524	0.05237	[0.0440, 0.0608]	+0.0565
SlimPajama-100M, $H = h = 64$					
42	0.4938	0.0181	0.01801	[0.0165, 0.0196]	+0.0638
99	0.3908	0.0508	0.05079	[0.0489, 0.0528]	+0.0480
123	0.3599	0.0187	0.01869	[0.0166, 0.0209]	+0.0474

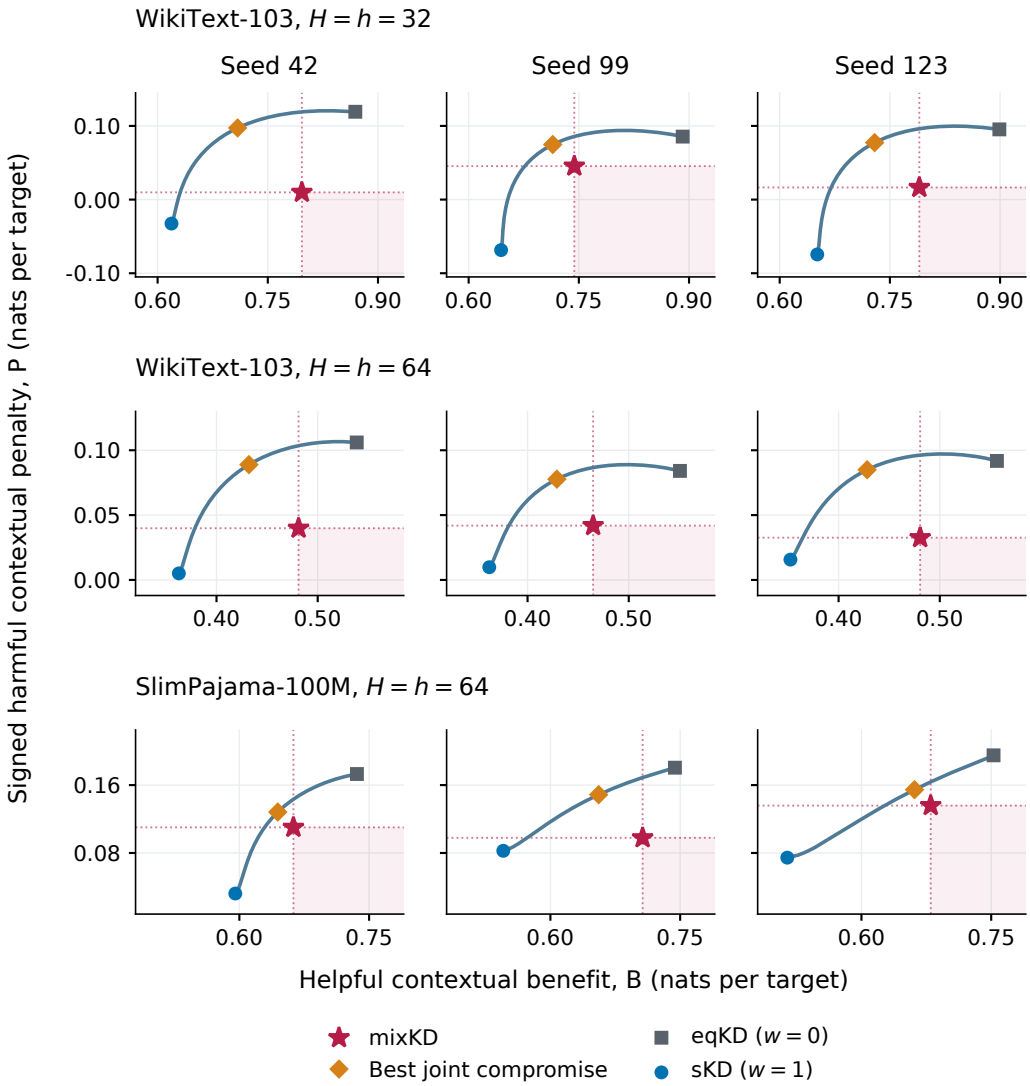

Figure 4: Contextual responses attainable by constant probability mixtures of the same-seed sKD and eqKD students. Each panel shows one seed and setting. Shading marks the region with at least mixKD's helpful benefit and no greater harmful penalty, not a region of better prediction. No ensemble curve enters that region. Lines connect evaluated weights; numerical exclusion over the full interval is established separately by the bounds in Table 29. Markers identify mixKD, the component students, and the best evaluated joint compromise. These are empirical curves, not confidence regions. Negative signed penalties mean contextual benefit on the donor-defined harmful population.

D ADDITIONAL SOURCE-PROVIDED WIKITEXT-103 VALIDATION

This evaluation uses the same internally selected checkpoints as the test evaluation, without training, reselection, or changes to mixture weights or analysis rules. Both horizons and all three seeds are included. Baseline and eqKD checkpoints and their full-input predictions are reused across horizons, so these are not additional training replications.

The entire source-provided validation file supplies 241 blocks and 216,177 common-tail targets. It lies within the historical internal held-out partition, not training or checkpoint-selection validation (Appendix B). This is an additional evaluation population at fixed checkpoints, not a claim of never-evaluated text or a switch to the standard WikiText training protocol. Labels are reconstructed from the other two baseline seeds on these targets and held fixed across methods. The original 1,000 paired block resamples quantify within-seed evaluation uncertainty, not training-seed variability.

