BEFORE THE KL: AUDITING INTERFACES FOR CROSS-TOKENIZER DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

In alignment-based cross-tokenizer knowledge distillation (CTKD), the loss acts on virtual distributions constructed by an alignment and scoring interface. We show that alignment and native predictive equivalence are insufficient to certify fidelity to the learning objective. For fixed-support linear scores, we prove that invariance to position-wise logit shifts holds exactly when candidates assign equal total weight to each native position. An executable paired audit connects this criterion to the training signal. A pinned public SimCT implementation violates the condition by aggregating raw logits where its published formula uses log probabilities. We isolate this difference under shared alignment, candidate support, the raw-teacher mask and loss reduction. In a frozen Falcon3-to-Pythia audit on 224 texts, native probabilities change by at most 1.19×10^{-7}, yet median complete-gradient cosine is 0.8437. This separation demonstrates that native prediction checks can miss changes in the supervision delivered to the student. A single-seed matched short-training diagnostic produces distinct parameter paths. Raw-logit and log-probability evaluators reverse the ranking of the same final students; native NLL per response byte favors the raw-trained student. These findings establish score construction and its gradient path as explicit parts of a reproducible CTKD objective. Formula fidelity, evaluator preference and native predictive performance require separate validation.

1 INTRODUCTION

Cross-tokenizer knowledge distillation (CTKD) transfers supervision between models that segment the same text differently. Subword tokenization assigns model-specific units to text (Sennrich et al., 2016; Kudo & Richardson, 2018); a teacher token can cover several student tokens, and vocabulary indices need not share meaning. Existing methods bridge this mismatch through token or feature alignment, rank comparisons, path likelihoods and queried event probabilities (Zhang et al., 2024; Boizard et al., 2025; Minixhofer et al., 2025; Phan et al., 2026; On et al., 2026). Each constructs the object on which the loss acts. The interface is therefore part of the learning objective's mathematical specification.

This distinction exposes a gap between matching native predictions and reproducing a learning objective. Replacing logits by log probabilities preserves native softmax probabilities but subtracts a different constant at each position. When candidates aggregate positions differently, these constants can alter their relative scores. Text, alignment and native predictions can all agree despite a changed training signal. We study interface fidelity, the preservation of the declared score formula in the executed construction.

Our executable audit fixes parameters, inputs, alignment and candidate support, then follows a scoring intervention through native predictions, virtual distributions and complete student gradients. This tests whether predictive equivalence survives the construction of supervision. Matched training and evaluation separate the rule that produces a student from the rule that judges it.

For fixed-support linear scores, invariance holds exactly when all candidates assign the same total weight to each position. The criterion identifies when a representation choice can change the learning objective. A pinned public SimCT implementation violates it by averaging raw logits across span positions but reading only the first position for overlaps. Its published formula uses log probabilities

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.

(Sun et al., 2026). We compare the raw-logit and log-probability (LP) gauges under shared runtime support and masking.

On 224 fixed GSM8K texts, a Falcon3-to-Pythia audit finds a maximum native probability difference of 1.19×10^{-7} alongside a median complete-gradient cosine of 0.8437. These measurements establish that prediction-level agreement can conceal a change in optimization direction. Single-seed matched short training separates parameter paths. Raw and LP evaluators rank the same final students oppositely in every fold; native NLL favors the raw-trained arm. The comparison establishes that reproducing a score formula and preferring a trained student are distinct claims. Neither substitutes for direct evaluation of native predictive performance.

Our contributions connect a general invariance criterion, an executable audit and controlled SimCT evidence. The criterion identifies structural sensitivity; the audit isolates its realized effect; the experiments trace that effect through gradients, training and shared evaluation. Together, they establish why objective fidelity requires verification beyond token alignment and native predictions.

2 Related work

Distillation objectives and training distributions. Soft-target distillation transfers a teacher's output distribution (Hinton et al., 2015); sequence-level distillation transfers generated outputs (Kim & Rush, 2016). Imitation learning and on-policy distillation address the mismatch between training and generation (Lin et al., 2020; Agarwal et al., 2024). MiniLLM and f-divergence distillation study the discrepancy being minimized (Gu et al., 2024; Wen et al., 2023). Our audit concerns the score construction supplied to a chosen loss, complementing these objective and sampling choices.

Alignment and representation-based distillation. DSKD, contextual dynamic mapping, SeDi and span representation alignment bridge token or feature spaces (Zhang et al., 2024; Chen et al., 2025; Wang et al., 2026b; Dao et al., 2026). Residual-learning distillation aligns compressed hidden states through cross-model attention and combines residual supervision with fine-tuning (On et al., 2026). ULD compares ranks and ALM observed-path likelihoods (Boizard et al., 2025; Minixhofer et al., 2025). The relevant implementation checks depend on these declared objects.

Probability queries and support. Exact byte and token-to-character scoring formulate cross-tokenizer comparisons as event queries (Phan et al., 2025; Vieira et al., 2025). Phan et al. (2026) derive exact BPE conversion and an approximate general scorer, retaining unqueried mass in partial KL. BLD, X-Token and BPM use byte or projected interfaces (Singh et al., 2026; Sreenivas et al., 2026; Wang et al., 2026a); predictive querying and Sparse Logit Sampling study queried supervision (Boyd et al., 2022; Anshumann et al., 2025). These methods motivate explicit event definitions and accounting for probability outside the queried support.

Fidelity and assessment. SimCT provides a published scoring formula and public runtime to compare under fixed alignment (Sun et al., 2026). Distillation Traps separates supervision completeness from usefulness (Zhan et al., 2026). We distinguish formula fidelity from evaluator preference: useful raw-logit supervision can still differ from a declared log-probability rule. Complete gradients and shared evaluators measure the consequences of that difference.

3 Method

Let T and S denote teacher and student models. Their alignment interface constructs shared candidate scores and virtual distributions for the loss (Figure 1). We analyze this construction before specifying the paired audit.

3.1 Native outputs and virtual score construction

At native position j, let $z_j(v)$ be a logit, $p_j = \mathrm{softmax}(z_j)$, and $L_j = \log \sum_v \exp z_j(v)$; the log probability is $z_j(v) - L_j$. The interface constructs candidate scores a_i and virtual probabilities $q_i = \mathrm{softmax}(a/\tau)_i$. We ask whether equivalent representations of p_j produce the same q.

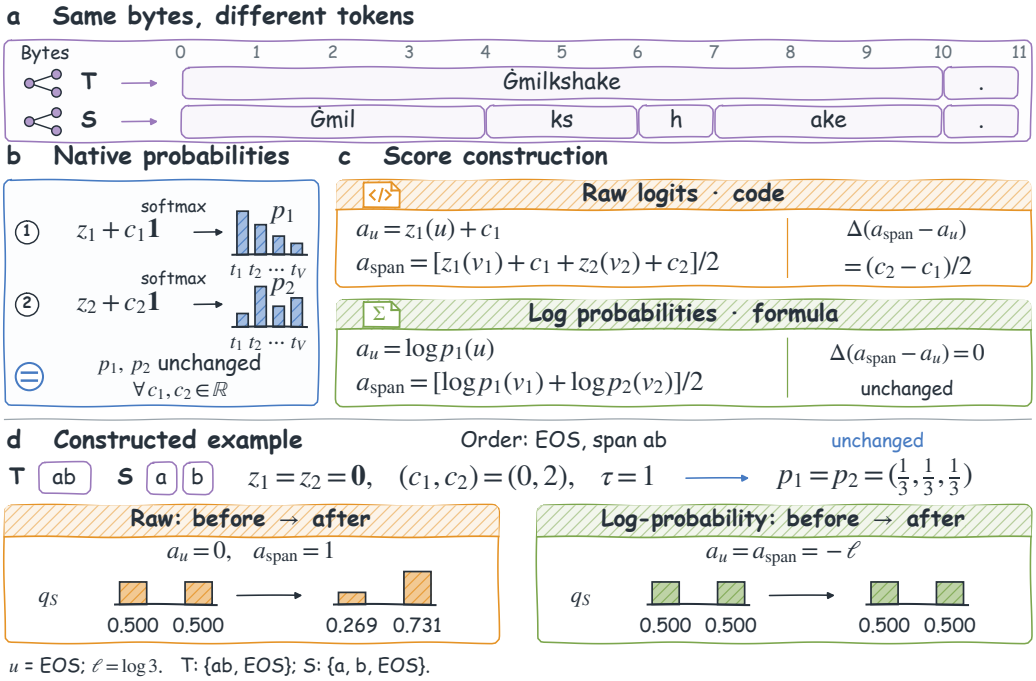


Figure 1: From token coordinates to the virtual scoring interface. **a**, An illustrative segmentation represents the same 11 UTF-8 bytes with two teacher tokens and five student tokens; offsets are relative to the displayed text. The dotted G marks a leading space in a token. **b,c**, Position-wise shifts preserve native probabilities but change the raw relative span/overlap score by $(c_2 - c_1)/2$; log-probability scores are invariant. **d**, Independent constructed example with text `ab`, fixed candidates (EOS, span `ab`) and shifts $(0, 2)$; scores are shown after the shift. Raw virtual probabilities change from $(0.500, 0.500)$ to $(0.269, 0.731)$; log-probability scores preserve them. Panels a and d are illustrative examples.

SimCT makes the comparison concrete because its candidate set contains both overlap tokens and aligned spans. For a span with realized tokens $v_1, \ldots, v_k$ and an overlap token u, the published formula (Sun et al., 2026, Eq. 7) and pinned runtime differ in the quantities they aggregate:

	Span	Overlap
Published	$\frac{1}{k}\sum_{j=1}^{k} \log p_j(v_j)$	$\log p_1(u)$
Pinned runtime	$a_{\text{span}} = \frac{1}{k}\sum_{j=1}^{k} z_j(v_j)$	$a_{\text{overlap}}(u) = z_1(u)$
Raw position weights	$(1/k, \ldots, 1/k)$	$(1, 0, \ldots, 0)$

(1)

Exponentiating the published span score gives the geometric mean of the conditional token probabilities; their product instead gives the likelihood of the realized token path. The mean therefore defines a length-normalized path score, not an exact byte-event probability. Its uniform position weights differ from the overlap score's first-position weight. Replacing z_j by $\log p_j$ consequently removes different combinations of normalizers despite identical alignment. The LP gauge follows this declared score formula. Here, formula fidelity concerns candidate scores under the shared runtime support and mask.

3.2 Why native equivalence need not survive alignment

A native softmax distribution is unchanged when a constant c_j is added to every logit at position j. We call this freedom in the native representation its *gauge*. The interface preserves that equivalence only if the induced candidate-score changes cancel in the virtual softmax. In Figure 1d, shifts $(0, 2)$ change raw scores from $(0, 0)$ to $(0, 1)$ and virtual probabilities from $(0.500, 0.500)$ to

$(0.269, 0.731)$. Native probabilities remain uniform; the log-probability construction also preserves its virtual distribution.

Position-wise log normalization chooses $c_j = -L_j$. Let a' denote the resulting scores. The change in the span-minus-overlap gap is determined by the normalizers removed from the two candidates:

$$(a'_{\text{span}} - a'_{\text{overlap}}) - (a_{\text{span}} - a_{\text{overlap}}) = L_1 - \frac{1}{k}\sum_{j=1}^{k} L_j. \tag{2}$$

Exponentiating Equation 2 divided by τ gives the multiplicative change in span/overlap odds. Virtual normalization cancels in this ratio. At $k = 1$ the gap change is zero; at $k > 1$ its direction and magnitude depend on the realized normalizers. A longer span permits sensitivity but does not guarantee it. The following criterion identifies the general structure.

Proposition 1 (Fixed-support linear-score invariance) *Fix a nonempty finite candidate set, finite logits, logit-independent weights and temperature $\tau > 0$. For*

$$a_i(z) = \sum_{j,v} w_{ijv} z_j(v), \qquad b_{ij} = \sum_{v} w_{ijv},$$

$\mathrm{softmax}(a(z)/\tau)$ *is invariant under all shifts $z_j(v) \mapsto z_j(v) + c_j$ if and only if the vectors b_i are identical across candidates.*

Exact response to position shifts. Write $b_i = (b_{i1}, \ldots, b_{im})$ for the weights of candidate i across m native positions. A shift vector c changes its score by $b_i^\top c$, so the perturbed virtual distribution is

$$\widetilde{q}_i(c) = \frac{q_i \exp(b_i^\top c/\tau)}{\sum_\ell q_\ell \exp(b_\ell^\top c/\tau)}, \qquad \log \frac{\widetilde{q}_i(c)}{\widetilde{q}_\ell(c)} - \log \frac{q_i}{q_\ell} = \frac{(b_i - b_\ell)^\top c}{\tau}. \tag{3}$$

This gives the criterion directly: equal position-weight vectors produce a common offset; unequal vectors admit a shift that changes a candidate ratio. SimCT assigns $(1/k, \ldots, 1/k)$ to spans and $(1, 0, \ldots, 0)$ to overlaps. Actual sensitivity depends on both these weight differences and the current virtual probabilities (Appendix A.1). Invariance alone does not establish formula fidelity: distinct scoring rules can both preserve native equivalence, so the declared and executed constructions must still be compared.

