INTERMEDIATE DECODER LADDERS: BEYOND TEMPERATURE CALIBRATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Learned readouts of intermediate language-model states are usually fitted to the final layer's full predictive distribution. We test whether this target still lowers the observed next-token negative log-likelihood (NLL) once a competing readout fitted to the teacher's argmax receives its own validation-selected temperature. Holding the intermediate states, frozen head, readout family and search budget fixed, we fit six decoder families on Qwen3-8B and Qwen3.5-9B and fix every fit and temperature before scoring 1,200 additional deduplicated documents. The full target lowers observed NLL by 0.124 and 0.224 nats, with simultaneous 95% intervals excluding zero and all 240 configurations agreeing in sign; per-arm temperature scaling removes 72% and 61% of the uncalibrated gap but not the remainder. A Llama-3.1-8B-Instruct/WikiText replication gives 0.266–0.271 nats across three training seeds, and its full affine readouts keep a 0.56–0.72-nat gain when rescored without refitting on 400 untouched articles, both in the first 128 tokens and in a later window. On Llama, reassigning whole teacher distributions among same-argmax positions removes 74–81% of the gain, and swapping only the tail while keeping each position's argmax probability removes 61–71%. The gain is largest for full affine maps, which stay 0.14–0.16 nats ahead of rank-256 maps under an extended budget, and the fitted readout beats the best calibrated cheap readout by 2.4–2.7 nats. For observed NLL, the full teacher distribution is the better of the two tested targets.

1 INTRODUCTION

A language model's intermediate state can be mapped to vocabulary probabilities through its frozen output head. The raw logit lens decodes directly; learned lenses first translate the state (nostalgebraist, 2020; Belrose et al., 2023). A successful translation can recover final-model predictions long before the last block. Yet "recovery" can mean matching the teacher's top token, approximating its whole distribution, or predicting the token actually observed in text, and a decoder can improve one of these quantities while worsening another.

This ambiguity matters when interpreting a depth profile. Hard-target supervision rewards only the teacher's most likely token. Full-distribution supervision retains its alternatives, but an observed log-loss improvement might merely reflect less overconfidence, a better class prior, or generic softening. Likewise, whether a restricted decoder suffices depends on how well the larger map is optimized and on which low-rank comparators are fitted. A useful comparison separates these explanations before attributing a gain to the represented information.

We therefore ask: *does a full-distribution fitting procedure improve observed-token log loss after both it and the hard-target procedure receive their own validation-selected temperature, and how does the answer depend on readout capacity?* The design holds the intermediate states, normalization site, frozen head, parameterization and search ranges fixed. Six map families cross two targets on Qwen3-8B and Qwen3.5-9B at four depths and five document partitions; all readouts, temperatures and the primary contrast are fixed before 1,200 additional deduplicated documents are scored. A Llama-3.1-8B-Instruct/WikiText experiment tests the direction under a joint change of model family and text source and contains the target, capacity and baseline controls, whose fitted readouts are finally rescored, without refitting, on untouched articles.

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.