D.1 PREDICTION AND POPULATION ACCOUNTING

mixKD improves helpful-position prediction over sKD and harmful-position prediction over eqKD in all six seed–horizon comparisons, with positive paired intervals. It remains worse than eqKD on helpful positions. Overall mixKD CE is lower in five of six comparisons, with positive intervals in those five. The seed-42 disadvantage at $H = h = 32$ has an interval spanning zero. Point rankings match test, although interval outcomes differ; at $H = h = 32$, seed 123's interval excludes zero here but not on test. Discordant/zero positions remain in the tables. Multiplying each conditional improvement by its count divided by 216,177 recovers its population contribution; the three contributions sum to overall improvement before rounding.

Table 30: Source-provided WikiText-103 validation CE and mixKD improvement over eqKD. CE is nats per common-tail target; positive improvement favors mixKD. Bounds are canonical paired-block 95% intervals. Baseline and eqKD full-input scores are reused across horizons.

$H = h$	Seed	Baseline	sKD	eqKD	mixKD	mixKD improvement [95% CI]
32	42	3.20553	3.15824	3.06860	3.07111	-0.002515 [-0.005982, +0.001033]
32	99	3.20430	3.14788	3.06968	3.06414	+0.005541 [+0.001906, +0.009101]
32	123	3.21070	3.15980	3.07930	3.07257	+0.006731 [+0.002149, +0.011067]
64	42	3.20553	3.13004	3.06860	3.06103	+0.007566 [+0.003952, +0.011413]
64	99	3.20430	3.13096	3.06968	3.05722	+0.012467 [+0.008702, +0.016214]
64	123	3.21070	3.13178	3.07930	3.06693	+0.012372 [+0.008286, +0.016528]

Table 31: Source-provided validation conditional full-input CE improvements at $H = h = 32$. Positive values favor mixKD over the named comparator. All three donor-defined populations are retained, with target counts and paired-block 95% intervals. Entries are conditional means, not population-weighted contributions.

Seed	Population	n	Over sKD [95% CI]	Over eqKD [95% CI]
42	Helpful	101,483	+0.15794 [+0.14999, +0.16553]	-0.06003 [-0.06579, -0.05420]
42	Harmful	38,566	-0.01566 [-0.02408, -0.00748]	+0.08908 [+0.08055, +0.09842]
42	Discordant/zero	76,128	+0.04480 [+0.03950, +0.05020]	+0.02776 [+0.02225, +0.03329]
99	Helpful	101,433	+0.15394 [+0.14476, +0.16261]	-0.04365 [-0.04841, -0.03828]
99	Harmful	38,865	-0.01300 [-0.02111, -0.00485]	+0.08837 [+0.08015, +0.09605]
99	Discordant/zero	75,879	+0.03945 [+0.03281, +0.04565]	+0.02888 [+0.02297, +0.03455]
123	Helpful	101,511	+0.16410 [+0.15590, +0.17294]	-0.04164 [-0.04787, -0.03584]
123	Harmful	38,864	-0.02114 [-0.02980, -0.01269]	+0.08839 [+0.07929, +0.09714]
123	Discordant/zero	75,802	+0.03984 [+0.03388, +0.04603]	+0.02964 [+0.02358, +0.03543]

Table 32: Source-provided validation conditional full-input CE improvements at $H = h = 64$, using the same conventions as Table 31. Populations change with the diagnostic cutoff h but are shared across methods within each comparison.

Seed	Population	n	Over sKD [95% CI]	Over eqKD [95% CI]
42	Helpful	88,867	+0.12538 [+0.11769, +0.13366]	-0.03915 [-0.04521, -0.03352]
42	Harmful	43,972	-0.00605 [-0.01299, +0.00100]	+0.05781 [+0.05018, +0.06675]
42	Discordant/zero	83,338	+0.04849 [+0.04291, +0.05364]	+0.03087 [+0.02536, +0.03651]
99	Helpful	88,403	+0.13111 [+0.12303, +0.13946]	-0.02651 [-0.03258, -0.02072]
99	Harmful	43,954	+0.00900 [-0.00013, +0.01803]	+0.05765 [+0.05010, +0.06509]
99	Discordant/zero	83,820	+0.04719 [+0.04153, +0.05326]	+0.02989 [+0.02487, +0.03507]
123	Helpful	88,588	+0.12207 [+0.11314, +0.13054]	-0.02508 [-0.03089, -0.01899]
123	Harmful	43,957	+0.00055 [-0.00652, +0.00810]	+0.06618 [+0.05801, +0.07356]
123	Discordant/zero	83,632	+0.03804 [+0.03261, +0.04350]	+0.02376 [+0.01894, +0.02905]

D.2 CONTEXTUAL EFFECTS AND ENSEMBLE BOUNDARY

Signed mean G_M is reported separately from prediction improvement. mixKD obtains more mean contextual benefit than sKD on helpful positions and a smaller penalty than eqKD on harmful positions. At $H = h = 32$, sKD has positive mean G_M on the donor-defined harmful population, illustrating that donor labels do not fix the sign of every evaluated model's effect. Intervals on the contextual means are not paired between-method intervals.

The equal probability prediction ensemble has lower CE than mixKD in every comparison, with paired intervals excluding zero. As on test, this limits an accuracy-superiority interpretation of mixKD but does not negate the complementary benefits combined in one student. No measured speed or cost advantage is inferred.

Table 33: Signed mean G_M at $H = h = 32$ on source-provided validation, with existing pointwise block-bootstrap 95% intervals. Positive values indicate contextual benefit and negative values a penalty. Counts are in Table 31.