For a particular input, unequal position weights can still yield no change if the realized offsets cancel in every candidate difference. Proposition 1 instead characterizes invariance over all offsets. The aggregation weights identify where a native-equivalent refactor can change the objective. The paired audit then measures the realized effect on frozen texts.

A single constant shared by all native positions would not expose this failure: both SimCT candidate types have total weight one. Every candidate score would receive the same increment, which virtual softmax removes. The relevant test is therefore position-dependent normalization, rather than a global shift of the logits.

3.3 Executable paired interface audit

Given a declared formula, pinned code, model/tokenizer checkpoints and frozen texts, the audit follows four steps:

1. **Specify and reconstruct.** Verify the input representation and independently compute declared scores or test their required properties.
2. **Freeze and intervene.** Fix parameters, text, tokenization, alignment, ordered candidates, teacher mask and loss reduction. Supply raw logits or differentiable log probabilities to the same scoring primitives.
3. **Verify native equivalence.** Compare native probabilities, argmax and target cross-entropy to check the intended probability-preserving intervention.
4. **Measure and report.** Save per-text virtual-distribution differences, losses and gradients over all trainable student parameters before aggregation.

The output pairs each text's configuration with formula/runtime checks, native controls and interface/gradient differences. A normalization refactor should pass the native checks and then be tested at the constructed distribution and complete-gradient levels. Passing the first check alone does not establish equivalence of the executed objective. Appendix Figure S1 summarizes the workflow; the supplementary materials provide the implementation, configurations and result-checking tools.

Separate diagnostics vary context or support rather than the SimCT score representation. Their definitions and results are confined to Appendix C.2.

3.4 Controlled loss, training and evaluation

Both gauges $h \in \{\mathrm{raw}, \mathrm{lp}\}$ use the same reverse-KL loss form for response x. We fix the virtual temperature at 1 and write the retained-segment objective as

$$\mathcal{L}^h(x) = \frac{1}{\max(n_S, n_T)} \sum_r M_r D_{\mathrm{KL}}\big(q^h_{S,r} \,\|\, q^h_{T,r}\big), \qquad M_r = \mathbf{1}\{Z^{\mathrm{raw}}_{T,r} \geq 1/2\}. \tag{4}$$

Here $q^h_{M,r}$ is model M's virtual distribution on segment r, and n_M counts response tokens. The runtime safeguard defines $G_r = 1/Z_{T,r}$, where $Z_{T,r}$ is computed from native teacher probabilities. Its configured threshold $G_r \leq 2$ is exactly $Z_{T,r} \geq 1/2$. We compute this mask from the raw teacher once and reuse it in both arms. Teacher targets and masks are detached. Student normalizers retain their gradients.

This intervention tests score-level fidelity while intentionally holding support and masking fixed. It is not a full re-execution of every score-dependent runtime decision. In this pinned implementation, recomputing Z_T after position-wise log normalization gives the same value in exact arithmetic because its inputs are native probabilities. Numerical differences near the threshold remain a separate check. Appendix B.4 provides the derivation and archived threshold results.

From score derivatives to the complete gradient. At a retained segment, let $J^h_i = \nabla_\theta a^h_{S,i}$ be the student score derivative. For the linear construction in Proposition 1, position-wise log normalization changes this derivative through the native normalizers:

$$J^{\mathrm{lp}}_i = J^{\mathrm{raw}}_i - \sum_j b_{ij} \nabla_\theta L_j, \qquad \nabla_\theta L_j = \sum_v p_j(v) \nabla_\theta z_j(v). \tag{5}$$

For SimCT, the span-minus-overlap derivative consequently changes by $\nabla_\theta L_1 - k^{-1} \sum_j \nabla_\theta L_j$. Detaching the normalizers would remove this term and define a different computation graph even if forward scores agreed.

To see how it enters the loss, define $d^h_i = \log q^h_{S,i} - \log q^h_{T,i}$ and $D^h = \sum_i q^h_{S,i} d^h_i$ for one segment. With a detached teacher and virtual temperature τ, differentiating reverse KL yields

$$\begin{aligned} \frac{\partial D^h}{\partial a^h_{S,i}} &= \frac{q^h_{S,i}}{\tau}(d^h_i - D^h), \\ \nabla_\theta \mathcal{L}^h(x) &= \frac{1}{\max(n_S, n_T)} \sum_r M_r \sum_i \frac{q^h_{S,ri}}{\tau}(d^h_{ri} - D^h_r) J^h_{ri}. \end{aligned} \tag{6}$$

At $\tau = 1$, Equation 6 gives the audited gradient. Teacher probabilities enter d^h, student probabilities weight and center it, and J^h maps it into parameter space. Native equivalence does not ensure agreement in these quantities. Equal position-weight vectors within each model make virtual probabilities identical and cancel the common Jacobian correction because $\sum_i q_{S,i}(d_i - D) = 0$. Unequal vectors need not cancel. Our paired intervention measures the combined teacher/student effect.

Paired readouts. Let g_h concatenate the gradients of all trainable student parameters for one text. We compute one cosine from the full dot product and norms, not an average of layer-wise cosines. The implementation accumulates these sums across parameter tensors, which is algebraically equivalent to flattening and concatenating them. With $\epsilon = 10^{-30}$, the reported changes are

$$C_g = \frac{g^\top_{\mathrm{raw}} g_{\mathrm{lp}}}{\max(\|g_{\mathrm{raw}}\|_2 \|g_{\mathrm{lp}}\|_2, \epsilon)}, \qquad R_g = \frac{\|g_{\mathrm{lp}} - g_{\mathrm{raw}}\|_2}{\max(\|g_{\mathrm{raw}}\|_2, \epsilon)}, \qquad R_{\mathcal{L}} = \frac{|\mathcal{L}^{\mathrm{lp}} - \mathcal{L}^{\mathrm{raw}}|}{\max(|\mathcal{L}^{\mathrm{raw}}|, \epsilon)}. \tag{7}$$

Thus relative L2 change R_g uses the raw-gradient norm, and relative loss change $R_{\mathcal{L}}$ uses the absolute raw loss. The three readouts separate direction, magnitude and scalar objective value. A small loss difference does not bound the gradient difference. We record gradients before clipping and optimizer updates.

For each text, span TV uses the lower median over multi-token segments before filtering. Sorting m segment values increasingly, this is the value at index $\lceil m/2 \rceil$ (one-based): the smaller central value when m is even. The final corpus summary uses the ordinary median across texts, averaging the central pair for an even count. Complete gradients include all retained segments, including one-to-one segments. Training arms then share initialization and optimizer settings but evolve separately. Final students are compared under common raw-logit, LP and native NLL evaluators.

4 EXPERIMENTS

The main experiments follow one SimCT scoring intervention from native controls to fixed-weight gradients, optimization paths and shared-evaluator preferences. Supplementary task and scope diagnostics are reported separately in the appendix.

4.1 EXPERIMENTAL SETUP

Main audit. Falcon3-1B-Base teaches Pythia-160M (Biderman et al., 2023) on 224 GSM8K question/answer texts (Cobbe et al., 2021): the first records from local shards 0 to 2 in numeric order. Texts were not filtered by span exposure.

Teacher selection. With Pythia fixed, tokenizer-only span counts ranked Qwen3 first (4,774 spans) and Falcon3 second (4,084). The initial Qwen3 audit failed its native-CE tolerance check and all-segment TV criterion. Its corrected run remains post-failure evidence. Falcon3 then served as a prospective confirmation of the corrected protocol, frozen before any Falcon3 logits or gradients were inspected. Both tokenizations reconstruct response bytes exactly, use prompt context without added BOS/EOS, and score only the response. The fixed-parameter audit uses float32 and evaluation mode with dropout disabled. This is a single frozen-checkpoint comparison.

Matched training. Matched short training uses one seed and four disjoint held-out blocks of 56 texts. Each fold has 168 training texts; these sets overlap. This diagnostic is post hoc. Each arm reloads the same weights and performs 56 AdamW updates (Loshchilov & Hutter, 2019), three texts per step in frozen order. The optimizer settings and update schedule are shared within each fold, so score representation is the intervention at initialization. All evaluators score the same endpoints and held-out texts. Appendix B.3 gives the loss reductions; the supplementary materials provide the complete training configuration.

Readouts. Per-text span TV is the lower median over multi-token segments before safeguard filtering; full gradients include every retained segment. We report ordinary medians across texts. Figure 2d to f uses within-fold medians and sentence IQRs; Table 2 also reports the median of four fold medians. These are not independent-seed intervals. Additional evaluation protocols and gradient-sample selection are documented in Appendix E.

4.2 RUNTIME FIDELITY AND NATIVE CONTROLS

The pinned primitive exactly matched an independent raw-logit reconstruction of Equation 1; all 13 formula fixtures returned their expected outcomes. Switching representation changed native probabilities by at most 1.19×10^{-7} and target cross-entropy by 1.91×10^{-6}, with unchanged argmax. All trainable student parameters received finite gradients. Both arms used the same 14,620 retained segments out of 14,693; 73 were masked (0.50%). Thus the next comparison holds predictions, alignment and segment participation fixed.

4.3 VIRTUAL DISTRIBUTIONS AND COMPLETE GRADIENTS

Fixed-weight consequence. Despite near-identical native predictions, median complete-gradient cosine was 0.8437 and relative L2 change 0.5577. Table 1 connects virtual distributions, gradients and loss changes within each fold.

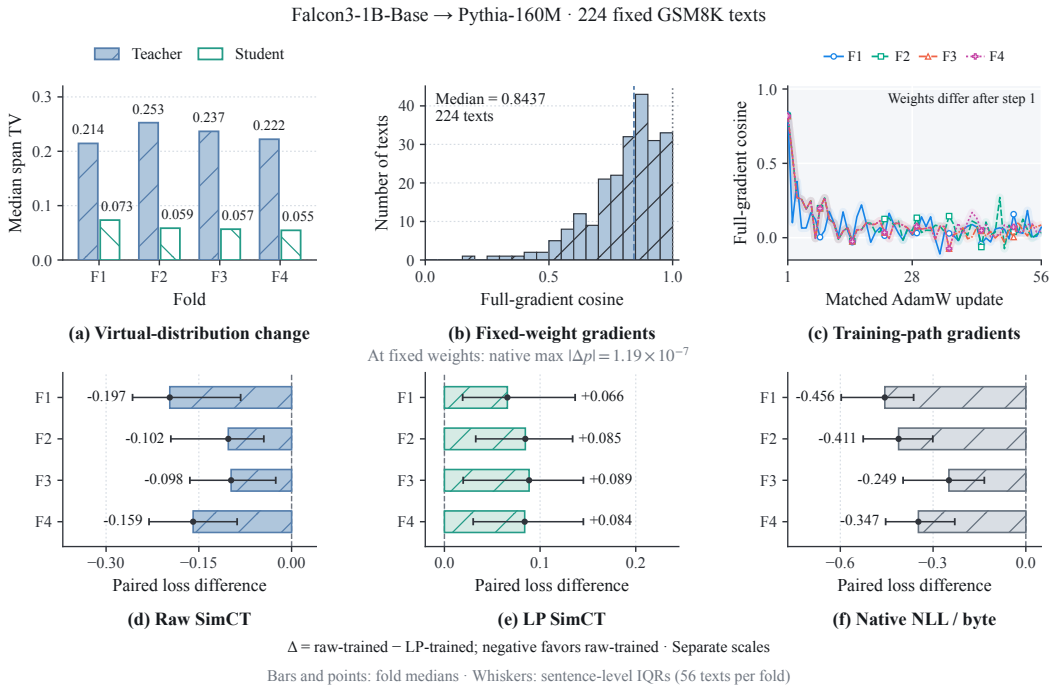


Figure 2: **Paired SimCT audit on 224 fixed GSM8K texts.** **a**, Fold medians of per-text teacher/student span TV (56 texts/fold). **b**, Full-gradient cosine at fixed weights (bin width 0.05; dashed: median 0.8437; dotted: cosine 1). **c**, All 56 gradient updates per fold; hollow markers identify selected steps. Parameters differ after update 1. **d to f**, Paired final-student loss differences under the named evaluators. Bars show within-fold medians; whiskers show sentence-level IQRs, not confidence intervals over runs. Negative/positive differences favor raw-/LP-trained students, respectively. Evaluator scales differ. The folds share one seed, with overlap among their training texts.

Table 1: **Fixed-parameter gauge effects by fold.** Each fold contains 56 texts; entries are medians over texts. Span TV uses each text's median over multi-token spans; complete gradients use the full retained loss. The final row pools all 224 texts.