Relation to prior work and contributions. Lens studies (Belrose et al., 2023; Trimigno et al., 2026; Pettyjohn et al., 2026) fit the teacher distribution and report teacher Kullback–Leibler divergence (KL), agreement or the perplexity of the elicited predictions, without an argmax arm or a per-arm temperature. Distillation studies that vary the target (Hinton et al., 2015; Müller et al., 2019; Yuan et al., 2020; Tang et al., 2020; Furlanello et al., 2018) train the student end to end, in most cases a classifier and in Furlanello et al. (2018) also a language model scored by perplexity. Our design compares both targets for frozen-head readouts, gives each arm its own post-fit temperature (Guo et al., 2017), and scores observed-token NLL on new documents after every fit is fixed. Three results bear on the account of soft targets as regularizers or learned label smoothing in this setting. Per-arm temperature scaling removes the part of the gap that one logit scale can absorb, yet 0.124 and 0.224 nats remain on Qwen. On Llama, frozen-head readouts fitted to targets that keep each position's argmax with an entropy-matched uniform remainder, or with the permuted remainder of Born-Again Networks, have 0.71–1.12 nats higher NLL than readouts fitted to the full distribution. A target that keeps every position's argmax and its probability but swaps the remaining tail costs 0.35–0.45 nats. For these readouts, fitting the argmax target thus leaves observed log loss that the full target recovers. We contribute five tested statements about fitted readout procedures. **1. A residual gain that temperature does not remove.** The full target lowers observed-token NLL by 0.124 and 0.224 nats on the two Qwen checkpoints after per-arm calibration, with simultaneous intervals excluding zero and 240/240 configurations agreeing; resampling fitting seeds as well as documents keeps every family's interval below zero. The Llama/WikiText replication gives 0.266 nats, and the Llama full affine readouts keep 0.56–0.72 nats on 400 untouched articles in an early and a later 128-token window. Continuous or nine-bin conditional temperatures leave 0.55–0.65 nats for Llama full affine maps. **2. Target and readout family interact.** Under our fixed fitting budget, in all eight Qwen model–depth cells full affine gives the largest calibrated gain, 0.25–0.63 nats, and a d-parameter scale map the smallest, 0.04–0.12 nats. On Llama, a bias-only readout gains 0.03–0.05 nats against 0.55–0.65 for full affine. Within each setting the gaps are approximately the readouts' teacher-KL gaps times one fitted coefficient, which predicts held-out depths, families and control targets with mean absolute errors of 0.002–0.044 nats and, frozen in advance, the gaps of two later-registered targets with mean absolute errors of 0.005–0.017 nats, as Section 4.6 shows. **3. Sensitivity to reassigning teacher distributions.** On Llama/WikiText, a registered shuffle that reassigns whole teacher distributions within same-argmax, entropy-ordered blocks of eight and leaves the class marginal unchanged removes 74–81% of the soft-over-hard gain, and the registered tail swap above removes 61–71%, with every fit seed agreeing. **4. A finite-budget capacity comparison on observed NLL.** Rank 32 closes 53.5% and 64.3% of the Qwen rank-4-to-full gap. Rank 256 stays 0.13–0.35 nats behind full affine, and 0.14–0.16 nats with an extended budget and an identity-preserving zero start. **5. Value over the cheapest readouts.** Selected on validation NLL alone, the full-target affine readout beats a calibrated logit lens, bias-only readouts and a constant teacher-marginal predictor, and on the untouched articles the best cheap readout by 2.4–2.7 nats. Predeclared confirmations are the primary Qwen effects, the first Llama seed and the untouched-article contrasts; the seed-resampled intervals are post hoc, the exposed-article temperature checks are sensitivity analyses, and the teacher-proxy relation is post hoc apart from its registered prospective test.

2 Related work

Predictive lenses and their capacity. The logit lens decodes an intermediate state with the frozen output head (nostalgebraist, 2020). Tuned Lens learns an identity-initialized affine translator for each layer, applied before the frozen final normalization and trained by KL divergence to the final-layer distribution; it reports the perplexity of the elicited predictions and also describes a bias-only "debiased" logit lens; its 2025 revision reports much lower KL with the Muon optimizer (Belrose et al., 2023). Linear and normalized narrow shortcuts map intermediate states to final-layer representations for early exit (Yom Din et al., 2024; Seshadri, 2024). Low-Rank Lens replaces the dense map with a residual update $W_U(h + ABh)$ of rank $r = \lfloor H/k \rfloor$, $k \in \{4, 16, 32, 64\}$, trained and evaluated by KL to the final layer (Trimigno et al., 2026); its $k = 32$ is not our fixed $r = 32$. OmniLens sweeps biased identity-plus-rank-r translators with a zero-initialized up-projection against a full-rank lens on GPT-2, by teacher KL and top-1 agreement (Pettyjohn et al., 2026). Affine lenses trained by KL distillation to the final layer, following Tuned Lens, have also served as geometric probes of predictive information across depth (Lombardo et al., 2026). Our study holds these readout families fixed and varies the fitting target and its calibration. Among these studies we found none with an argmax-target arm, a

separate post-fit temperature for each arm, or a capacity comparison on observed-token NLL. Our primary insertion site follows the frozen normalization; the matched Llama controls cross both sites and the presence of a bias.