Seed	Method	Helpful [95% CI]	Harmful [95% CI]	Discordant/zero [95% CI]
42	Baseline	0.8544 [0.8284, 0.8823]	-0.1834 [-0.1933, -0.1736]	0.1304 [0.1236, 0.1372]
42	sKD	0.6155 [0.5963, 0.6374]	0.0287 [0.0207, 0.0370]	0.1818 [0.1754, 0.1886]
42	mixKD	0.7874 [0.7633, 0.8135]	-0.0200 [-0.0281, -0.0116]	0.2119 [0.2052, 0.2189]
42	eqKD	0.8632 [0.8385, 0.8909]	-0.1341 [-0.1429, -0.1251]	0.1735 [0.1671, 0.1797]
99	Baseline	0.8432 [0.8186, 0.8694]	-0.1758 [-0.1861, -0.1661]	0.1281 [0.1214, 0.1348]
99	sKD	0.6371 [0.6163, 0.6593]	0.0642 [0.0561, 0.0728]	0.2082 [0.2008, 0.2154]
99	mixKD	0.7290 [0.7064, 0.7549]	-0.0561 [-0.0645, -0.0477]	0.1709 [0.1648, 0.1773]
99	eqKD	0.8782 [0.8515, 0.9056]	-0.1018 [-0.1112, -0.0921]	0.2007 [0.1931, 0.2083]
123	Baseline	0.8279 [0.8044, 0.8554]	-0.1757 [-0.1848, -0.1667]	0.1260 [0.1196, 0.1324]
123	sKD	0.6460 [0.6255, 0.6682]	0.0623 [0.0546, 0.0699]	0.2084 [0.2013, 0.2162]
123	mixKD	0.7713 [0.7478, 0.7983]	-0.0315 [-0.0399, -0.0228]	0.1934 [0.1861, 0.2000]
123	eqKD	0.8885 [0.8640, 0.9166]	-0.1070 [-0.1175, -0.0973]	0.1898 [0.1828, 0.1977]

Table 34: Signed mean G_M at $H = h = 64$ on source-provided validation, with the same conventions as Table 33. Counts are in Table 32.

Seed	Method	Helpful [95% CI]	Harmful [95% CI]	Discordant/zero [95% CI]
42	Baseline	0.5380 [0.5178, 0.5592]	-0.1349 [-0.1425, -0.1278]	0.0716 [0.0670, 0.0766]
42	sKD	0.3571 [0.3432, 0.3723]	-0.0032 [-0.0087, 0.0027]	0.0906 [0.0860, 0.0954]
42	mixKD	0.4798 [0.4612, 0.4991]	-0.0386 [-0.0449, -0.0323]	0.1143 [0.1090, 0.1200]
42	eqKD	0.5350 [0.5153, 0.5565]	-0.1121 [-0.1189, -0.1052]	0.0915 [0.0866, 0.0969]
99	Baseline	0.5310 [0.5115, 0.5503]	-0.1318 [-0.1389, -0.1250]	0.0732 [0.0685, 0.0779]
99	sKD	0.3546 [0.3412, 0.3701]	-0.0112 [-0.0164, -0.0062]	0.0900 [0.0854, 0.0946]
99	mixKD	0.4593 [0.4418, 0.4789]	-0.0414 [-0.0473, -0.0357]	0.1018 [0.0961, 0.1073]
99	eqKD	0.5454 [0.5262, 0.5654]	-0.0938 [-0.1013, -0.0866]	0.1036 [0.0981, 0.1094]
123	Baseline	0.5185 [0.5002, 0.5382]	-0.1314 [-0.1387, -0.1244]	0.0676 [0.0630, 0.0722]
123	sKD	0.3479 [0.3342, 0.3623]	-0.0182 [-0.0234, -0.0131]	0.0794 [0.0755, 0.0834]
123	mixKD	0.4702 [0.4534, 0.4907]	-0.0382 [-0.0442, -0.0317]	0.1058 [0.1005, 0.1113]
123	eqKD	0.5433 [0.5241, 0.5655]	-0.0954 [-0.1031, -0.0880]	0.1019 [0.0969, 0.1069]

Table 35: Source-provided validation comparison with the equal-probability ensemble of trained sKD and eqKD students. Excess loss is $\mathrm{CE}_{\mathrm{mixKD}} - \mathrm{CE}_{\mathrm{ensemble}}$; positive values favor the ensemble. Intervals are the existing paired direct CE-difference bounds, not the attribution-gain intervals used above.

$H = h$	Seed	Ensemble CE	mixKD CE	Excess loss [95% CI]
32	42	3.01570	3.07111	+0.055414 [+0.052564, +0.058474]
32	99	3.01124	3.06414	+0.052906 [+0.049778, +0.055935]
32	123	3.02739	3.07257	+0.045179 [+0.041532, +0.048938]
64	42	3.00851	3.06103	+0.052527 [+0.049966, +0.055332]
64	99	3.00908	3.05722	+0.048135 [+0.044971, +0.051428]
64	123	3.01383	3.06693	+0.053101 [+0.050234, +0.055886]

E Population and Dependence Robustness

These WikiText-103 analyses are post-test sensitivities using cached scores, not a new prespecified confirmation. Checkpoints, target alignment, and the original zero-margin consensus comparison are unchanged. Donor pairs are 99/123 for evaluated seed 42, 42/123 for seed 99, and 42/99 for seed 123. Their overlap and shared architecture/data do not make them independent semantic adjudicators. These checks were not repeated on SlimPajama-100M.