	Span distribution		Complete gradient		Loss
Fold	Teacher TV	Student TV	Cosine	Rel. L2	Change (%)
F1	0.2144	0.0734	0.8596	0.5218	5.38
F2	0.2526	0.0585	0.8323	0.5724	5.37
F3	0.2366	0.0567	0.8563	0.5357	4.55
F4	0.2221	0.0546	0.8115	0.6021	6.25
All texts	**0.2282**	**0.0585**	**0.8437**	**0.5577**	**5.38**

Consistency across folds. Table 1 resolves the pooled result into four 56-text subsets. Gradient cosine medians range from 0.8115 to 0.8596. Relative L2 changes range from 0.5218 to 0.6021. Every fold exhibits the discrepancy at fixed weights. The largest teacher span TV occurs in F2 (0.2526). The lowest gradient cosine occurs in F4 (0.8115). These summaries capture different aspects of the intervention. The table's final row pools texts rather than averaging fold medians, preserving the aggregation used for the primary result.

Span TV measures movement in probability space. Complete gradients additionally depend on the reverse-KL coefficients and score Jacobians in Equation 6. Contributions from different segments can reinforce or cancel when summed. This structure explains why span TV and gradient cosine need not rank folds identically. Reporting both connects the forward-score discrepancy to its parameter-space consequence without treating one readout as a substitute for the other.

From affected spans to the full loss. Teacher/student span TV medians were 0.2282/0.0585, with a 5.38% median relative loss change. Multi-token spans occur in every text and account for 4,084 of 14,693 aligned segments (27.8%). Invariant one-to-one segments drive all-segment TV medians toward zero, yet the complete retained loss still produces different gradients. Both models' representations change, so the effect includes their combined contribution.

Span exposure. A post-hoc analysis associates the multi-token segment fraction with lower gradient cosine ($\rho = -0.5372$) and higher relative L2 ($\rho = 0.5375$); associations with span TV and loss change are near zero. Appendix C.1 retains sentence-bootstrap intervals. These are within-corpus associations, not interventions on exposure.

4.4 Matched short optimization paths

Across 224 matched updates, median gradient cosine was 0.0660 and relative L2 change 1.2588 (Figure 2c). After update 1, gradients are evaluated at different arm parameters: unlike the fixed-weight 0.8437 comparison, this readout includes path separation. Final inter-arm distance is 1.8457 to 1.8819 in parameter L2, or 1.0285 to 1.0605 times the arms' mean displacement from initialization (Table 2). The scoring intervention produces distinct parameter paths under a common update schedule.

Normalizing endpoint separation by mean displacement puts the difference on the scale of the observed training movement. Values near one mean that the two endpoints are separated by roughly as much as each has moved from initialization, on average. Endpoint distance measures accumulated separation. Gradient cosine compares the local directions encountered during training.

4.5 Evaluator-dependent preference

Raw SimCT favors the raw-trained arm in every fold; LP SimCT favors the LP-trained arm. Median fold differences (raw-trained minus LP-trained) are -0.1308 and $+0.0842$ (Table 2). Evaluator identity changes the preferred endpoint without changing either student.

A reversal of fold medians does not specify how often the evaluators disagree on the same text. Figure 3 pairs the two loss differences underlying Table 2: both coordinates subtract the LP-trained student's score from the raw-trained student's score. Opposite signs indicate disagreement. The expected reversal occurs on 155/224 texts (69.2%); 40 favor raw under both evaluators and 29 favor LP under both, with no reverse disagreement. The scatter exposes the 69 agreements hidden by four reversed fold summaries. Difference magnitudes use each evaluator's own scale.

The disagreement appears in every fold: 42/56, 35/56, 37/56 and 41/56 texts show the raw-negative, LP-positive pattern. Thus the fold-level reversals reflect a majority pattern within each split, rather than a reversal on every held-out text. The shared-evaluator comparison reveals both this recurring direction and its variation across texts. It does not require different evaluators to have comparable absolute loss scales.

Native NLL per response byte favors raw in every fold, with median difference -0.3795. This native-model diagnostic complements the interface scores. Students should therefore be compared under common named scoring rules, with native-model outcomes reported alongside them. Formula fidelity, evaluator preference and model quality require distinct readouts; neither evaluator's lower loss establishes task superiority.

4.6 Interpreting score fidelity and evaluation

The fixed-parameter and trained-student comparisons answer different questions. At fixed parameters, the two computations receive the same text, tokenization, candidates and retained segments. The gradient discrepancy therefore identifies a score-representation effect under those controls. Once optimization begins, each arm encounters its own parameter state. The short-training comparison measures the accumulated separation of these parameter paths.

The shared-evaluator comparison holds the final students fixed and changes the rule used to score them. Each evaluator favors the arm trained with its own representation in all four folds. Such agreement is not an independent measure of model quality.

Table 2: **Final students under shared evaluators.** Within-fold paired medians over 56 held-out texts; Δ is raw-trained minus LP-trained. Final row: median of the four fold medians.

	Path	Paired score difference		
Fold	Sep.	Δ_{raw}	Δ_{lp}	Δ_{NLL}
F1	1.0605	-0.1970	+0.0658	-0.4562
F2	1.0285	-0.1024	+0.0846	-0.4115
F3	1.0323	-0.0978	+0.0885	-0.2487
F4	1.0343	-0.1592	+0.0839	-0.3475
Median	**1.0333**	**-0.1308**	**+0.0842**	**-0.3795**

Sep.: final distance divided by mean displacement from initialization; Δ_{NLL}: native NLL per response byte. Negative differences favor raw-trained students; positive differences favor LP-trained students. Evaluator scales differ. Both panels use the same four folds, one seed and overlapping training sets.

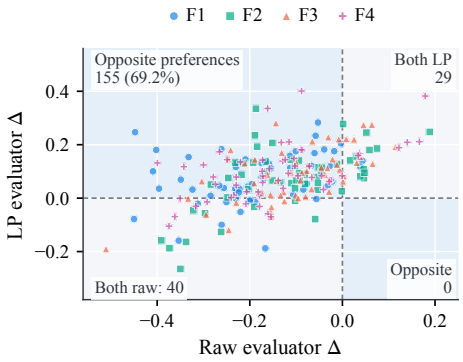


Figure 3: **Evaluator preferences by text.** Text-level view of the paired losses in Table 2. Each point compares the same students on one text. Zero lines separate preference directions; shaded quadrants indicate disagreement.

Native NLL per response byte supplies a separate readout and favors the raw-trained student. The log-probability arm follows the declared score formula under the shared support and mask; this property does not imply superior native prediction.

These distinctions suggest three checks when reproducing or refactoring SimCT. First, compare the executed candidate scores with the declared formula after fixing alignment. Second, test the constructed distributions and complete gradients at identical parameters, even when native prediction checks pass. Third, evaluate both trained students under the same explicitly named rule. Report native predictive or task outcomes separately from interface losses. A discrepancy at one stage should not be inferred solely from a measurement at another.

A discrepancy first detected in the virtual distributions points to the construction layer. A native-prediction discrepancy beyond numerical tolerance signals a failed control and must be resolved before attributing the gradient difference to that layer. This order makes the audit useful for debugging as well as for reporting.

Scope diagnostics. Separate context and support diagnostics vary controls held fixed in the SimCT intervention. They illustrate further uses of the audit but use different constructions, data and metrics (Appendix C.2). A separate three-seed task evaluation found no reliable task-level difference under the frozen grader. Its grading errors, uncertainty and missing controls are documented with the results in Appendix E.

5 Conclusion

Our audit separates declared formulas, runtime scores, training and evaluation. The position-weight criterion explains why native-equivalent scores can yield different virtual distributions and complete gradients. Pinned SimCT exhibits this discrepancy and its optimization consequences. Shared-evaluator comparisons show why score-level formula fidelity and native predictive performance must be measured separately. Reproducing tokenizer alignment is therefore one part of reproducing a distillation objective. The score construction and its gradient path must also be specified, and endpoint quality assessed under a common evaluation rule. These checks make a discrepancy observable before it is interpreted as a training benefit or failure.

Future work can test the persistence of these effects across additional model pairs, independent seeds and longer training. Verified task grading and shared baselines would connect the audit to predictive utility. Explicit mask recomputation and public reference artifacts would strengthen numerical validation and independent reproduction of the complete pipeline.

AI USE STATEMENT

Generative AI tools assisted with writing and presentation, mathematical formulation, methodological feedback, result interpretation, and supporting code. Teacher-generated training data are described in Appendix D.1. The authors take responsibility for the final text, claims and artifacts.

REPRODUCIBILITY STATEMENT

Section 3 specifies the score constructions, invariance criterion and paired audit controls; Section 4 describes the experimental settings and comparisons. Appendix A provides proofs, and Appendix B explains sample selection and metric aggregation. The supplementary materials provide pinned code, complete configurations, archived results, plotting scripts and an implementation guide. Appendix E describes the additional evaluation protocols and grading limits. Checks of archived results are distinct from rerunning the models, which requires the corresponding data, weights and dependencies. Remaining availability constraints, including access to the exact fixed-parameter and short-training input shards, are documented in Appendices B.1, D.1 and E.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-Policy Distillation of Language Models: Learning from Self-Generated Mistakes. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/5be69a584901a26c521c2b51e40a4c20-Abstract-Conference.html.

Anshumann, Mohd Abbas Zaidi, Akhil Kedia, Jinwoo Ahn, Taehwak Kwon, Kangwook Lee, Haejun Lee, and Joohyung Lee. Sparse Logit Sampling: Accelerating Knowledge Distillation in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18085–18108, 2025. doi: 10.18653/v1/2025.acl-long.885. URL https://aclanthology.org/2025.acl-long.885/.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. On the Cross-lingual Transferability of Monolingual Representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4623–4637, 2020. doi: 10.18653/v1/2020.acl-main.421. URL https://aclanthology.org/2020.acl-main.421/.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. The Belebele Benchmark: A Parallel Reading Comprehension Dataset in 122 Language Variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 749–775, 2024. doi: 10.18653/v1/2024.acl-long.44. URL https://aclanthology.org/2024.acl-long.44/.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. Pythia: A Suite for Analyzing Large Language Models Across Training and Scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 2397–2430, 2023. URL https://proceedings.mlr.press/v202/biderman23a.html.

Nicolas Boizard, Kevin El Haddad, Céline Hudelot, and Pierre Colombo. Towards Cross-Tokenizer Distillation: The Universal Logit Distillation Loss for LLMs. *Transactions on Machine Learning Research*, 2025. URL https://openreview.net/forum?id=bwRxXiGO9A.

Alex Boyd, Samuel Showalter, Stephan Mandt, and Padhraic Smyth. Predictive Querying for Autoregressive Neural Sequence Models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 23751–23764, 2022. doi: 10.52202/068431-1725. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/9622163c87b67fd5a4a0ec3247cf356e-Abstract-Conference.html.

Yijie Chen, Yijin Liu, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. Enhancing Cross-Tokenizer Knowledge Distillation with Contextual Dynamical Mapping. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 8005–8018, 2025. doi: 10.18653/v1/2025.findings-acl.419. URL https://aclanthology.org/2025.findings-acl.419/.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. XNLI: Evaluating Cross-lingual Sentence Representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2475–2485, 2018. doi: 10.18653/v1/D18-1269. URL https://aclanthology.org/D18-1269/.

Quoc Phong Dao, Hoang Son Nguyen, Pham Khanh Chi, Tung Nguyen, Linh Ngo Van, Diep Thi-Ngoc Nguyen, and Trung Le. SRA: Span Representation Alignment for Large Language Model Distillation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 32961–32975, 2026. doi: 10.18653/v1/2026.acl-long.1522. URL https://aclanthology.org/2026.acl-long.1522/.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge Distillation of Large Language Models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/8ac015d409635f196f9e3e9dcfb9a94e-Abstract-Conference.html.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring Mathematical Problem Solving With the MATH Dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/be83ab3ecd0db773eb2dc1b0a17836a1-Abstract-round2.html.

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.

Yoon Kim and Alexander M. Rush. Sequence-Level Knowledge Distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1317–1327, 2016. doi: 10.18653/v1/D16-1139. URL https://aclanthology.org/D16-1139/.

Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 66–71, 2018. doi: 10.18653/v1/D18-2012. URL https://aclanthology.org/D18-2012/.

Alexander Lin, Jeremy Wohlwend, Howard Chen, and Tao Lei. Autoregressive Knowledge Distillation through Imitation Learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 6121–6133, 2020. doi: 10.18653/v1/2020.emnlp-main.494. URL https://aclanthology.org/2020.emnlp-main.494/.

Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1711.05101.

Benjamin Minixhofer, Ivan Vulić, and Edoardo Maria Ponti. Universal Cross-Tokenizer Distillation via Approximate Likelihood Matching. In *Advances in Neural Information Processing Systems*, volume 38, pp. 79297–79326, 2025. doi: 10.52202/085713-2653. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/720f9f5dc751eb56952ae4fee2398f73-Abstract-Conference.html.

Thinh On, Hengzhi Pei, Leonard Lausen, and George Karypis. Knowledge Distillation for Large Language Models through Residual Learning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/608fe7e32f7b773545cc1d656a0fdc98-Abstract-Conference.html.

Buu Phan, Brandon Amos, Itai Gat, Marton Havasi, Matthew J. Muckley, and Karen Ullrich. Exact Byte-Level Probabilities from Tokenized Language Models for FIM-Tasks and Model Ensembles. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/5ed5c3c846f684a54975ad7a2525199f-Abstract-Conference.html.