Soft targets and what carries their benefit. Full-distribution supervision comes from knowledge distillation (Hinton et al., 2015), and multi-exit networks distill later exits into earlier ones to improve early-exit accuracy (Phuong & Lampert, 2019). Menon et al. (2021) show that a teacher closer to the Bayes class probabilities lowers the variance of the student's objective. Other analyses credit much of the benefit to regularization: label smoothing improves calibration but makes teachers worse distillers (Müller et al., 2019); distillation behaves like learned label smoothing, and weak or hand-designed teachers can still help (Yuan et al., 2020); teacher knowledge decomposes into label-smoothing, class-relationship and instance-specific parts (Tang et al., 2020). Students trained on the dark knowledge with permuted predictions (DKPP) targets of Born-Again Networks, which permute the non-argmax teacher probabilities, still improve over their baselines (Furlanello et al., 2018). Born-Again Networks also distill Penn Treebank language models, scored by perplexity, and DistillLens aligns the logit-lens distributions of student and teacher layers (Dhakal et al., 2026); in both the student network is trained. We ask the question for a readout that shares the teacher's frozen head, scored by observed-token NLL after calibration, and compare entropy-matched and DKPP targets directly with the full target; in Section 4.6 both fall short of it. Because optimizing against an imperfect teacher can drift from the true objective (Tiapkin et al., 2025), we treat the relation between teacher-KL and observed-NLL gaps as an empirical property of our settings.

Calibration of final and intermediate predictions. Temperature scaling is a strong one-parameter post-hoc calibrator (Guo et al., 2017). Fan et al. (2024) relate a teacher's calibration to distillation outcomes; we leave the teacher unchanged and calibrate both fitted readouts identically after fitting. Entropy and token-frequency neurons regulate the confidence of final predictions (Stolfo et al., 2024), and the prediction-head bias encodes word frequency (Kobayashi et al., 2023); these findings motivate our raw-lens, bias-only and constant-marginal anchors. Joshi et al. (2025) track calibration across layers through untrained logit-lens projections of multiple-choice answer tokens, and Zhou et al. (2025) find late-intermediate layers better calibrated than the final layer and combine layers without training. Our readouts are instead fitted and scored over the full vocabulary on observed next tokens, each with its own global or confidence-conditional temperature.

Predictive versus causal interpretation. Probe results depend on the probe's capacity and need controls (Hewitt & Liang, 2019; Voita & Titov, 2020; Belinkov, 2022), and intervention methods patch representations to decode what they encode or whether they are actually used (Ghandeharioun et al., 2024; Zhang & Nanda, 2024). Our readouts measure how accessible a prediction is under a fixed fitting procedure; Appendix L shows why a small error in probability values does not bound the error in intervention derivatives.

3 Controlled decoder comparisons

3.1 A fixed site and distinct endpoints

For a decoder with L blocks and width d, let $h_{k,t}$ be the residual state after block k at position t, and let $x_{k,t} = \mathrm{LN}(h_{k,t})$ use the frozen final normalization. With frozen head W_{lm}, the teacher distribution, its hard target and the prediction of a readout f are

$$\begin{aligned} p_{L,t} &= \mathrm{softmax}(W_{\mathrm{lm}} x_{L,t}), \\ y^*_t &= \arg\max p_{L,t}, \\ q^f_{k,t} &= \mathrm{softmax}(W_{\mathrm{lm}} f(x_{k,t})). \end{aligned}$$

The raw lens sets f to the identity.

We report teacher agreement $\Pr[\arg\max q^f = y^*]$, teacher forward KL $\mathbb{E}\,\mathrm{KL}(p_L \| q^f)$, observed next-token accuracy $\Pr[\arg\max q^f = y]$, and observed NLL $-\mathbb{E}\log q^f(y)$. Here y is the next token in the document, rather than the teacher's prediction. Teacher endpoints include every retained position; observed endpoints omit the final position of each document because it has no retained successor. All comparisons within an endpoint use identical positions.

3.2 FAMILIES AND FITTING TARGETS

The families share x_k and the frozen head:

Family	Map $f(x)$	Parameters
Positive scalar	sx, $s > 0$	1
Bias only	$x + \beta$	d
Scale only	$\gamma \odot x$	d
Diagonal affine	$\gamma \odot x + \beta$	$2d$
Identity plus rank r	$(I + UV^\top)x$	$2rd$
Full affine	$x + Wx + b$	$d^2 + d$

The primary natural-text study fits scale, diagonal, their rotated counterparts, rank four and full affine. A predeclared secondary rank-32 bridge adds $64d$ parameters at layers 8 and 16, under the identical train/validation fitting recipe. Full affine contains the other affine families at this fixed insertion site, but a finite-budget fitted comparison is not a comparison of globally optimal risks.