E.1 Label rules and donor agreement

We test each non-native donor alone and two-donor consensus at margins $\tau \in \{0, .01, .025, .05\}$ nats. Helpful membership requires the applicable donor effects above τ, harmful membership below $-\tau$; single-donor rules use strict signs. Opposite-sign and remaining low-margin/zero positions stay in the residual population. The $\tau = 0$ consensus rule recovers the primary partition. No evaluated-baseline disagreements or residual targets are silently removed.

Both focal conditional prediction improvements retain positive paired interval bounds in all 36 seed–horizon–rule cells, separately for each contrast. Tables 36–37 retain every rule and its coverage, rather than selecting a favorable margin. Donor agreement does not ensure that the evaluated seed's own baseline experiences the same sign or exceeds the same margin at every selected position. The populations describe reproducibility across the chosen donors, not independently validated semantic helpfulness or harmfulness.

Table 36: WikiText-103 test label sensitivity at $H = h = 32$. d denotes one donor and τ a two-donor consensus margin. Improvements are conditional nats per target with pointwise paired 95% intervals. Coverage is percent of all 247,572 targets.

Seed	Rule	mixKD over sKD Helpful improvement	mixKD over eqKD Harmful improvement	Helpful coverage (%)	Harmful coverage (%)
42	$d = 99$	+0.1344 [+0.1281, +0.1414]	+0.0547 [+0.0485, +0.0606]	64.5	35.5
42	$d = 123$	+0.1330 [+0.1263, +0.1399]	+0.0556 [+0.0494, +0.0614]	64.3	35.6
42	$\tau = 0$	+0.1646 [+0.1565, +0.1734]	+0.0835 [+0.0751, +0.0922]	46.7	17.9
42	$\tau = 0.01$	+0.1841 [+0.1753, +0.1934]	+0.0907 [+0.0818, +0.1003]	41.3	16.3
42	$\tau = 0.025$	+0.1912 [+0.1820, +0.2012]	+0.0941 [+0.0846, +0.1036]	39.3	15.3
42	$\tau = 0.05$	+0.1998 [+0.1903, +0.2106]	+0.0974 [+0.0875, +0.1082]	36.8	14.0
99	$d = 42$	+0.1260 [+0.1189, +0.1340]	+0.0592 [+0.0540, +0.0645]	64.4	35.6
99	$d = 123$	+0.1270 [+0.1196, +0.1349]	+0.0555 [+0.0504, +0.0612]	64.3	35.6
99	$\tau = 0$	+0.1565 [+0.1479, +0.1658]	+0.0844 [+0.0767, +0.0925]	46.7	18.0
99	$\tau = 0.01$	+0.1749 [+0.1658, +0.1845]	+0.0917 [+0.0835, +0.1003]	41.4	16.6
99	$\tau = 0.025$	+0.1822 [+0.1727, +0.1924]	+0.0951 [+0.0867, +0.1038]	39.4	15.5
99	$\tau = 0.05$	+0.1912 [+0.1815, +0.2017]	+0.1008 [+0.0920, +0.1101]	37.0	14.2
123	$d = 42$	+0.1320 [+0.1258, +0.1389]	+0.0599 [+0.0537, +0.0659]	64.4	35.6
123	$d = 99$	+0.1329 [+0.1266, +0.1400]	+0.0561 [+0.0501, +0.0623]	64.5	35.5
123	$\tau = 0$	+0.1659 [+0.1582, +0.1747]	+0.0848 [+0.0764, +0.0921]	46.7	17.9
123	$\tau = 0.01$	+0.1858 [+0.1775, +0.1952]	+0.0914 [+0.0826, +0.0996]	41.3	16.5
123	$\tau = 0.025$	+0.1927 [+0.1842, +0.2023]	+0.0960 [+0.0869, +0.1046]	39.3	15.5
123	$\tau = 0.05$	+0.2021 [+0.1933, +0.2118]	+0.1025 [+0.0925, +0.1117]	36.9	14.2

Table 37: WikiText-103 test label sensitivity at $H = h = 64$. d denotes one donor and τ a two-donor consensus margin. Improvements are conditional nats per target with pointwise paired 95% intervals. Coverage is percent of all 247,572 targets.