Truong Buu Phan, Ashish Khisti, and Karen Ullrich. Cross-Tokenizer Likelihood Scoring Algorithms for Language Model Distillation. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/40d50fe20d5e6922dbb52d7da6de03f1-Abstract-Conference.html.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural Machine Translation of Rare Words with Subword Units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1715–1725, 2016. doi: 10.18653/v1/P16-1162. URL https://aclanthology.org/P16-1162/.

Avyav Kumar Singh, Yen-Chen Wu, Alexandru Cioba, Alberto Bernacchia, and Davide Buffelli. Cross-Tokenizer LLM Distillation through a Byte-Level Interface. In *Proceedings of the Second Workshop on Customizable NLP: Progress and Challenges in Customizing NLP for a Domain, Application, Group, or Individual (CustomNLP4U)*, pp. 84–96, 2026. doi: 10.18653/v1/2026.customnlp4u-1.9. URL https://aclanthology.org/2026.customnlp4u-1.9/.

Sharath Turuvekere Sreenivas, Adithyakrishna Venkatesh Hanasoge, Mingyu Yang, Ali Taghibakhshi, Saurav Muralidharan, Ashwath Aithal, and Pavlo Molchanov. X-Token: Projection-Guided Cross-Tokenizer Knowledge Distillation. *arXiv preprint arXiv:2605.21699*, 2026. URL https://arxiv.org/abs/2605.21699.

Jie Sun, Mao Zheng, Mingyang Song, Qiyong Zhong, Yilin Cheng, Bichuan Feng, Pengfei Liu, Junfeng Fang, and Xiang Wang. SimCT: Recovering Lost Supervision for Cross-Tokenizer On-Policy Distillation. *arXiv preprint arXiv:2605.07711*, 2026. URL https://arxiv.org/abs/2605.07711.

Tim Vieira, Benjamin LeBrun, Mario Giulianelli, Juan Luis Gastaldi, Brian DuSell, John Terilla, Timothy J. O'Donnell, and Ryan Cotterell. From Language Models over Tokens to Language Models over Characters. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 61391–61412, 2025. URL https://proceedings.mlr.press/v267/vieira25a.html.

Hao Wang, Kun Yuan, Wenlin Zhong, Minglei Zhang, Han Xiao, Ming Sun, and Honggang Qi. Cross-Tokenizer On-Policy Distillation via Byte-Prefix Marginalization. *arXiv preprint arXiv:2607.22334*, 2026a. URL https://arxiv.org/abs/2607.22334.

Huazheng Wang, Yongcheng Jing, Haifeng Sun, Jingyu Wang, Jianxin Liao, Leszek Rutkowski, and Dacheng Tao. Bridging the Tokenizer Gap: Semantics and Distribution-Aware Knowledge Transfer for Unbiased Cross-Tokenizer Distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 33494–33502, 2026b. doi: 10.1609/aaai.v40i39.40637. URL https://ojs.aaai.org/index.php/AAAI/article/view/40637.

Yuqiao Wen, Zichao Li, Wenyu Du, and Lili Mou. f-Divergence Minimization for Sequence-Level Knowledge Distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10817–10834, 2023. doi: 10.18653/v1/2023.acl-long.605. URL https://aclanthology.org/2023.acl-long.605/.

Weixiao Zhan, Yongcheng Jing, Leszek Rutkowski, and Dacheng Tao. Distillation Traps and Guards: A Calibration Knob for LLM Distillability. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 19814–19833, 2026. doi: 10.18653/v1/2026.acl-long.908. URL https://aclanthology.org/2026.acl-long.908/.

Songming Zhang, Xue Zhang, Zengkui Sun, Yufeng Chen, and Jinan Xu. Dual-Space Knowledge Distillation for Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 18164–18181, 2024. doi: 10.18653/v1/2024.emnlp-main.1010. URL https://aclanthology.org/2024.emnlp-main.1010/.

A THEORETICAL ANALYSIS

This appendix gives the fixed-support score-invariance proof used in Section 3.2, followed by the candidate-local support certificate used in Section 4.6. The invariance rule connects virtual distributions to parameter gradients. The support certificate bounds a forward-KL derivative with respect to one candidate logit.

A.1 PROOF OF THE LINEAR-SCORE INVARIANCE RULE

For Proposition 1, the increment under a position-wise shift is $\Delta a_i = \sum_j b_{ij} c_j$. Finite softmax outputs agree exactly when all score differences agree, since $\log(q_i/q_k) = (a_i - a_k)/\tau$. Thus invariance for every c is equivalent to $(b_i - b_k)^\top c = 0$ for every c and every i, k, which holds exactly when $b_i = b_k$. This includes the one-candidate case. In contrast, native log-softmax removes each position's additive shift before any span aggregation.

Odds and local sensitivity. Equation 2 implies the exact virtual-odds identity

$$\frac{q'_{\text{span}}/q'_{\text{overlap}}}{q_{\text{span}}/q_{\text{overlap}}} = \exp\left[\frac{1}{\tau}\left(L_1 - \frac{1}{k}\sum_{j=1}^{k} L_j\right)\right]. \tag{8}$$

For the shifted virtual distribution in Equation 3, the local response is

$$\left.\frac{\partial \log \widetilde{q}_i(c)}{\partial c_j}\right|_{c=0} = \frac{1}{\tau}\left(b_{ij} - \sum_{\ell} q_\ell b_{\ell j}\right). \tag{9}$$

A candidate responds to a position offset when its position weight differs from the virtual-distribution-weighted mean. Thus weight mismatch identifies possible sensitivity; current probabilities determine its local magnitude.

From scores to parameter gradients. Fix the text, alignment and candidate support, and write the student's span-minus-overlap gap as $d^h(\theta)$ in gauge h. Differentiating Equation 2 gives

$$\nabla_\theta(d^{\text{lp}} - d^{\text{raw}}) = \nabla_\theta L_1 - \frac{1}{k}\sum_{j=1}^{k} \nabla_\theta L_j, \qquad \nabla_\theta L_j = \sum_v p_j(v) \nabla_\theta z_j(v).$$

Here $p_j = \text{softmax}(z_j)$ is the native student distribution. Position-wise log normalization can therefore change both virtual distributions and the student's score Jacobian, which enter the reverse-KL parameter gradient. The fixed-parameter gauge audit changes both teacher and student score representations and measures the combined loss-gradient effect; it does not decompose teacher-target, student-distribution and Jacobian contributions.

A.2 CANDIDATE-LOCAL SUPPORT-REMOVAL CERTIFICATE

Proposition A1. For the forward categorical KL at temperature 1, let p_T, p_S be teacher and student probabilities on one retained candidate. Let m_T, m_S be removed masses. The student-logit gradient coordinate is $p_S - p_T$. Assume $0 \le m_T, m_S < 1$, $p_T \le 1 - m_T$, and $p_S \le 1 - m_S$. Define

$$g = p_S - p_T, \qquad \widetilde{g} = \frac{p_S}{1 - m_S} - \frac{p_T}{1 - m_T},$$

$$\delta_S = \frac{p_S m_S}{1 - m_S}, \qquad \delta_T = \frac{p_T m_T}{1 - m_T}, \qquad B = \delta_S + \delta_T.$$

These quantities bound the gradient change:

$$|\widetilde{g} - g| = |\delta_S - \delta_T| \le B.$$

Consequently, $B \le \tau$ certifies $|\widetilde{g} - g| \le \tau$. If $|g| > B$, then $\widetilde{g}$ is nonzero and has the same sign as g.

Proof. Subtracting the full coordinate from the renormalized coordinate gives $\widetilde{g} - g = \delta_S - \delta_T$. Both increments are nonnegative, so the triangle inequality gives the bound. The tolerance statement is immediate. If $g > 0$, then $\widetilde{g} \geq g - |\widetilde{g} - g| \geq g - B > 0$; if $g < 0$, then $\widetilde{g} \leq g + |\widetilde{g} - g| \leq g + B < 0$. This proves the sign statement. The proposition certifies a candidate-logit coordinate.

This forward-KL coordinate bound does not apply to the reverse-KL full parameter gradients in the SimCT experiment. We use B as a simple, conservative audit diagnostic: it ignores cancellation between δ_S and δ_T. When both increments are known, their absolute difference is exact; no computational advantage or tightness claim is made for B.

B AUDIT PROTOCOLS AND REPRODUCIBILITY

The main evidence combines the SimCT invariance criterion with a fixed-parameter gauge audit and matched short training. The Falcon3 gauge audit is prospective. Matched short training is post hoc. The earlier corrected Qwen3 gauge diagnostic remains post-failure evidence. Code, complete configurations and artifact documentation are provided separately in the supplementary materials.

Additional experiments examine frozen-logit interfaces, support certificates, matched-target likelihoods, output-head gradients and short context-training paths. Their no-BOS policy encodes text without prepended special tokens and scores positions after the first native token. Probability outside the tokenizer vocabulary is retained as an explicit residual. These context and support diagnostics complement the SimCT evidence.

B.1 REPRODUCIBILITY SCOPE

The supplementary materials organize the scoring implementation, frozen audit inputs, per-text results and figure-generation tools in a separate implementation guide. The reported tables and figures are recomputed from archived outputs. This verifies their numerical consistency without rerunning training or inference. Model execution additionally requires the corresponding weights, dependencies and input texts. Independent public reacquisition of the exact local GSM8K shards remains unresolved. The scientific comparisons below distinguish these levels of reproduction.

B.2 INTERFACE COMPARABILITY AND REPRODUCTION DEPTH

Table A1 separates formula inspection, primitive execution and task reproduction. Each row identifies the evidence available for one interface family.

B.3 SIMCT OBJECTIVE, COMPUTATION GRAPH AND METRIC AGGREGATION

Selection and evidence status. Three choices are distinct. First, tokenizer-only checks found only four multi-token segments in the initial 224 XNLI inputs and 14 under natural-prefix framing; before model logits were inspected, the protocol switched to GSM8K question/answer responses with response-token loss masks. Second, the text manifest took the first 224 records across the three local shards in numeric order, without filtering individual records by span counts. Third, with the Pythia student fixed, locally complete teacher tokenizers were ranked by multi-token span counts on this frozen manifest: Qwen3 ranked first (4,774 spans), and Falcon3 was the recorded runner-up (4,084 spans). The Qwen3-to-Pythia gauge diagnostic subsequently failed numerical and TV gates. Its corrected run remains post-failure evidence. The corrected Falcon3 protocol was frozen before inspecting its logits or gradients, reusing all 224 texts. The configuration records the corpus-framing and teacher-selection history separately from the fixed-order text rule. The preceding post-failure Qwen3 corrected diagnostic also showed sensitivity (span TV 0.2453/0.0590; gradient cosine 0.8487).

For a response x, let $q^h_{S,r}$ and $q^h_{T,r}$ be the normalized virtual score vectors of aligned segment r in gauge h (raw or LP). With n_S, n_T selected response-token counts, the executed temperature-1 objective is

$$\mathcal{L}^h(x) = \frac{1}{\max(n_S, n_T)} \sum_r M_r \sum_i q^h_{S,r,i} \big(\log q^h_{S,r,i} - \log q^h_{T,r,i}\big), \qquad M_r = \mathbf{1}\{Z^{\mathrm{raw}}_{T,r} \geq 1/2\}.$$

Table A1: Interface comparability and reproduction depth. Formula fixtures classify the stated object. Runtime/task indicates execution of an upstream primitive and reproduction of a published task table, respectively. The SimCT trajectories use a local fixed-text adapter.

Interface	Compared object / type	Formula check	Runtime / task	Supported comparison
ALM	Observed-path likelihood and complement; binary score	Independent fixture	No / no	Formula properties; no full training reproduction
ULD	Sorted probabilities; rank coordinates	Independent fixture	No / no	Token-identity invariance of rank object
SimCT	Overlap/span continuation scores; operational	Paper/code fixtures	Primitive / no	Pinned gauge, gradient and short-path behavior
Exact scoring	Queried byte outcomes; event probabilities	Byte oracles	One path / no	First-/two-byte agreement on one pinned path
BPM	Byte-prefix mass mapping; projection	Style fixture	No / no	Scatter property; no full mapping reproduction
X-Token	Projected teacher mass; target coordinates	Style fixture	No / no	Conditional mass preservation and truncation

This is reverse KL, $D_{\mathrm{KL}}(q_S^h \| q_T^h)$, with equal unit weights for retained segment losses before token-count normalization. The teacher distribution and safeguard mask are detached; the student's $z_j - \log \sum_v \exp z_j(v)$ retains its computation graph, including the log-normalizer derivative. Both arms use the same raw-teacher mask. This isolates candidate scoring under the shared runtime mask. Each training update averages three response losses. The frozen gauge-audit records contain 14,693 aligned segments, including 4,084 multi-token spans (27.8%); all 224 texts contain at least one span. The shared teacher mask removes 73 segments and retains 14,620. Span TV is measured before this filtering. The gradient uses retained segments.