The two fitting targets are

$$\begin{aligned} \mathcal{L}_{\mathrm{hard}}(f) &= -\mathbb{E} \log q^f(y^*), \\ \mathcal{L}_{\mathrm{soft}}(f) &= \mathbb{E}\, \mathrm{KL}(p_L \| q^f). \end{aligned} \tag{1}$$

Soft training computes the full vocabulary distribution at teacher temperature one. Minimizing its cross-entropy is equivalent to minimizing forward KL because teacher entropy does not depend on f. No top-k approximation, label smoothing or observed-token supervision enters either objective.

Common natural-text budget. Every family and objective uses Adam without weight decay, batch size 256 and thirty epochs. Learning rates are $\{10^{-1}, 3\times10^{-2}, 10^{-2}, 3\times10^{-3}, 10^{-3}, 3\times10^{-4}, 10^{-4}, 3\times10^{-5}, 10^{-5}, 3\times10^{-6}\}$. Selection minimizes the corresponding validation objective over learning rate and epoch; exact ties retain the first rate and earliest epoch. Paired hard/soft fits share initialization and shuffle seeds. The original six-family comparison selects 480 fits; two select a rate boundary and 107 select epoch 30. Changing the target also changes its validation selection criterion, so the estimand is the complete fitting-and-selection procedure.

3.3 CONFIDENCE AND BASIS CONTROLS

For each selected readout, a post-fit scalar multiplies its logits. We compare identity scaling with two validation selections over 33 positive inverse temperatures logarithmically spaced from $1/16$ to 16 with adjacent ratio $2^{1/4} \approx 1.19$: minimum teacher KL, or minimum observed NLL. The latter explicitly uses observed validation labels. Neither uses test outcomes. All 1,040 selected Qwen scalars lie inside the grid, at inverse temperatures 0.42–1; Qwen effects refer to this grid, and Section 3.7 examines continuous optimization on Llama. A positive scalar preserves the argmax; we also check every selected scaling against the unscaled finite-arithmetic argmax.

For a fixed random orthogonal matrix Q, rotated diagonal and scale maps are $Q\,\mathrm{diag}(\gamma)Q^\top x + \beta$ and $Q\,\mathrm{diag}(\gamma)Q^\top x$, with the same parameter counts as their native counterparts; they compare fitted risks under two coordinate restrictions.

3.4 DATA, ARITHMETIC AND UNCERTAINTY

Natural text uses a fixed 1,000-document FineWeb-Edu pool (Penedo et al., 2024): normalized unique documents from one shard, truncated to 128 tokens without concatenation. Five seeds, 42, 1042, 2042, 3042 and 4042, define 400/200/400 document partitions. The six-family comparison covers Qwen3-8B (Yang et al., 2025) and Qwen3.5-9B (Qwen Team, 2026) at layers 8, 16, 24 and 31. Every family is fitted on the same normalized states with the frozen output head. This pool is development data, inspected before this study; its complete 2,400-row comparison is in Appendix I and does not enter the confirmation below.

3.5 CONFIRMATION ON ADDITIONAL DOCUMENTS

We select 1,200 additional documents from the same immutable FineWeb-Edu shard by a fixed, output-independent pseudorandom ordering. Eligibility requires at least 256 normalized characters. Selection excludes the 1,000-document development pool, normalized whole-document duplicates, identical 128-token prefixes under either tokenizer, and five-word-shingle Jaccard similarity at least 0.8 against development or newly selected documents. This separates the documents from the development pool, not from checkpoint pretraining.

All 480 original readouts and 40 rank-32 readouts, including their three scalar choices, are fixed before any confirmation endpoint is computed. All selected readouts are scored on the same intermediate states: 151,707 observed-token positions for Qwen3-8B and 151,756 for Qwen3.5-9B.

The primary effect for model m is

$$\Delta_m = \frac{1}{120} \sum_{c=1}^{120} \left[\mathrm{NLL}_{m,c,\mathrm{soft}}(\hat{\alpha}^{\mathrm{NLL}}_{m,c,\mathrm{soft}}) - \mathrm{NLL}_{m,c,\mathrm{hard}}(\hat{\alpha}^{\mathrm{NLL}}_{m,c,\mathrm{hard}}) \right], \tag{2}$$

where c enumerates the six original families, four depths and five original fit seeds, and each NLL is a token-weighted document aggregate. Both scalars $\hat{\alpha}^{\mathrm{NLL}}$ were selected on the original validation observed-NLL objective. Rank 32 remains secondary and is excluded from this mean.