Seed	Rule	mixKD over sKD Helpful improvement	mixKD over eqKD Harmful improvement	Helpful coverage (%)	Harmful coverage (%)
42	$d = 99$	+0.1067 [+0.1009, +0.1136]	+0.0404 [+0.0345, +0.0463]	60.3	39.7
42	$d = 123$	+0.1069 [+0.1007, +0.1137]	+0.0436 [+0.0375, +0.0500]	60.1	39.9
42	$\tau = 0$	+0.1344 [+0.1270, +0.1430]	+0.0592 [+0.0514, +0.0673]	40.8	20.5
42	$\tau = 0.01$	+0.1536 [+0.1453, +0.1633]	+0.0655 [+0.0569, +0.0745]	35.0	18.2
42	$\tau = 0.025$	+0.1623 [+0.1535, +0.1728]	+0.0707 [+0.0618, +0.0801]	32.6	16.8
42	$\tau = 0.05$	+0.1731 [+0.1640, +0.1840]	+0.0767 [+0.0671, +0.0865]	29.8	14.9
99	$d = 42$	+0.1042 [+0.0974, +0.1110]	+0.0410 [+0.0354, +0.0468]	60.2	39.8
99	$d = 123$	+0.1048 [+0.0986, +0.1118]	+0.0409 [+0.0357, +0.0460]	60.1	39.9
99	$\tau = 0$	+0.1323 [+0.1241, +0.1411]	+0.0603 [+0.0526, +0.0678]	40.7	20.4
99	$\tau = 0.01$	+0.1518 [+0.1428, +0.1616]	+0.0645 [+0.0562, +0.0728]	34.9	18.3
99	$\tau = 0.025$	+0.1598 [+0.1502, +0.1701]	+0.0684 [+0.0595, +0.0770]	32.6	16.8
99	$\tau = 0.05$	+0.1708 [+0.1601, +0.1815]	+0.0728 [+0.0635, +0.0819]	29.8	15.0
123	$d = 42$	+0.1034 [+0.0972, +0.1108]	+0.0450 [+0.0399, +0.0508]	60.2	39.8
123	$d = 99$	+0.1022 [+0.0961, +0.1094]	+0.0466 [+0.0408, +0.0521]	60.3	39.7
123	$\tau = 0$	+0.1320 [+0.1241, +0.1417]	+0.0663 [+0.0585, +0.0741]	40.9	20.4
123	$\tau = 0.01$	+0.1522 [+0.1432, +0.1630]	+0.0732 [+0.0646, +0.0820]	35.0	18.2
123	$\tau = 0.025$	+0.1601 [+0.1505, +0.1712]	+0.0767 [+0.0678, +0.0858]	32.7	16.8
123	$\tau = 0.05$	+0.1721 [+0.1621, +0.1841]	+0.0815 [+0.0717, +0.0914]	29.9	15.0

E.2 Label-independent magnitude and tokenwise contraction

The all-target magnitude reduction is

$$D_{\mathrm{abs},M} = \mathrm{mean}(|G_B| - |G_M|).$$

It uses every evaluated target without donor labels or baseline-magnitude thresholds. Both sKD and mixKD reduce mean absolute contextual effect relative to the baseline at both horizons in every seed. sKD has the largest point-estimated reduction, while eqKD increases mean magnitude in seeds 99 and 123. The estimates use float64 full/restricted loss algebra; donor effects and eligibility retain the recorded score arithmetic.

Table 38: WikiText-103 test all-target magnitude reduction relative to the evaluated seed's baseline, with original paired 95% intervals. Positive means smaller mean absolute contextual effect, not a larger fraction of contracting positions. No helpful/harmful classification or threshold exclusion is used.

$H = h$	Seed	sKD	eqKD	mixKD
32	42	+0.1396 [+0.1324, +0.1471]	+0.0056 [+0.0015, +0.0097]	+0.0404 [+0.0357, +0.0453]
32	99	+0.0950 [+0.0879, +0.1017]	-0.0296 [-0.0340, -0.0255]	+0.0903 [+0.0857, +0.0946]
32	123	+0.0898 [+0.0831, +0.0966]	-0.0359 [-0.0401, -0.0318]	+0.0382 [+0.0330, +0.0436]
64	42	+0.1135 [+0.1092, +0.1188]	+0.0150 [+0.0119, +0.0181]	+0.0417 [+0.0381, +0.0458]
64	99	+0.1142 [+0.1099, +0.1190]	-0.0093 [-0.0126, -0.0058]	+0.0515 [+0.0482, +0.0549]
64	123	+0.1315 [+0.1271, +0.1365]	-0.0146 [-0.0177, -0.0116]	+0.0340 [+0.0302, +0.0376]

Average attenuation does not imply more tokenwise contraction. Literal contraction means $|G_M| < |G_B|$, with ties counted as non-contraction. On SlimPajama-100M test seed 99, mixKD reduces the mean harmful contextual penalty from 0.1805 nats under eqKD to 0.0979 nats. Yet mixKD contracts fewer positions than eqKD, 52.57% versus 54.89%, in the shared eligible harmful subset requiring the evaluated baseline's $|G_B| > .025$. Its magnitude-reduction difference from eqKD on that subset is -0.008575 nats, with paired interval $[-0.009508, -0.007693]$. Crucially, the mean penalties use the entire donor-defined harmful population, whereas contraction frequencies and this magnitude contrast use the additional baseline-magnitude restriction. Neither result establishes uniform tokenwise attenuation, and comparing them does not identify a causal mechanism.

E.3 Block dependence and interval interpretation

Because reliable document assignments are unavailable, the dependence check groups four adjacent evaluation blocks in their original order. The 276 blocks form 69 non-overlapping clusters. Both schemes use 1,000 paired PCG64 draws with percentile endpoints, preserve model/donor alignment, recompute conditional denominators within each draw, and retain target weighting. Reported intervals are pointwise and conditional on trained checkpoints, not simultaneous or training-seed intervals.

Both focal conditional improvements retain positive lower bounds in every seed and horizon under four-block resampling. At $H = h = 64$, all three overall improvements retain positive intervals; at $H = h = 32$, only seed 99's overall interval is positive. A supporting magnitude ordering is less robust. The $H = h = 32$ seed-99 mixKD-minus-sKD all-target magnitude-reduction difference is -0.004642 nats, with block interval $[-0.008820, -0.000435]$ but four-block interval $[-0.010541, +0.000609]$. We therefore do not claim uniformly interval-supported sKD/mixKD magnitude ordering.

Table 39: WikiText-103 test dependence sensitivity at $H = h = 32$ for the primary prediction contrasts. Slash names compare the first method with the second; positive favors the first. Both columns give prediction improvement and pointwise 95% intervals on the indicated population.