For each aligned segment, TV is half the sum of absolute differences between raw and LP virtual distributions. A multi-token segment has length greater than one under either tokenizer. For each text, the recorded span TV is the lower median over these segments, before safeguard filtering (the behavior of the executed tensor median for an even number of segments). Overall estimates then use the ordinary median across the 224 text records; Figure 2a shows within-fold medians; Appendix Figure S2a uses ordinary medians within each 56-text fold. Full-gradient cosine is computed per text from all trainable student parameters, then summarized by these same text/fold groupings. For raw and LP gradient vectors $g_{\mathrm{raw}}, g_{\mathrm{lp}}$, relative L2 change is $\|g_{\mathrm{lp}} - g_{\mathrm{raw}}\|_2 / \max(\|g_{\mathrm{raw}}\|_2, 10^{-30})$; relative loss change is $|\mathcal{L}^{\mathrm{lp}} - \mathcal{L}^{\mathrm{raw}}| / \max(|\mathcal{L}^{\mathrm{raw}}|, 10^{-30})$. Matched short training records gradients after backward and before clipping and the AdamW step; these are gradient comparisons, not comparisons of the actual optimizer updates.

Native token NLL per response byte is $-\sum_{k=1}^{n_S} \log p_S(x_k \mid \mathrm{prompt}, x_{<k}) / |\mathrm{UTF8}(x)|$, using the selected response targets and the student's native tokenization. It is a token-path likelihood normalized by response length, not a sum over all tokenizations of a byte event. For each final student pair, evaluator differences are paired on the same held-out text. Figure 2d to f and Appendix Figure S2d to f use medians and linearly interpolated 25th/75th percentiles of the 56 differences within each fold; Table 2 uses the median of the four fold medians.

Paired execution controls. Both arms use unit virtual temperature and the same raw-teacher safeguard. Overlap tokens have a shared deterministic coordinate order; this common permutation leaves reverse KL unchanged. Each arm and fold reloads identical student weights. The update schedule and optimizer settings are matched, and dropout is disabled during gradient computation. Complete configurations and the pinned implementation revision are provided in the supplementary materials.

The corrected Qwen3 gauge diagnostic replaced its failed original numerical gate with a dtype-based tolerance and separated multi-token spans from invariant one-to-one segments. It remains post-failure evidence; the fixed-parameter gauge audit froze this corrected protocol before Falcon3 logits or gradients were inspected.

B.4 Safeguard invariance and available threshold summaries

The pinned safeguard uses native probabilities, not the raw virtual-score softmax. Let $\mathcal{O}_T$ be the teacher overlap-token set and let I_r indicate a multi-token segment under either tokenizer. For teacher span tokens $v_1, \ldots, v_k$, the executed quantity is

$$Z_{T,r} = \min\left\{1,\ \sum_{u\in\mathcal{O}_T} p_{T,1}(u) + I_r \exp\left(\frac{1}{k}\sum_{j=1}^{k}\log p_{T,j}(v_j)\right)\right\}, \qquad G_r = Z_{T,r}^{-1}.$$

The mask at threshold $\gamma > 0$ is $M_r(\gamma) = \mathbf{1}\{G_r \le \gamma\} = \mathbf{1}\{Z_{T,r} \ge 1/\gamma\}$. Thus the configuration value 2 and Equation 4's cutoff $1/2$ describe the same rule. At fixed alignment, overlap set and span identities, position-wise shifts preserve every probability in $Z_{T,r}$. Recomputing this safeguard from raw logits or their log-probability representation gives identical masks in exact arithmetic.

The primary audit shares the mask computed from raw teacher logits. It does not separately measure numerical mask disagreements after recomputation. Near-threshold floating-point effects require that explicit check; the native-probability bound alone does not rule them out. The delivered records do not contain a paired shared-mask versus recomputed-mask comparison or a corresponding full-gradient ablation. Neither the exact-arithmetic argument nor the following threshold summaries are presented as that experiment.

Table A2: **Archived safeguard-threshold summaries for Falcon3/Pythia.** Each condition contains 224 valid text records. Cosine and relative L2 are medians; losses are means. These supplementary aggregates have a slightly different base result from the primary audit and do not replace it.

Threshold γ	Gradient cosine	Relative L2	Raw loss	LP loss
1.25	0.843558	0.557189	1.392981	1.322109
2 (base)	0.843989	0.557189	1.399126	1.328607
4	0.843989	0.557189	1.403588	1.332696
Disabled	0.844593	0.559313	1.407511	1.336653

The supplementary results package contains threshold settings 1.25, 2, 4 and disabled at temperature 1, float32 precision and full overlap support. Their median gradient cosines range from 0.843558 to 0.844593 (Table A2). This narrow range is descriptive evidence across the delivered threshold settings. The base cosine 0.843989 differs slightly from the primary audit's 0.843676. The package supplies condition definitions and aggregates, but not the executed runner or per-text masks for these variants. Consequently, it cannot establish whether masks were recomputed separately in each gauge. The supplementary implementation guide locates the archived condition definitions and result summaries.

B.5 Multilingual implementation and aggregation details

Chunks lie between consecutive shared byte boundaries. Vocabulary maps exclude empty, invalid-UTF-8 and non-unique byte representations. The no-BOS audits omit chunks starting at token index zero on either side; the remaining eligibility and exposure rules are in Appendix C.2. The archived mean-log-probability alternative is not the source of the reported primary competition masses.

The model-vocabulary residual is $\rho_{M,j} = \sum_{v\ge|\mathcal{V}_M|} P_M(v \mid h_{M,j})$, where $|\mathcal{V}_M|$ is the tokenizer entry count. It is measured separately, not added to Z_M. The no-BOS contract retains these IDs in the native softmax. The original breadth audit instead masks them before softmax and prepends the configuration-level BOS when available. The log-gap implementation caps each removed mass at $1-10^{-15}$ before evaluating $-\log(1-m_M)$. Tolerance certification uses ≤ 0.005; sign certification uses a strict inequality. These labels concern the forward-KL candidate coordinate in Appendix A.2.

The joint-policy diagnostic matches records by pair, sample ID, parallel-group ID, language and chunk index, then retains exposure in both contracts: 29,943 records, versus 43,500 exposed no-BOS repair audit records. The coverage denominators are 15,662 for the Falcon3 support audit and 13,052 for the TinyLlama support audit. Texts without exposed spans contribute no rows to these denominators. Each exposed span has equal weight. We do not rebalance contributions across sentences, languages or pairs.

The model audits select the first 32 distinct group IDs in the frozen 128-group NFC manifest, retaining seven languages per group. The existing 5,000 bootstrap replicates sample 32 group IDs with replacement, retain their language/pair/span records and recompute pooled statistics; bounds are percentile intervals. Pair intervals restrict the same draws to that pair. Figure S3a,b uses seven shared partners; c includes all seven Falcon3 and eight TinyLlama pairs.

B.6 Limitations

The prospective complete-gradient result concerns one model pair; short training uses one seed and overlapping folds. The supplementary task result covers three paired seeds but retains frozen-grader errors and a question-conditional interval. Its delivered outputs do not independently reproduce the training environment or establish continuity with every historical continuation artifact. Context/support diagnostics have separate constructions and evidence states. Reacquiring the original GSM8K shards from a public source and reproducing the full upstream training stack remain open tasks.

C Additional Experiment Results

Figures S1 and S2 give the paired workflow and fold-level results. The following subsections report span exposure and context/support diagnostics; Appendix D records failures and the status of the historical continuation status.

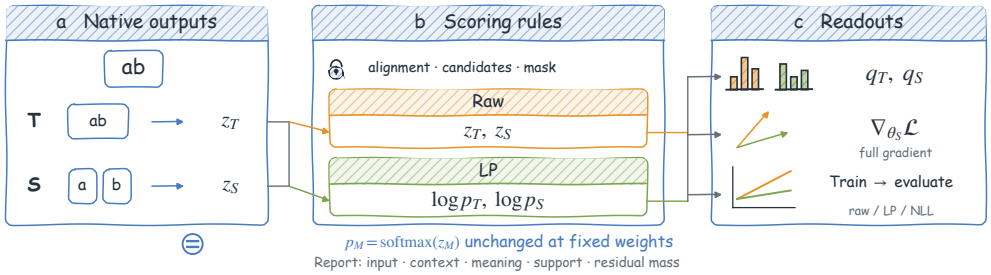


Figure S1: Paired scoring audit of the pinned SimCT interface. Both models' outputs enter raw-logit and LP log-probability rules with shared alignment, candidates and teacher mask. Native probabilities are preserved at fixed weights; virtual distributions and student gradients are compared before short training and named evaluation. Token blocks are schematic. The lower strip lists reporting obligations; Figure 2 reports the measurements.

C.1 Post-hoc association between span exposure and gradient change

A post-hoc sentence-level analysis defined span exposure as the fraction of aligned segments that were multi-token spans. Exposure was associated with lower complete-gradient cosine (Spearman $\rho = -0.5372$, sentence-bootstrap 95% interval $[-0.6338, -0.4272]$) and higher relative L2 change ($\rho = 0.5375$, $[0.4288, 0.6336]$). Its associations with teacher span TV, student span TV and relative loss change were close to zero (Table A3). These 224-sentence associations describe the audited pair rather than an intervention on span exposure. Each interval uses 10,000 sentence-bootstrap replicates, conditional on the audited model pair, checkpoints and texts. The two gradient associations describe a common pattern, not independent replications.

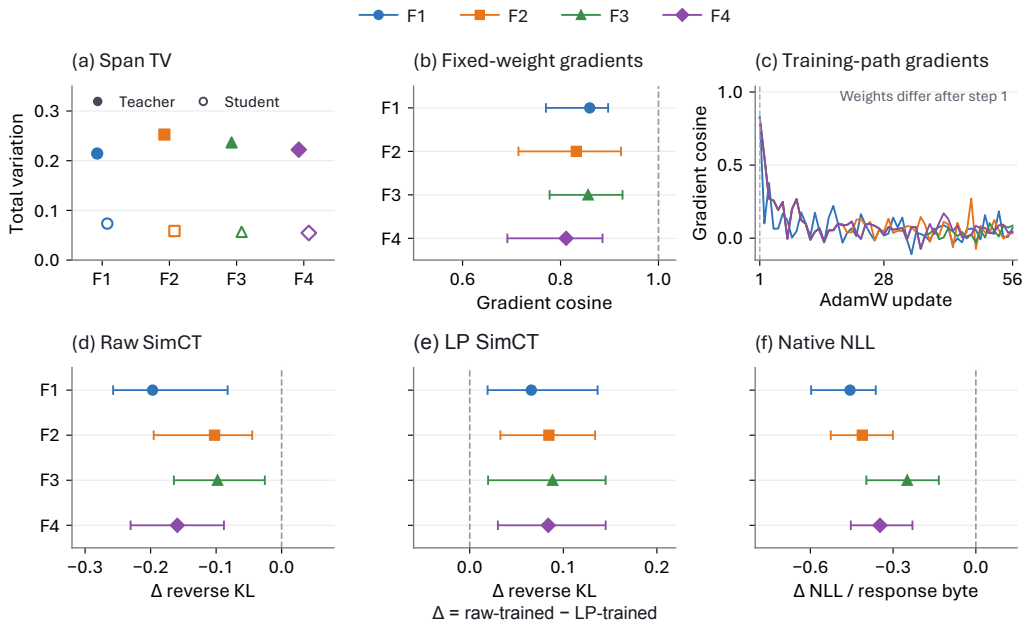


Figure S2: Paired SimCT audit of Falcon3-1B-Base and Pythia-160M on 224 fixed GSM8K texts. Colors and marker shapes identify folds. **a**, Fold medians of span-only TV (filled: teacher; hollow: student); native probability change is at most 1.2×10^{-7}. **b**, Full-gradient cosine at identical weights (prospective comparison; overall median 0.8437). **c**, Gradient cosine at 56 matched updates per fold (post hoc), not a parameter-distance curve; after update 1 the weights differ. **d to f**, Paired held-out differences for the same final students under raw SimCT, LP SimCT and native token NLL per response byte, with separate scales; negative differences favor raw-trained students and positive differences favor LP-trained students. Points in **b,d to f** are within-fold medians; bars are sentence-level interquartile ranges, not confidence intervals over training runs.

Table A3: Post-hoc association between fixed-parameter gauge audit span exposure (span segments divided by all aligned segments) and sentence-level outcomes. All 224 sentences enter each Spearman correlation; intervals use 10,000 sentence-bootstrap replicates.

Outcome	ρ	95% bootstrap interval
Teacher span TV	-0.0348	$[-0.1715,\ 0.1043]$
Student span TV	$+0.0636$	$[-0.0699,\ 0.1968]$
Relative loss change	$+0.0623$	$[-0.0689,\ 0.1928]$
Complete-gradient cosine	-0.5372	$[-0.6338,\ -0.4272]$
Complete-gradient relative L2	$+0.5375$	$[0.4288,\ 0.6336]$

C.2 Context and support: definitions and summary evidence

This subsection collects the definitions, evidence ledger and coverage plot for Section 4.6. The following subsections give the experiment narratives: input reconstruction in Appendix C.3, joint context/support decisions in Appendix C.5, prospective Falcon3 results in Appendices C.7 to C.8, and sequential TinyLlama coverage failures in Appendix C.9. These path-score diagnostics sum log probabilities and concern forward-KL candidate coordinates; the SimCT gauge experiment uses a separate construction and objective.