We resample the 1,200 documents jointly across every configuration, recompute token-weighted losses, and form 5,000 percentile bootstrap draws. Per-model 97.5% intervals give nominal simultaneous 95% coverage for the two declared effects by Bonferroni adjustment. The prespecified directional criterion is that both upper endpoints lie below zero. These intervals condition on the frozen fits and the sampled documents; family/depth strata and alternate calibration rules are descriptive.

3.6 AN UNRELATED MODEL–CORPUS REPLICATION

We repeat the fixed-site, hard/soft comparison with Llama-3.1-8B-Instruct (Grattafiori et al., 2024) on WikiText-103 (Merity et al., 2017). This one compound replication changes model family and corpus together. Article-level roles are 400 training, 200 calibration and 400 confirmation articles from the public training shard, disjoint by identity, normalized text and token sequence; each article contributes its first 128 tokens.

Diagonal affine, rank 32 and full affine are fitted at layers 8 and 16 with the identical ten-rate, thirty-epoch recipe and three scalar rules. All 12 selected fits and scalars of the first seed, 42, are fixed before the confirmation states are extracted. Two further training seeds use the same states and change only initialization and minibatch order. For each seed the primary statistic is the equal mean of six calibrated soft-minus-hard NLL contrasts, with a 10,000-draw 400-article bootstrap and nominal 95% intervals; seed variation is reported separately.

3.7 MATCHED MECHANISM CONTROLS

On the exposed Llama/WikiText articles, a matched full-affine comparison crosses layers 8/16, translation before/after final normalization, common 30/60-epoch maxima and objective-specific/common observed-NLL selection. It replaces the scalar grid with bounded continuous optimization and tests biased/unbiased rank-32 and rank-128/256 maps; Appendix D gives the details. These are sensitivity checks on fitted procedures, not additional held-out data.

Separate comparisons at layers 8 and 16 fit five teacher-derived targets: the full distribution, its argmax, an entropy-matched distribution with uniform nonmaximum mass, the DKPP random permutation of its nonmaximum probabilities, and a registered stratified shuffle. The shuffle deranges complete teacher distributions within same-argmax blocks of consecutive entropy rank that hold at least two positions; blocks hold eight positions, a short trailing block is merged into the preceding one, and Appendix E gives the construction. In total, 47,239 fitting positions receive another position's distribution and 3,961 singleton positions keep their own. All targets use the same biased affine map before final normalization and four paired fit seeds. The 1,000 articles have disjoint roles: 400 for fitting, 100 for scalar fitting, 100 for checkpoint selection and 400 for evaluation. Each target receives ten learning rates, 60 epochs and twelve checkpoint evaluations. Global or nine-bin

confidence-conditional temperatures are fitted on the scalar subset before a common observed-NLL selector compares checkpoints on the selection subset. The same roles, seeds and selectors define the cheap references: bias-only maps $x + \beta$ under either target, the raw lens with a global or conditional temperature, and two constant predictors with fitted temperatures, namely the mean teacher distribution and the smoothed observed unigram of the fitting articles. A readout policy and a cheap comparator are chosen by minimum seed-mean NLL on the checkpoint-selection articles. A registered extended-budget comparison with 12 rates and 120 epochs refits full affine and rank-256 maps, the latter from the original small random start and from an identity-preserving start with a zero output factor, as described in Appendix D.

Untouched articles and a later window. The same fixed-order selection, excluding duplicates of all 1,000 earlier articles, supplies the first 400 unused articles with at least 512 Llama tokens. Every readout was fixed on the earlier articles' roles before being scored on this cohort, which Appendix C describes. Without refitting, each readout is scored on the first 128 tokens and on tokens 385–512. The predeclared endpoints are full-minus-argmax NLL at both depths and calibrated selectors, and the validation-selected policy minus the cheap comparator; each family of four contrasts has 98.75% intervals from a 20,000-draw article bootstrap.

4 RESULTS

4.1 A DISTRIBUTIONAL BENEFIT BEYOND SCALAR CALIBRATION

All 520 readouts in 40 cells yield 1,560 scalar evaluations. With each target's observed-NLL calibration, soft lowers NLL by 0.124 and 0.224 nats on Qwen3-8B and Qwen3.5-9B, and both simultaneous intervals in Table 1 exclude zero.