Seed	Contrast and population	Paired blocks	Four-block clusters
42	mixKD/eqKD overall	-0.00043 [-0.00483, +0.00387]	-0.00043 [-0.00549, +0.00444]
42	mixKD/sKD helpful	+0.16462 [+0.15646, +0.17299]	+0.16462 [+0.15245, +0.17764]
42	mixKD/eqKD harmful	+0.08352 [+0.07521, +0.09163]	+0.08352 [+0.07470, +0.09324]
99	mixKD/eqKD overall	+0.00423 [+0.00045, +0.00784]	+0.00423 [+0.00006, +0.00904]
99	mixKD/sKD helpful	+0.15654 [+0.14679, +0.16634]	+0.15654 [+0.14430, +0.17016]
99	mixKD/eqKD harmful	+0.08439 [+0.07629, +0.09202]	+0.08439 [+0.07681, +0.09243]
123	mixKD/eqKD overall	+0.00312 [-0.00067, +0.00665]	+0.00312 [-0.00082, +0.00710]
123	mixKD/sKD helpful	+0.16594 [+0.15814, +0.17404]	+0.16594 [+0.15453, +0.17640]
123	mixKD/eqKD harmful	+0.08475 [+0.07734, +0.09190]	+0.08475 [+0.07696, +0.09275]

Table 40: WikiText-103 test dependence sensitivity at $H = h = 64$ for the primary prediction contrasts. Slash names compare the first method with the second; positive favors the first. Both columns give prediction improvement and pointwise 95% intervals on the indicated population.

Seed	Contrast and population	Paired blocks	Four-block clusters
42	mixKD/eqKD overall	+0.00889 [+0.00489, +0.01306]	+0.00889 [+0.00400, +0.01297]
42	mixKD/sKD helpful	+0.13438 [+0.12612, +0.14195]	+0.13438 [+0.12494, +0.14491]
42	mixKD/eqKD harmful	+0.05922 [+0.05136, +0.06699]	+0.05922 [+0.05061, +0.06667]
99	mixKD/eqKD overall	+0.00989 [+0.00592, +0.01353]	+0.00989 [+0.00610, +0.01361]
99	mixKD/sKD helpful	+0.13228 [+0.12451, +0.14124]	+0.13228 [+0.12079, +0.14424]
99	mixKD/eqKD harmful	+0.06026 [+0.05289, +0.06801]	+0.06026 [+0.05287, +0.06778]
123	mixKD/eqKD overall	+0.00983 [+0.00602, +0.01370]	+0.00983 [+0.00516, +0.01424]
123	mixKD/sKD helpful	+0.13204 [+0.12403, +0.14114]	+0.13204 [+0.12154, +0.14338]
123	mixKD/eqKD harmful	+0.06630 [+0.05893, +0.07370]	+0.06630 [+0.05827, +0.07454]

Interpretation boundary. Uniform shrinkage $G_M = cG_B$ with $0 < c < 1$ produces $\Delta G_M = (c-1)G_B$, negative for positive baseline effects and positive for negative ones. Opposite signed changes therefore do not establish semantic selectivity. Cross-seed labeling does not remove this possibility. Prediction improvement also depends on restricted-input performance through Eq. 5; these measurements describe joint outcomes, not an identified linguistic or causal mechanism.

F PREDICTIVE-ANCHORING DIAGNOSTIC

This diagnostic asks whether the direction of short-context guidance at a fixed student is consistent with predictive anchoring (Section 3.3). It measures local output-space directions at selected final checkpoints, not how the behavior develops during training.

Comparison and sample. At each selected mixKD student trained with $H = 64$, we compare three targets. The original full-context teacher receives its full causal input, the separately pretrained short teacher receives physical 64-token windows, and the same full teacher receives those windows with its weights fixed. The full teacher is the CE-only baseline used by eqKD, not the trained eqKD student. Runtime restriction isolates an immediate input-access change at fixed teacher weights, including sensitivity to a shorter input sequence. The common student is not neutral, because it was trained with both original teachers.

We use seeds 42, 99, and 123 and a frozen inventory of 40 source-provided validation blocks per corpus, shared across methods and seeds. WikiText-103 uses its source-provided validation, not the internal selection split. Each block provides 1,024 CE positions and 897 aligned KD targets, giving 35,880 diagnostic targets per seed. The same student is also evaluated on physical 64-token windows. Full input means the available causal prefix within the block, not 1,024 preceding tokens at every position.

Residual magnitude and direction. Let $z_{F,i}$ and $z_{64,i}$ be the student's unsoftened logits for the same target under full and restricted input, $p_{T,i} = \mathrm{softmax}(z_{F,i}/T)$, and $q_{T,i}$ the alternative teacher distribution at $T = 2$. We report mean residual size $\|p_{T,i} - q_{T,i}\|_2$. With one block per diagnostic batch and equal comparison coefficient 0.5 for each alternative,

$$
\begin{aligned}
g_{\mathrm{KD},i} &= \frac{0.5T}{897}(p_{T,i} - q_{T,i}),\\
d_i &= (z_{F,i} - z_{64,i}) - \frac{1}{V}\sum_{v=1}^{V}(z_{F,i,v} - z_{64,i,v})\mathbf{1},\\
R_i &= g_{\mathrm{KD},i}^{\top} d_i, \qquad c_i = \frac{R_i}{\|g_{\mathrm{KD},i}\|_2 \|d_i\|_2}.
\end{aligned}
\tag{11}
$$

Here V is vocabulary size and centering removes a vocabulary-wide logit shift. Positive R_i means that KD descent along $-g_{\mathrm{KD},i}$ has a component opposing the student's full-minus-restricted logit change. We report its frequency and mean directional cosine c_i; no cosine is undefined in this sample. These are not simulated parameter updates. Actual mixKD training assigns coefficient 0.25 to each teacher's T^2-scaled KL term.