C.2.1 Definitions and aggregation

Multilingual audit metrics. Let C contain nonempty valid-UTF-8 strings with exactly one token ID in each tokenizer. An eligible valid-UTF-8 byte-aligned span $x \notin C$ uses multiple tokens on at least one side and starts at a scored position. It is *prefix-exposed* when $R_x = \{c \in C : c \prec x \text{ or } x \prec c\} \neq \emptyset$, where $\prec$ means

proper byte prefix. For side $M \in \{T, S\}$ (recorded left/right), $u_M(c)$ is the next-token probability at that side's span start under the recorded context/support policy and $s_M(x)$ is the product of the realized conditional token probabilities over the span. The candidate distribution on $C \cup \{x\}$ has

$$Z_M = s_M(x) + \sum_{c \in C} u_M(c), \quad p_M = \frac{s_M(x)}{Z_M}, \quad m_M = \frac{\sum_{c \in R_x} u_M(c)}{Z_M}.$$

Competition mass is $m = \max(m_T, m_S)$. Removing R_x and renormalizing yields the absolute span log-gap change $G = |\log(1 - m_T) - \log(1 - m_S)|$. This candidate mass is distinct from native model-vocabulary residual on IDs without tokenizer entries. Define the forward-KL coordinate $g = p_S - p_T$ and $B = \sum_{M \in \{T,S\}} p_M m_M/(1 - m_M)$. The materiality, tolerance-certification and sign-certification labels are

$$y_{\text{mat}} = \mathbf{1}\{m \geq .01\}, \qquad y_{\text{tol}} = \mathbf{1}\{B \leq .005\}, \qquad y_{\text{sign}} = \mathbf{1}\{|g| > B\}.$$

Coverage is the fraction of *all exposed spans* satisfying the corresponding certificate label, without materiality filtering. Global medians and fractions pool span records; pair/language summaries restrict that pool rather than averaging sentence means. Intervals resample whole parallel sentence groups, keeping their languages and pairs together. Label-change rates use paired disagreements over common exposed records (Appendix B.5).

Changes in these labels count changes in certification or materiality status between audit policies. In particular, a change in y_{sign} does not measure whether the observed gradient coordinate changed sign. An uncertified label ($y_{\text{sign}} = 0$) leaves the sign undetermined. It does not establish a gradient sign flip.

Table A4: Evidence states and principal no-BOS outcomes. Intervals are 95% sentence-cluster intervals over 32 parallel groups. TinyLlama was selected after the Falcon3 audit and is sequential rather than an independent holdout.

Protocol	State	Exposed	Materiality: mass; fraction $\geq .01$	Certificate / scope
No-BOS repair	Post hoc	43,500	.0717; 87.62%	Repair scope; not prospective
Falcon3	Pass	15,662	.1089 [.1014,.1161]; 92.99% [.9219,.9384]	55.25% tolerance; 88.22% sign; zero violations
TinyLlama	Pass / fail	13,052	.1127 [.0977,.1280]; 88.21% [.8649,.8964]	50.09% tolerance; 84.67% sign; sign-coverage gate failed

C.2.2 Evidence ledger and coverage summary

Table A5 retains the original evidence states, including failed coverage requirements. Figure S3 shows the corresponding checkpoint and partner summaries. The intervals resample sentence groups; the four-fold gradient and training summaries use one seed and overlapping training sets.

C.3 Input-domain and tokenizer scope are empirical obligations

Structural discovery uses XQuAD and Belebele (Artetxe et al., 2020; Bandarkar et al., 2024). Confirmation uses 128 parallel XNLI groups in seven languages (896 texts); model audits take the first 32 groups from that frozen NFC manifest (224 texts) (Conneau et al., 2018).

The raw-source confirmation passed all registered structural statistics but failed byte exactness because Hindi strings changed under Unicode normalization. The failure was preserved. A separately frozen NFC-domain confirmation then reconstructed all 896 evaluated rows exactly and reproduced the tokenizer-pair signature across the held-out XNLI corpus. Thus, the choice between raw and NFC input was not metadata: it changed whether the stated interface existed.

The seven-tokenizer scope audit evaluated all 21 nominal pairs over 4,704 pair-sample rows. Exactly one pair, Qwen2.5/Qwen3, had an identical tokenizer interface on every frozen sample. We treated it as a same-interface negative control rather than a CTKD effect pair. The corrected scope therefore contained 20 tokenizer-distinct pairs, not 21.

Table A5: Evidence summary. P: prospective; S: sequential after the Falcon3 audit; D: post-hoc or post-failure diagnostic. Intervals resample 32 sentence groups (5,000 replicates). Training and fixed-parameter audits report descriptive fold summaries. Times are recorded runner seconds on RTX 5070 Ti Laptop (L) or RTX 6000D (R). NR denotes analyses that reuse logits and have no separate recorded runtime.

Evidence	Scope / samples	Metric and estimate	95% CI	Time	Evidence target
Falcon3 mass; P, pass	Falcon3; 7 pairs; 15,662 spans	Competition mass .1089	[.1014,.1161]	201 L	Materiality
Falcon3 certificate; P, pass	Same Falcon3 spans	Tolerance/sign coverage 55.25/88.22%	[53.42,57.27]%; [87.01,89.22]%	NR	Local certificate
TinyLlama mass; S, pass	TinyLlama; 8 pairs; 13,052 spans	Competition mass .1127	[.0977,.1280]	195 L	Sequential replication
TinyLlama certificate; S, fail	Same TinyLlama spans	Sign coverage 84.67% $< 85\%$; 5/8 passing pairs $< 6/8$	[83.54,85.90]%	NR	Coverage failure; soundness retained
Context; D	7 models; 1,568 sentence-model rows	Absolute NLL change .4153 nats	[.3703,.4537]	89 L	Matched targets
Context training; D	Qwen2.5 to Pythia; 4×96 steps	Gradient cosine .1724; BCE direction 3/4 folds	Not estimated	411 L	Short CTKD path
Fixed gauge; P, pass	Falcon3 to Pythia; 224 texts	Gradient cosine .8437; loss change 5.38%	Not estimated	50 R	Fixed parameters
Short training; D	Same pair; 4×56 steps	Separation 1.0333; gauge-dependent preference 4/4 folds	Not estimated	112 R	Path / evaluator

C.4 The original current-model protocol failed despite favorable effects

The original breadth run found material effects on 30,433 exposed spans: all 14 observed new effect pairs and seven languages passed their directional checks. The original breadth audit nevertheless remained FAIL. It preregistered 15 new pairs and 21 combined pairs, and Qwen3's padded-output mass reached 0.0553, above the 0.001 ceiling.

The scope audit did not prepend BOS. The original model runner instead prepended the configured BOS token whenever one was defined. Six models used a BOS-prepended policy; OLMo2 did not. We retain the original breadth audit and the downstream BOS-prepended results under their original statuses and report the input policy explicitly.

Held-out approximation studies reinforce this separation. The two-byte diagnostic passed all 15 pair-level direction checks but failed its padded-output gate and remained INVALID. The one-step scatter diagnostic was invalid and directionally negative: held-out spanning mass fell to 0.0298, and only 9/15 pair cosine gates passed. We therefore reject a robust one-step-scatter claim.

C.5 Explicit context and residual accounting recover the seven-model scope

The no-BOS repair audit specified no-BOS inputs and explicit residual accounting after the original breadth audit failure. It attempted all 21 nominal model pairs, identified the Qwen2.5/Qwen3 negative control, and evaluated 20 effect pairs on the same 224 samples. The negative control produced zero effect rows as required. Across the effect pairs, the no-BOS repair audit retained 65,298 divergent spans; 43,500 were prefix-exposed. The pooled median maximum competition mass was 0.0717, the median absolute log-gap change was 0.0553, and 87.6% of exposed spans had mass

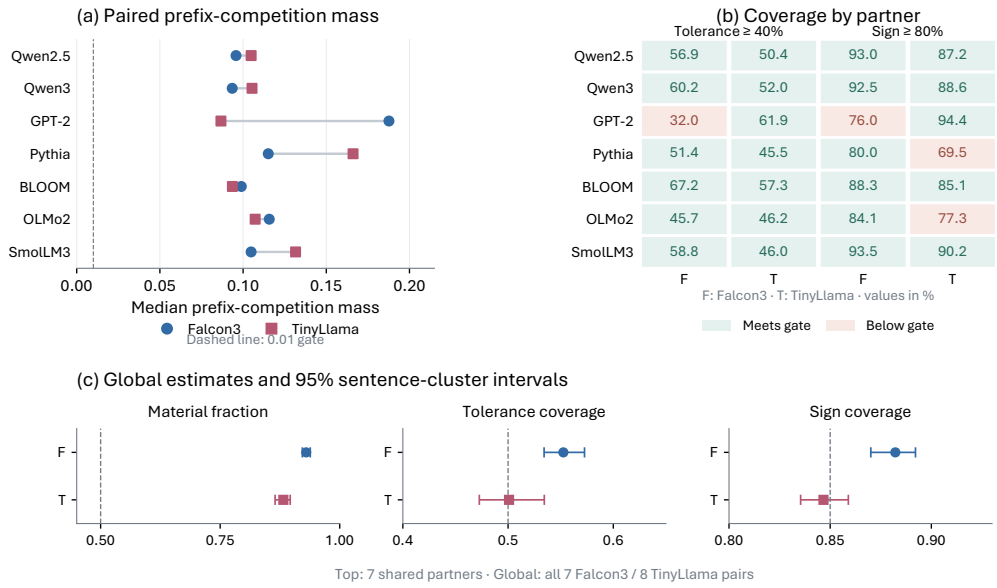


Figure S3: Prefix competition persists while certificate coverage remains checkpoint-dependent. **a**, Median competition mass for seven shared partners; circles denote Falcon3, squares TinyLlama, and gray lines join paired estimates. The dashed gate is 0.01. **b**, Coverage percentages; F/T denote Falcon3/TinyLlama. Cell colors indicate whether unrounded estimates meet pair-level gates (40% tolerance, 80% sign). **c**, Global estimates over all registered pairs (Falcon3: seven; TinyLlama: eight), with original 95% sentence-cluster intervals (32 groups, 5,000 replicates). Axes have separate ranges; gates are 0.5, 0.5 and 0.85. TinyLlama sign coverage is 84.67%, below its point-estimate gate, but its interval crosses the gate. Intervals quantify sentence-group uncertainty. The TinyLlama study is sequential.

at least 0.01. All 20 effect pairs passed both registered pair-level materiality checks, and all seven languages passed their mass criterion.

The explicit residual exposed a model-specific issue without invalidating the registered no-BOS repair audit ceiling. Qwen3 assigned at most 0.0122 probability to model output identifiers outside its tokenizer; every other model's maximum was at most 6.56×10^{-5}. The no-BOS repair audit's frozen ceiling was 0.10, and peak allocated GPU memory was 9.53 GB on a 12 GB RTX 5070 Ti Laptop GPU. All 16 checks passed. Because the no-BOS repair audit was defined after observing the original breadth audit's failure, its recorded status is POST HOC PASS. It provides post-hoc support within the repaired protocol.

We next tested whether the policy change affected analysis decisions. A matched post hoc diagnostic covered 29,943 exposed spans across 14 overlapping pairs. Pair rankings remained high, with Spearman 0.947 for median mass and 0.969 for median log-gap. The joint contract nevertheless altered 5.29% of materiality labels, 5.17% of tolerance-certification labels, and 3.53% of sign-certification labels. Qwen3-containing pairs had higher change rates than other pairs. This diagnostic cannot isolate BOS prepending from residual masking, and it is not prospective. It establishes a bounded analysis-decision consequence despite stable aggregate rankings.

C.6 TOKENIZER PREFLIGHT PREVENTS REDUNDANT MODEL EVIDENCE

We screened candidate checkpoints using tokenizer behavior before downloading their full weights. StableLM-2 and OLMo2 used the same tokenizer interface on all 224 frozen samples, so the registered preflight failed and the StableLM-2 weight was not downloaded. Falcon3 used a 131,072-entry tokenizer, inserted no BOS, reconstructed every input exactly and differed from each existing tokenizer. It therefore entered the prospective model stage. This sequence separates model-family breadth from event-interface breadth and avoids counting two checkpoints as independent CTKD evidence when their tokenizer events are identical.

C.7 PROSPECTIVE FALCON3 CONFIRMATION

The Falcon3 materiality protocol was frozen while Falcon3's weight was unavailable locally and before any candidate logits. The no-BOS smoke used 3.36 GB peak allocated GPU memory. It scored ten positions without inserting the configuration-level BOS and observed zero model-vocabulary residual. We then evaluated Falcon3 against the seven predefined tokenizer partners.