The benefit is consistent over the declared strata. Every one of the 120 configurations per model has lower calibrated NLL under soft supervision. Depth-level mean reductions range from 0.113 to 0.132 nats for Qwen3-8B and 0.186 to 0.268 for Qwen3.5-9B, as Figure 1 shows. These configuration summaries share documents; the uncertainty statement belongs to Equation 2.

How much remains after confidence correction? Without calibration, mean soft-minus-hard NLL is -0.436 and -0.578 nats. With the original teacher-KL-selected scalars it is -0.142 and -0.225; with observed-NLL-selected scalars it is -0.124 and -0.224. The last comparison attenuates the raw gap by 71.6% and 61.2% and leaves the confirmed residual. Temperature preserves all selected argmax predictions in all 1,560 evaluations, so the remaining distributional change is not equivalent to selecting a different top token through the scalar. Mean teacher KL also improves, by 0.103 and 0.216 nats after observed-NLL calibration. Our claims concern log loss: the argmax target is the one fitted for top-1 agreement and remains the natural choice when top-1 accuracy is the endpoint; accuracy and agreement of every readout are in the supplementary tables.

Which families retain the benefit? The residual depends on the readout family. Appendix Table 6 reports every family, with rank 32 separated by its early-depth scope. Among the original six, mean reductions range from 0.049 to 0.324 nats on Qwen3-8B and from 0.101 to 0.528 on Qwen3.5-9B. In each of the eight model–depth cells, full affine has the largest calibrated reduction, 0.25–0.63 nats, and the d-parameter scale map the smallest, 0.04–0.12 nats; every family retains a reduction in all of its configurations. Also resampling the five fitting seeds gives 95% primary intervals $[-0.126, -0.122]$ and $[-0.227, -0.221]$, and in Appendix F every family's interval excludes zero. The equal-family mean was fixed before these outcomes; Section 4.6 relates the ordering to teacher KL.

4.2 A STRONGER LOW-RANK CAPACITY BRIDGE

To tighten the low-rank comparison with full affine, we add a fixed rank-32 condition at the two early depths, using only the original training and validation data and the same ten-rate, thirty-epoch recipe. All 40 selected rank-32 weights and their scalars are fixed before confirmation.

Table 2 shows a substantial intermediate regime. For soft fits, rank 32 closes 53.5% of the rank-four-to-full NLL gap on Qwen3-8B and 64.3% on Qwen3.5-9B, averaged over layers 8 and 16. Full affine

Table 1: Primary confirmation on 1,200 additional documents. Both targets use their own observed-NLL-selected scalar. Means equally weight 120 configurations per model; each configuration uses token-weighted losses on the same documents. Intervals are per-model 97.5% document-bootstrap confidence intervals (CI), which jointly give 95% coverage for the two NLL effects.

Model	Hard NLL	Soft NLL	Δ NLL [CI]	Lower NLL
Qwen3-8B	5.372	5.248	-0.124 $[-0.126, -0.122]$	120/120
Qwen3.5-9B	4.999	4.775	-0.224 $[-0.228, -0.221]$	120/120

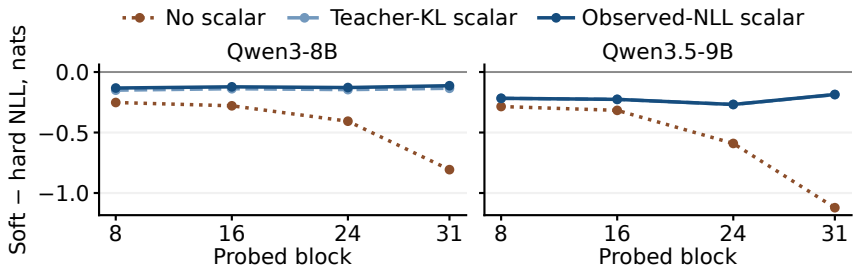


Figure 1: Confirmation on 1,200 additional documents: soft-minus-hard observed NLL, averaged over the six original families and five frozen fit seeds at each depth, for the three fixed calibration rules, of which the observed-NLL scalar is primary. These descriptive means share documents; Table 1 gives the document-level inference.

still improves NLL by 0.318 and 0.287 nats over rank 32. On the rank-32 bridge, too, soft supervision lowers calibrated NLL in all ten configurations per model.