Finding and its boundary. At the same mixKD student, short-context guidance opposes the contextual logit change more often than full-context guidance in every seed of both corpora. Restricting the full teacher's runtime input also increases opposition in every seed. Table 42 retains paired intervals for both contrasts. The mean cosines are small, so these frequencies do not establish strong vector alignment or overwhelming pressure. Moreover, WikiText-103 does not show a consistent sKD-over-eqKD opposition-frequency ordering when each teacher is evaluated against its own trained student. The fixed-student result cannot be transferred to that different-student comparison.

Intervals use 1,000 paired bootstrap resamples of the shared 40 blocks, with seed 20260918 and no pooling across training seeds. They quantify uncertainty within this sample, not training-seed uncertainty or guaranteed whole-corpus coverage.

Table 41: Reference comparisons at the same selected mixKD student within each seed. Entries are descriptive arithmetic means of three seedwise estimates. Residuals are softened-probability L2 distances at $T = 2$; opposition is the percentage of targets with $R_i > 0$. "Same full" reuses the full teacher's weights with restricted input. No pooled confidence interval is attached to these means.

Teacher weights / input	Residual L2	Opposition (%)	Mean cosine
WikiText-103			
Full / full prefix	0.03028	36.6	-0.00793
Short / 64 tokens	0.03134	66.2	$+0.01574$
Same full / 64 tokens	0.03322	78.0	$+0.02624$
SlimPajama-100M			
Full / full prefix	0.02387	37.1	-0.00917
Short / 64 tokens	0.02600	75.2	$+0.02799$
Same full / 64 tokens	0.02496	84.1	$+0.04088$

Table 42: Paired changes in opposition frequency at the common mixKD student, in percentage points with 95% block-bootstrap intervals. Both contrasts subtract the full teacher on full input. The second changes only that teacher's runtime input to 64-token windows.

Seed	Short teacher minus full teacher	Full teacher on 64 tokens minus full input
WikiText-103		
42	29.26 [27.64, 31.07]	42.24 [40.57, 43.79]
99	31.10 [29.08, 33.05]	42.28 [40.61, 43.82]
123	28.34 [25.84, 30.56]	39.39 [37.62, 41.03]
SlimPajama-100M		
42	38.72 [35.49, 41.76]	48.85 [45.67, 51.94]
99	38.14 [34.99, 41.29]	46.61 [43.73, 49.48]
123	37.50 [34.48, 40.72]	45.38 [42.68, 48.18]

The fixed-student contrast supports the anchoring interpretation at the measured checkpoints. It does not isolate the historical effect of short-context pretraining from runtime restriction, and opposition is not intrinsically beneficial because distant text can help or harm prediction. The host student was trained with both references, and all measurements are taken after checkpoint selection. These observations therefore do not establish when the directions arose, how parameter updates followed them, or whether they caused the learned contextual behavior.

G Supplementary WikiText-2 Comparisons

G.1 Reversal of the sKD–eqKD ordering

This supplementary comparison at $H = h = 128$ uses a four-layer, width-512, four-head decoder with 64,123,985 parameters, GPT-2 BPE (50,257 tokens), pre-layer normalization, RoPE, GELU feed-forward width $4d$, zero dropout, final layer normalization, and separate embedding/output matrices. Source-provided train, validation, and test texts were concatenated before contiguous 80/10/10 token partitioning, rather than using the original benchmark splits. The common-tail validation probe covers 293 non-overlapping 1024-token blocks, or 262,821 targets per method and seed. Both conditions use matched fresh initialization, validation-CE selection, $\alpha = .5$, $T = 2$, and the same 897 KD positions. They differ in teacher regime, with $H = h = 128$ for sKD and equal causal input for eqKD.

Table 43: WikiText-2 full-input prediction improvement at $H = h = 128$ over the same-seed baseline, in nats per common-tail target. The last two columns compare sKD with eqKD using the original paired 95% block-bootstrap intervals. These probe losses are distinct from all-position checkpoint-selection losses.

Seed	eqKD improvement	sKD improvement	sKD over eqKD	95% interval
42	.051153	.080795	.029642	[.024835, .034656]
99	.070615	.122983	.052368	[.049002, .055765]
123	.019513	.098397	.078883	[.074110, .083608]

sKD has lower loss than eqKD in all three seeds, reversing their aggregate ordering in the main experiments. Architecture, corpus, and horizon differ, so this observation does not isolate why the balance changes. It establishes a boundary to a universal eqKD advantage, not a corpus-size or horizon law.

G.2 Separating teacher pretraining horizon from runtime access

Setup. This supplementary sKD experiment separates the teacher's pretraining horizon H_{train} from its runtime input length h_{run} during student distillation. It uses the WikiText-2 architecture and historical 80/10/10 partition described in Appendix G.1. We trained nine students, with seeds 42, 99, and 123 for each off-diagonal pair $(H_{\text{train}}, h_{\text{run}}) \in \{(64, 32), (128, 32), (128, 64)\}$. The corresponding diagonal comparisons use $(32, 32)$, $(64, 64)$, and $(128, 128)$. Frozen teachers were reused from the corrected pretraining runs. Students retained matched fresh initialization, a 1,024-token context, $\alpha = 0.5$, $T = 2$, and the same 897 KD target positions. Teacher inputs physically contained only h_{run} tokens. Checkpoints followed validation-CE selection with five-epoch early-stopping patience. The historical artifact names "Editor" and "Author" refer to the teacher and student, respectively.