The Falcon3 materiality audit passed all 16 registered checks. Across 224 samples, it retained 25,819 divergent spans, of which 15,662 were prefix-exposed. The pooled median maximum competition mass was 0.1089, median absolute log-gap change was 0.0859, and 92.99% of exposed spans carried at least 0.01 competition mass. All seven pairs passed both pair-level gates and all seven languages passed. Pair medians ranged from 0.0935 (Qwen3/Falcon3) to 0.1877 (GPT-2/Falcon3). Even the lowest median exceeded nine times the frozen 0.01 threshold. Falcon3's residual was zero, while Qwen3 remained the global maximum at 0.0122, below the frozen 0.10 ceiling. Peak allocated GPU memory was 10.11 GB and wall-clock time was 201 s.

Sentence-cluster resampling preserved this result. We ran 5,000 prospectively specified bootstrap replicates over 32 parallel groups. The global median mass had a 95% interval [0.1014, 0.1161]. The material-mass fraction interval was [0.9219, 0.9384]. These are sentence-group intervals, not model-family population intervals.

C.8 CANDIDATE-LOCAL SUPPORT CERTIFICATES AND CLUSTERED UNCERTAINTY

The earlier BOS-prepended discovery protocol tested three unseen BLOOM pairs. It certified 72.54% of exposed coordinates below tolerance 0.005 and 94.20% as sign-stable. No bound, tolerance, or sign certificate was violated. A 5,000-replicate parallel-group bootstrap gave 95% intervals [70.39%, 74.75%] and [93.04%, 95.22%], respectively. These results validate the inequality and its coverage under that intervention, but their model state is explicitly BOS-prepended-conditioned.

The Falcon3 materiality audit passed. Its frozen progression rule then allowed the support certificate audit. Across 15,662 exposed coordinates, 55.25% were certified below distortion tolerance 0.005 and 88.22% were certified candidate-logit sign stable. The corresponding 95% sentence-cluster intervals were [53.42%, 57.27%] and [87.01%, 89.22%]. Six of seven pairs passed each registered pair-coverage gate. GPT-2/Falcon3 was the sole pair below both thresholds, with 31.99% tolerance coverage and 76.02% sign coverage, and remains an explicit failure slice.

The certificate was informative as well as sound: median exact distortion was 0.002012, median upper bound was 0.003322, and median exact-to-bound ratio was 0.8564. No analytic bound was violated, no tolerance certificate was false, and no sign-certified coordinate flipped. These results prospectively quantify candidate-logit certificate coverage under the no-BOS policy.

C.9 SEQUENTIAL TINYLLAMA MATERIALITY AND CERTIFICATE COVERAGE

TinyLlama was selected only after the Falcon3 materiality audit and is therefore a sequential, not independent, replication. Two failed SentencePiece infrastructure attempts remain preserved. A separately frozen hybrid-byte preflight then reconstructed all 1,792 pair-sample rows before weight download or logits.

The TinyLlama materiality audit passed every breadth gate on the eight predefined TinyLlama pairs, retaining 13,052 prefix-exposed spans. Median maximum competition mass was 0.1127, median absolute log-gap change was 0.0949, and 88.21% of exposed spans carried at least 0.01 mass. All 8/8 pair mass and 8/8 pair log-gap gates passed, as did six of seven languages. The 5,000-replicate sentence-cluster interval for median mass was [0.0977, 0.1280]. The interval for material frequency was [0.8649, 0.8964].

The certificate did not fully replicate (Figure S3b,c). Tolerance coverage was 50.09%, but sign coverage was 84.67%, below the frozen 85% gate; only 5/8 pairs passed the sign gate rather than the required 6/8. The clustered intervals were [47.27%, 53.44%] and [83.54%, 85.90%]. The TinyLlama certificate audit remains FAIL for coverage, although all soundness checks had zero violations.

C.10 MATCHED-TARGET SCORING EXPOSES A BOUNDED MODEL CONSEQUENCE

The matched-target context diagnostic held target identity fixed while crossing no-BOS or BOS-prepended context with complete or tokenizer-masked output support. Every condition scored native targets at positions 1 onward. The initial run remains invalid after loading a Pythia placeholder tokenizer. A separately frozen corrected run changed only the tokenizer location and retained all scientific outcomes, thresholds and claim boundaries.

The corrected context diagnostic completed on seven BOS-defined checkpoints, 1,568 sentence-model rows and 72,863 matched target tokens. BOS prepending changed sentence mean NLL by a median absolute 0.4153 nats under complete output support. The 5,000-replicate sentence-cluster interval was [0.3703, 0.4537]. All 1,568 rows exceeded the preregistered descriptive level of 0.01 nats. The effect was not concentrated in one checkpoint or language. Model medians ranged from 0.1760 to 1.4736; language medians ranged from 0.2878 to 0.5233 (Figures S4b and S6a).

The matched cross-entropy logit-gradient vectors also changed. Their pooled median cosine between no-BOS and BOS-prepended contexts was 0.99834, with a 95% sentence-cluster interval of [0.99790, 0.99866]. Model medians ranged from 0.97485 for Pythia to 0.99958 for GPT-2. In contrast, tokenizer-vocabulary masking had a near-zero median NLL effect at these observed native targets. The largest absolute model median was 6.86×10^{-5} nats for Qwen3. Signed context effects differed by checkpoint. The diagnostic establishes a material change rather than uniform degradation. It remains a post-failure diagnostic of model likelihood and output-space gradients, not a parameter-gradient, training-trajectory or downstream-task experiment.

C.11 OUTPUT-HEAD GRADIENTS PRESERVE THE CONTEXT CONSEQUENCE

Table A6: **Context effects with matched target tokens on seven checkpoints (post-failure diagnostics).** Each model uses the same 224 XNLI texts. NLL change is the median sentence-mean absolute token-NLL difference in nats. Logit cosine takes a within-sentence token median, then a median over sentences; head metrics take sentence medians.

	Likelihood	Gradient cosine		Gradient change
Checkpoint	$\lvert\Delta\rvert$ NLL	Logits	Head	Head rel. L2
Qwen2.5†	0.2892	0.9988	0.9822	0.1888
Qwen3†	0.4750	0.9968	0.9604	0.2852
GPT-2†	0.1760	0.9996	0.9850	0.1769
Pythia	1.4736	0.9748	0.8827	0.5399
BLOOM†	0.4002	0.9981	0.9872	0.1614
SmolLM3†	1.0765	0.9985	0.9306	0.3777
TinyLlama	0.2037	0.9993	0.9682	0.2544

Contexts differ by prepending the model-defined BOS token. Head metrics compare observed-target output-row gradient slices. †: output-head contribution to a tied embedding parameter, not its complete gradient. These are context interventions, separate from the SimCT gauge comparison.

The output-head diagnostic propagated the matched-target intervention to every output-head row observed as a target in each sentence. The initial run remains FAIL: repeated half-precision head projection produced a GPT-2 reconstruction error of 0.0625, above its frozen 0.001 ceiling. A separately frozen corrected run instead captured the exact LM-head input with a forward pre-hook and changed no scientific outcome, threshold or claim boundary.

The corrected run covered 42,921 target rows across the same 1,568 sentence-model units (Figures S4c and S6c,d). The median output-head gradient cosine was 0.96717 (95% sentence-cluster interval, [0.96205, 0.97155]). Median relative L2 change was 0.25983 ([0.23974, 0.27790]). Cosine was at most 0.99 for 84.95% of rows, while 95.15% had relative change at least 0.10. Model medians ranged from 0.88274 to 0.98717. For tied embeddings, this is the output-head path contribution to a shared parameter, not its complete gradient. The output-head diagnostic establishes a parameter-path consequence, not an optimizer trajectory, trained-model effect or downstream result.

C.12 MATCHED CE AND CTKD TRAJECTORIES DIVERGE

The CE context-training diagnostic reloaded tied GPT-2 and untied Pythia-160M from identical weights for four frozen 16-step full-parameter SGD rotations. Across four 16-step rotations per model (128 matched steps in total), the medians of within-rotation gradient-cosine medians were 0.96893 and 0.31938; the corresponding relative L2 medians were 0.24988 and 1.24052. Final separation was 0.46698 and 1.52355 times mean net movement. All four split-mean no-BOS held-out NLL differences were positive for each model. Thus the context effect reaches full gradients in tied and untied models and short CE trajectories (Figure S4d).

The context-training diagnostic tested an actual teacher-student objective. The initial run failed before model loading and produced no scientific output. The corrected run froze only that correction. It matched Qwen2.5-0.5B teacher and Pythia-160M student likelihoods over 7,028 byte-identical aligned chunks, using four disjoint evaluation folds and 96 AdamW steps per fold. Initialization, text, chunk weights, order and optimizer were matched; only both models' context policy changed. Across 384 matched steps, median full-gradient cosine was 0.17240, relative L2 change was 1.18713, and final path separation was 0.96637 times mean net movement. At step 96, the median BOS-prepended minus no-BOS held-out CTKD BCE was 0.000996. Fold differences ranged from -0.000458 to 0.003882; 3/4 were positive (Figure S4e). The context-training diagnostic establishes CTKD trajectory sensitivity, not uniform harm. It remains one operational aligned-path objective, pair, optimizer and short horizon, with no downstream task.

Table A7: **Matched context-training diagnostic on Qwen2.5-0.5B and Pythia-160M (post-failure diagnostic).** Each fold uses 24 training and eight held-out parallel groups, 96 updates, and 56 held-out sentences. Both students are evaluated without an inserted BOS token.

	Optimization path		Common no-BOS evaluator	
Fold	Path cosine	Separation	Δ CTKD BCE	Δ NLL/byte
F1	0.1288	1.0158	-0.000458	-0.0098
F2	0.1072	0.9707	+0.000552	+0.0194
F3	0.1755	0.9621	+0.001441	+0.0660
F4	0.2313	0.9442	+0.003882	+0.0335

Differences are BOS-trained minus no-BOS-trained; positive values favor no-BOS training. Path cosine is the median over 96 updates at the arms' evolving parameters. Separation uses the same displacement-normalized definition as Table 2. This post-hoc diagnostic uses one seed and overlapping training groups; its held-out scores are byte-weighted, not sentence medians.

D PRESERVED FAILURES AND INCOMPLETE VALIDATION

The archive retains the original breadth failure and invalid two-byte and scatter confirmations, the StableLM-2 same-interface stop, TinyLlama's failed global-offset preflight and the TinyLlama sign-coverage failure. The initial context, output-head and context-training runs retain their tokenizer-path, reconstruction and import-path failures, respectively. The initial Qwen3 gauge run retains its numerical/TV failures. Corrections are recorded separately. These records are not pooled with the prospective fixed-parameter gauge audit; matched short training remains post hoc. Formula fixtures and runtime primitives remain distinct from complete training/task reproductions.

D.1 HISTORICAL CONTINUATION AND ARTIFACT STATUS

The original Qwen2.5-to-Phi continuation record covers one seed pair and 156 updates per arm, but no task evaluation. The available records do not establish an independently reproduced training run. The later three-seed evaluation is assessed separately in Appendix E. Its results do not retroactively validate the earlier run. Detailed provenance and missing-material records are retained in the supplementary implementation guide.

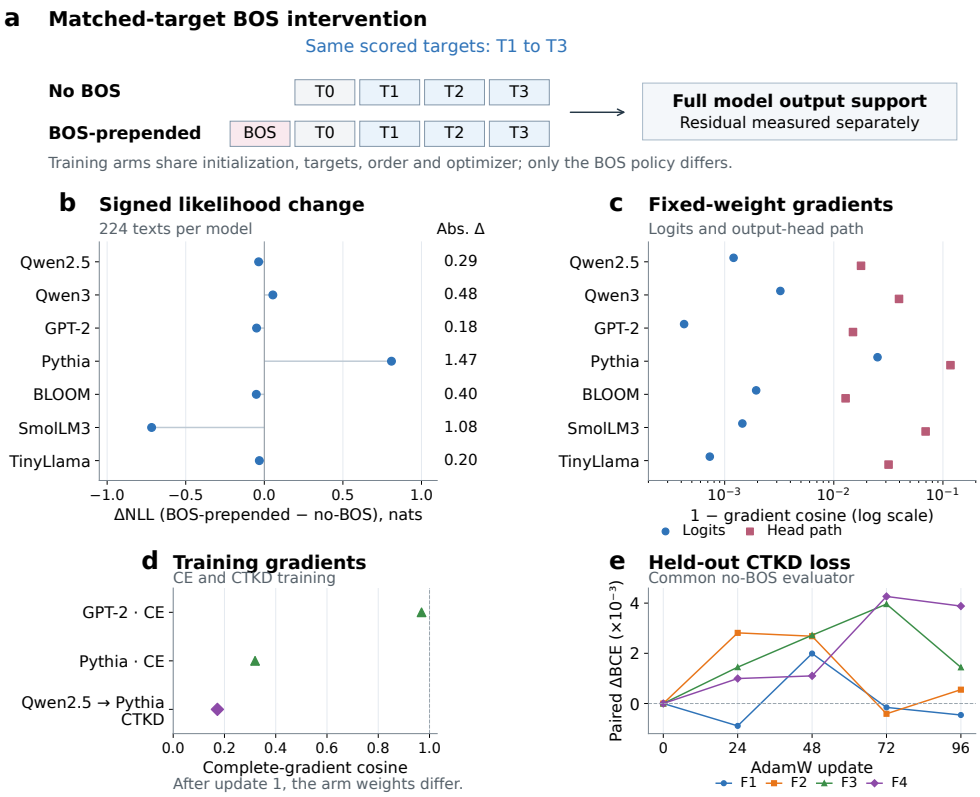


Figure S4: BOS prepending changes likelihoods, gradients and training outcomes. **a**, Blue tokens mark identical scored targets. **b**, Model medians of sentence-mean signed NLL change; median absolute change is shown separately. **c**, One minus median cosine for fixed-weight logit and output-head path gradients (likelihood and output-head diagnostics), not full parameter gradients. **d**, CE training uses the median of four rotation medians (16 steps each). CTKD training pools 384 steps. Weights differ after update 1. **e**, Byte-weighted held-out CTKD BCE difference (BOS-trained minus no-BOS-trained), both under the no-BOS evaluator. Lines connect five recorded evaluations per fold. F1 to F4 have disjoint evaluation sets, not independent model replications.