4.3 Replication across family, corpus and training seed

With all 120 candidate fits per seed completed, Table 3 shows mean calibrated NLL reductions of 0.266, 0.268 and 0.271 nats for the original seed and two optimization repeats. Table 8 shows that all six family–depth contrasts are negative under every seed. Under all three scalar rules, NLL and teacher KL improve in all 54 paired rows of Appendix B.

All three seeds keep full affine below rank 32, and no selected rate or scalar lies on a grid boundary; Appendix B.1 gives absolute anchors.

4.4 Fitted readouts against the cheapest calibrated alternatives

A fitted readout is useful only if it beats inexpensive corrections of the logit lens. Table 4 places the raw lens with a global or nine-bin conditional temperature, two constant predictors and bias-only readouts under both targets beside the full affine readout, using the same Llama roles, seeds and selectors. A learned bias improves on the teacher-marginal predictor by 0.18 and 0.78 nats. The full affine readout adds a further 2.72 nats over the best bias-only readout, averaged over depths and seeds, with 99.17% interval $[-2.78, -2.66]$, 7.02 nats over the calibrated raw lens, with interval $[-7.11, -6.93]$, and 2.91 and 3.49 nats over the constant teacher marginal. A readout policy chosen on validation NLL alone beats the validation-selected cheap readout by 2.74 nats in Table 24.

4.5 Confirmation on untouched articles and a later window

The Llama controls and cheap references above use previously inspected articles and 128-token prefixes. Table 5 rescores the same readouts, without refitting, on 400 untouched articles. In the prefix window the full target lowers calibrated NLL relative to the argmax target by 0.56–0.66 nats, matching the 0.55–0.65 nats on the exposed articles; in the later window the gain is 0.61–0.72 nats, and all eight intervals exclude zero. The entropy-matched, tail-permuted, stratified-shuffle and tail-swap targets remain 0.73–0.94, 0.99–1.31, 0.44–0.58 and 0.37–0.50 nats worse than the full target, and the validation-selected policy beats the validation-selected cheap readout by 2.43–2.73 nats.

Table 2: Early-depth capacity bridge on the additional documents. Means over layers 8/16 and five fit seeds, ten configurations, using the original observed-NLL-selected scalars. NLL in nats, accuracy in percent. Counts exclude the frozen backbone/head. These are descriptive fixed-budget comparisons.

Model	Family	Parameters	Hard NLL	Soft NLL	Soft accuracy
Qwen3-8B	Diagonal	$2d$	6.296	6.189	14.957
	Rank 4	$8d$	6.049	5.937	16.101
	Rank 32	$64d$	5.747	5.571	18.954
	Full affine	$d^2 + d$	5.594	5.253	20.818
Qwen3.5-9B	Diagonal	$2d$	6.263	6.105	17.685
	Rank 4	$8d$	5.802	5.515	19.894
	Rank 32	$64d$	5.445	4.998	23.390
	Full affine	$d^2 + d$	5.260	4.711	24.639

Table 3: Llama-3.1-8B-Instruct, 400 WikiText confirmation articles: equal mean of the six predeclared family–depth pairs under observed-NLL scalars, with nominal 95% article-bootstrap intervals per fitted seed. The last two seeds reuse these articles; no pooling or best-seed selection.

Training seed	Hard NLL	Soft NLL	Δ NLL [article CI]
First (42)	5.473	5.207	-0.266 $[-0.274, -0.259]$
Second	5.484	5.215	-0.268 $[-0.275, -0.261]$
Third	5.480	5.209	-0.271 $[-0.278, -0.263]$

4.6 Selection, capacity and class correspondence

All 48 matched depth/site/horizon/selection/calibration contrasts favor soft NLL; with a common observed-NLL selector, continuous calibration and a common 60-epoch maximum, the reduction in Appendix D is 0.510–0.668 nats at both depths and insertion sites.