Evaluation and accounting. All comparisons use the same 293 non-overlapping validation blocks and 262,821 target positions. These probe losses are not the all-position CE used for checkpoint selection, nor are they an independent test evaluation. We set the diagnostic cutoff $h = h_{\text{run}}$ and use the same-seed baseline in Eq. 5. Table 44 reports mean full-input prediction improvement $g_{\text{full},M}$, restricted-input improvement $g_{h,M}$, and change in contextual-gain ΔG_M. These correspond to

A_{gain}, L, and ΔG in the archived reports. Horizon comparisons at a fixed runtime length use the same restricted inputs and baseline; changing the runtime length also changes the diagnostic cutoff.

Table 44: WikiText-2 prediction improvement and contextual accounting, in nats per evaluated target. Each row is one trained student compared with its same-seed baseline. The last three columns satisfy $g_{\text{full},M} = g_{h,M} + \Delta G_M$ before rounding. Positive prediction improvement means lower CE; negative ΔG_M means less contextual-gain than the baseline. These are descriptive seedwise estimates, not a pooled significance claim.

H_{train}	h_{run}	Seed	$g_{\text{full},M}$	$g_{h,M}$	ΔG_M
32	32	42	0.0785	0.1450	-0.0665
32	32	99	0.0794	0.1442	-0.0648
32	32	123	0.0718	0.1470	-0.0752
64	32	42	0.0834	0.1563	-0.0729
64	32	99	0.1044	0.1624	-0.0580
64	32	123	0.0917	0.1653	-0.0737
128	32	42	0.0748	0.1505	-0.0757
128	32	99	0.1101	0.1554	-0.0453
128	32	123	0.0858	0.1670	-0.0812
64	64	42	0.0845	0.1305	-0.0459
64	64	99	0.1006	0.1354	-0.0349
64	64	123	0.0141	0.0763	-0.0621
128	64	42	0.1087	0.1406	-0.0319
128	64	99	0.1038	0.1313	-0.0275
128	64	123	0.0992	0.1476	-0.0485
128	128	42	0.0808	0.1038	-0.0230
128	128	99	0.1230	0.1342	-0.0113
128	128	123	0.0984	0.1234	-0.0250

Prediction at fixed runtime access. All nine off-diagonal students improve over their baselines. At runtime length 32, a teacher pretrained on 64 tokens improves student prediction over the 32-token-pretrained teacher in all three seeds; the 128-token-pretrained teacher does so in two. At runtime length 64, the 128-token-pretrained teacher improves over the 64-token-pretrained teacher in all three seeds. Longer pretraining context is not uniformly better, however. At runtime length 32, the 128-token-pretrained teacher produces less improvement than the 64-token-pretrained teacher in two seeds. The teacher's learned predictor therefore matters beyond its runtime input length, without establishing a monotonic horizon rule.

Contextual tradeoff. All nine off-diagonal runs have $g_{h,M} > 0$ and $\Delta G_M < 0$. Their overall prediction improvement combines better restricted-input prediction with reduced average contextual-gain. To examine strongly helpful positions, we retain targets where both other-seed baselines have positive contextual-gain and rank them by the smaller of those two gains. The highest of five approximately equal-sized groups, denoted Q5, identifies the strongest donor-defined helpful positions. Labels and rankings do not use the evaluated student's outputs and are shared across teacher pretraining horizons at a fixed runtime length. Ties are resolved by target index.

Table 45 shows that contextual-gain remains below the same-seed baseline even within Q5 in all nine runs. Intervals use the original 1,000 paired resamples of the 293 validation blocks, with bootstrap seed 20260904 and population/quintile assignments held fixed. They describe within-seed block uncertainty, not training-seed variability or simultaneous coverage across comparisons. This check does not establish selective preservation of helpful context. The donor labels describe predictive effects, not semantic relevance.

Table 45: Change in contextual-gain on the strongest donor-defined helpful quintile in the nine off-diagonal WikiText-2 runs. Units are nats per Q5 target; n_{Q5} is the number of targets. Original paired 95% block-bootstrap intervals are retained. Every estimate and interval is below zero.

H_{train}	h_{run}	Seed	n_{Q5}	ΔG_M	95% interval
64	32	42	22,367	-0.4437	[-0.4630, -0.4239]
64	32	99	21,974	-0.3613	[-0.3780, -0.3449]
64	32	123	22,800	-0.4410	[-0.4591, -0.4231]
128	32	42	22,367	-0.4589	[-0.4782, -0.4381]
128	32	99	21,974	-0.3073	[-0.3230, -0.2917]
128	32	123	22,800	-0.4677	[-0.4857, -0.4500]
128	64	42	19,719	-0.2911	[-0.3078, -0.2726]
128	64	99	19,235	-0.2276	[-0.2418, -0.2146]
128	64	123	20,079	-0.3200	[-0.3363, -0.3047]

Interpretation and scope. Matching teacher pretraining and runtime horizons is not necessary for the observed combination of improved prediction and reduced contextual-gain. Changing the teacher's pretraining horizon at fixed runtime access can change student performance, while the contextual tradeoff persists. These supplementary single-teacher results do not establish an ideal pretraining horizon, explain mixKD's difference from inference ensembles, or identify a representation-level mechanism. Their scope is this WikiText-2 architecture, partition, and evaluated horizon grid.