E Supplementary Evaluation and Reproducibility Checks

This section analyzes archived task outputs and additional interface summaries. We recomputed the reported statistics and checked the gradient-sample selection without rerunning training, reloading final weights or changing the grader. The supplementary materials contain the validation scripts, result files and complete configurations. The historical continuation status in Appendix D.1 remains separately recorded.

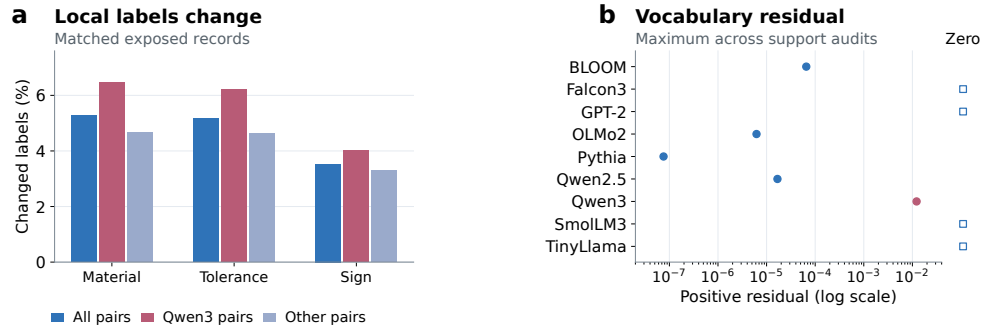
E.1 Task and Temperature Results

This separate Qwen2.5-7B-Instruct-to-Phi-4-mini-instruct study uses three paired training seeds. Table A8 reports MATH-500 scores, on a subset of MATH (Hendrycks et al., 2021), and fixed-answer temperature contrasts. The evaluation design and score reconstruction are documented below.

The task panel uses five sampling streams per model. The temperature panel rescores fixed reference answers at unchanged checkpoints, with no new responses or updates.

Checker exceptions or extraction failures affect 8.77% of outputs. This is substantial relative to the reported arm difference, so the task table is descriptive rather than evidence of a performance advantage. Appendix E.3 details the grading errors and missing controls.

Table A8 reports frozen-grader raw/LP mean pass@1 of 62.44%/61.80%. The LP-minus-raw difference is -0.64 points, with a question-bootstrap 95% interval $[-1.67, 0.40]$ conditional on the



c **Preserved failures and recorded follow-up**

Recorded run	Status	Recorded follow-up
Original breadth audit	FAIL	Separate repair audit
Exact two-byte diagnostic	INVALID	Exclude confirmation
One-step scatter diagnostic	INVALID	Retire route
StableLM-2 preflight	FAIL	Stop weight download
TinyLlama global offsets	FAIL	Freeze revised preflight
TinyLlama certificate coverage	FAIL	Retain coverage failure
Initial context preflight	INVALID PREFLIGHT	Freeze corrected protocol
Initial head reconstruction	FAIL	Freeze exact-hook protocol
Initial import preflight	FAIL RUNTIME	Freeze import-only correction

Figure S5: Contract sensitivity and immutable failures. **a**, Label changes under the joint context/residual intervention. **b**, Maximum model- vocabulary residual across the no-BOS repair, Falcon3 and TinyLlama support audits; positive values use a log axis and recorded zeros occupy a separate, non-numeric column. **c**, Preserved registered failures and their recorded follow-up. The joint-policy diagnostic is post hoc and cannot separate BOS prepending from residual masking.

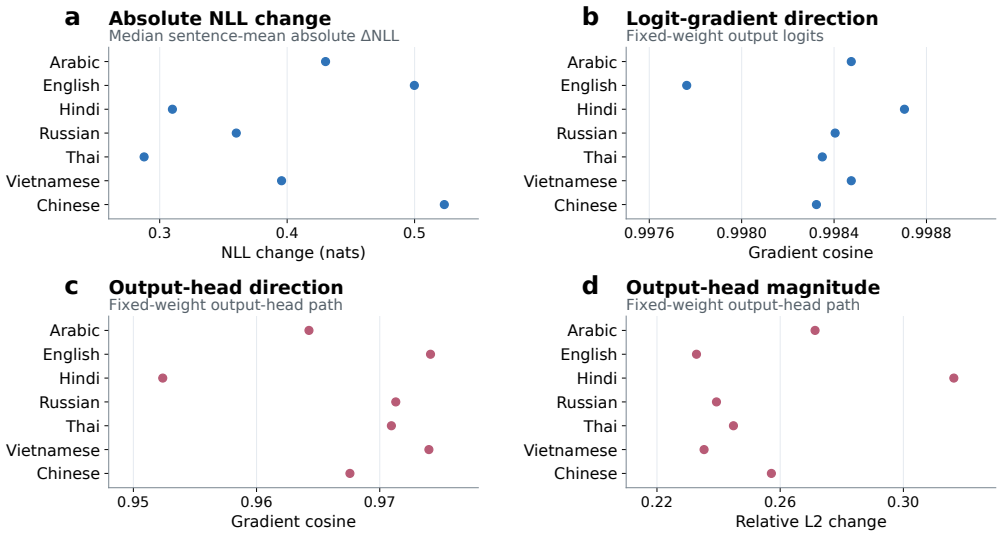


Figure S6: Context consequences span seven languages. Language-wise medians for **a**, Sentence mean absolute NLL change; **b**, logit-gradient cosine; **c**, output-head gradient cosine; and **d**, output-head relative L2 change. Each language contains 32 parallel groups evaluated over seven BOS-defined checkpoints. Each point pools the 224 sentence-model records for that language; axes use separate scales. Languages are categorical, so points are not joined.

six models. It establishes neither equivalence nor a reliable task advantage. The grader records 874 symbolic-checker exceptions and 441 extraction failures, including known equivalent-answer format errors; scores retain this grading limitation.

Table A8: **Separate three-seed task and evaluator comparison.** Task scores average five generation streams per model. Evaluator differences average three paired seeds and 500 reference answers per model. Here Δ is LP-trained minus raw-trained.

(a) MATH-500 pass@1 (%)

Seed	Raw	LP	Δ (pp)
R1	63.32	62.16	-1.16
R2	61.36	61.20	-0.16
R3	62.64	62.04	-0.60
Mean	**62.44**	**61.80**	-0.64

(b) Evaluator loss Δ

τ	Raw	LP
0.5	+0.0269	-0.0026
1	+0.0104	-0.0036
2	-0.0043	-0.0086

Negative loss Δ favors LP.
Higher task scores are better.

On fixed reference answers, temperature 0.5 or 1 makes each evaluator favor its own training gauge in all three pairs. At temperature 2, both favor LP. The reversal depends on the evaluation rule and temperature. Task performance is measured separately.

E.2 Training and task-evaluation design

Qwen2.5-7B-Instruct teaches Phi-4-mini-instruct in three paired training seeds. Both arms start from the same supervised-fine-tuning checkpoint and use matched continuation settings within each seed. The final scheduled checkpoint at update 156 is evaluated without selection by task score. Each student is tested on the same 500 questions with five paired sampling streams. The supplementary materials provide exact model revisions, seed values, optimizer settings and generation configurations.

The second and third seed pairs use a revised sequence-length rule, so the execution configuration is not identical across all three seeds. Within each seed pair, the limits are shared between arms, with no text truncation or sample dropping during training. The supplied outputs do not include complete final-weight inventories or the exact execution environment; numerical reconstruction does not by itself establish independent training reproduction.

E.3 Task-score reconstruction and grading limits

Each of the six output files contains 2,500 unique question/stream pairs. Across files, the same 500 question IDs are used with five streams; the streams are repeated generation measurements, not independent training runs. Recomputing frozen correctness labels gives Table A8: LP-minus-raw differences are -1.16, -0.16 and -0.60 percentage points. The mean is -0.64 points.

The supplied uncertainty procedure resamples question IDs jointly across arms, training seeds and streams, using 10,000 draws and linear 2.5%/97.5% quantiles. Replaying this procedure yields $[-1.6667, 0.4000]$ points, matching the delivered interval. This interval is conditional on the six fixed models and does not quantify uncertainty over a population of training seeds. The result does not satisfy the supplied impact rule requiring an absolute mean difference of at least two points and an interval excluding zero.

The frozen grader records 874 symbolic-checker exceptions and 441 extraction failures, with zero overlap between these two recorded categories. There are 325 length-truncated generations and no empty outputs. Checker exceptions are counted incorrect under the original rule. The records include mathematically equivalent ordered pairs rejected because of LaTeX formatting; numerical score reconstruction therefore does not establish grading correctness. The disjoint checker and extraction categories affect 1,315 of 15,000 outputs (8.77%). Truncations are reported separately and may overlap those categories. Their frequency is large relative to the 0.64-point arm difference; the question-bootstrap interval does not account for systematic grading errors. These results therefore do not support a task-level advantage, equivalence or absence of an effect. The delivery contains no no-distillation/SFT baseline scores and no verified contamination check between continuation data and MATH-500. We retain the frozen labels for traceability rather than treating numerical reconstruction as validation of the grader.

E.4 FIXED-ANSWER SCORES AND EVALUATION TEMPERATURE

The secondary delivery contains 500 fixed-answer records per model, 3,000 in total, with no new generations or updates. Each record identifies the training seed, arm and final step 156. Native NLL is normalized per student target token, not per response byte. SimCT scores use a common raw-teacher safeguard (threshold 2) and FP32 inputs; coordinate locks identify the paired raw/LP score comparison on a given endpoint.

We recomputed all 18 model/temperature rows of both question means and token-weighted summaries. At temperatures 0.5 and 1, each evaluator favors its own training arm in all three seed pairs. At temperature 2, both evaluators favor LP in every pair. The temperature panel in Table A8 averages the three paired differences at each temperature. Native NLL favors raw in all pairs under both supplied aggregations; its unit differs from the byte-normalized matched short training diagnostic. Evaluator scales are not compared across columns or temperatures.

E.5 GRADIENT SAMPLES AND LIMITS OF THE ADDITIONAL INTERFACE SUMMARIES

The fixed-parameter audit computes complete gradients over all trainable student parameters. The text manifest contains 224 unique GSM8K records with documented selection rules. This supplies the sample content; independent reacquisition of the original public shard snapshot remains a distinct reproducibility requirement. Exact execution settings are provided in the supplementary materials.

Table A9: **Delivered base-condition gradient summaries.** These rows are supplementary summaries, not additional independently reproduced complete-gradient audits. Sample selection was verified from the delivered manifest; the unresolved execution details are described in the text.

Pair label	Interface n	Gradient n	Cosine	Relative L2
Falcon3/Pythia	224	224	0.843989	0.5572
Qwen3/Pythia	224	16	0.859615	0.5137
Qwen2.5/Phi	224	16	0.980475	0.1995
Qwen2.5/Qwen3	224	16	1.000000	1.16×10^{-6}

The additional interface summary uses all 224 gradient samples for Falcon3 to Pythia and a shared deterministic 16-sample subset for each other pair. The subset is selected deterministically from stable text identifiers and shared across those model pairs; its construction is documented in the supplementary materials. Every pair has 224 interface records, so interface and gradient sample sizes must not be conflated.

The additional Falcon3 baseline has cosine 0.843989 and relative L2 0.557189, slightly different from the prospective fixed-parameter gauge audit's 0.843676 and 0.557666. These values are not substituted for the primary audit. Full checkpoint identifiers, per-sample gradient outputs and the executed parameter scope of the additional pair summaries are not all supplied; the fixed-parameter audit configuration is not assumed to describe every additional run. The near-overlap Qwen2.5/Qwen3 pair has vocabulary Jaccard 0.9999736 and no valid multi-token span-TV observations in its base summary. Its near-unit gradient cosine is a descriptive control under the delivered condition, not a general invariance proof.

Across all 38 interface conditions, the summary retains 8,066 configuration records: 7,787 valid, 277 with no exposure, and two unsupported. These are repeated configurations, not independent samples. The delivery includes temperature, safeguard, precision and support/residual variants, but the exact runner definitions for span bins and explicit virtual residual coordinates remain unresolved. Their original rows and statuses are retained in the result archive; we do not turn them into new main-text mechanism claims.