Which property of the teacher carries the gain? On the exposed articles, at both depths and under both calibrated selectors, the full target has lower NLL than the entropy-matched target by 0.715–0.813 nats and than the tail-permuted target by 0.980–1.115 nats, with every paired fit seed agreeing in sign in Appendix Table 22. The registered stratified shuffle gives 92.3% of positions another genuine teacher distribution from their same-argmax, entropy-ordered block, leaving the class marginal unchanged. Fitted identically, it improves on the argmax target, with 5.112 versus 5.257 nats at layer 8, but the full target beats it by 0.501 and 0.513 nats at layer 8 and 0.432 and 0.443 at layer 16. These 98.75% intervals exclude zero, also with fit seeds resampled in Appendix F; all four fit seeds agree, and the untouched articles of Table 5 repeat the gap. Under this procedure the reassignment removes 74–81% of the gain but does not hold each position's own confidence or entropy. A registered check in Appendix Table 20 confirms that the shuffled targets, like the full target, are producible by the frozen head at all 2,048 sampled positions to within 4×10^{-5} nats, so the cost concerns fitting one shared map to the reassigned pairs. A registered follow-up keeps every position's argmax and argmax probability exactly and swaps only the remaining tail between same-argmax positions of adjacent tail entropy, with a mean absolute entropy change of 0.11 nats. The full target still beats it by 0.35–0.45 nats, and a class-mean tail with each position's own confidence by 0.30–0.32; all eight intervals exclude zero, also with fit seeds resampled, as do the untouched-article gaps of 0.30–0.50 nats in Appendix H. Replacing position-specific tails thus costs observed NLL with argmax and confidence kept; as each replacement also alters other target properties and the fit, the fractions describe these operations, not one property's share.

How much expressivity remains useful? Before normalization, rank 256 comes within 0.134 and 0.146 nats of full affine at layers 8 and 16 with 12.5% of the parameters; after normalization the gaps are 0.353 and 0.288 nats, as Table 17 shows. A registered extended budget with one fit seed, 12 rates and 120 epochs, from either the original random start or an identity-preserving zero start, leaves rank 256 0.144–0.155 nats behind full affine at both depths in Table 18. Full affine thus keeps a finite advantage for these Adam recipes, at 2.7% more multiply-adds per position including the shared head; as optimizers alone change lens fits (Belrose et al., 2023), the gaps are recipe-specific.

Table 4: Cheapest calibrated alternatives on Llama-3.1-8B-Instruct/WikiText: 400 evaluation articles, pre-normalization site, seeded rows averaging four fits; NLL in nats, accuracy in percent. Parameters: raw lens 0 plus 1 or 9 scalars, constant V plus 1, bias d, full affine $d^2 + d$.

			Layer 8		Layer 16	
Readout	Target	Temperature	NLL	Acc.	NLL	Acc.
Raw lens	–	global	11.694	0.15	10.996	1.36
Raw lens	–	conditional	11.692	0.15	10.996	1.36
Observed unigram	–	global	7.696	4.51	7.696	4.51
Teacher marginal	–	global	7.519	4.51	7.519	4.51
Bias only	argmax	global	7.388	7.43	6.782	12.29
Bias only	argmax	conditional	7.385	7.43	6.774	12.19
Bias only	full	global	7.338	7.66	6.742	12.14
Bias only	full	conditional	7.335	7.66	6.748	12.23
Full affine	full	global	4.610	28.25	4.034	33.50
Full affine	full	conditional	4.593	28.25	4.019	33.50

Table 5: Llama readouts rescored without refitting on 400 untouched WikiText articles: full affine before normalization, conditional temperature, four fits averaged. Observed-NLL differences in nats, with 98.75% paired-article intervals for full minus argmax; the tail swap keeps each position's argmax and its probability; policy accuracy differences in points. Appendix C has all values.

		Full − argmax target		Full − control target				Policy − cheap	
Tokens	Layer	ΔNLL	Interval	Ent. m.	Tail perm.	Shuffle	Tail swap	ΔNLL	ΔAcc.
1–128	8	-0.640	$[-0.662, -0.619]$	-0.819	-1.119	-0.511	-0.451	-2.71	$+19.9$
1–128	16	-0.557	$[-0.576, -0.537]$	-0.729	-0.992	-0.450	-0.373	-2.73	$+21.1$
385–512	8	-0.716	$[-0.740, -0.692]$	-0.936	-1.308	-0.581	-0.497	-2.43	$+16.5$
385–512	16	-0.612	$[-0.632, -0.590]$	-0.817	-1.152	-0.509	-0.423	-2.53	$+17.6$

A teacher-proxy account of the gap size. If observed tokens were teacher samples, the observed-NLL difference of two fixed readouts would equal their teacher-KL difference in expectation; as our endpoints meet this only approximately, we test $\Delta\mathrm{NLL} \approx b\,\Delta\mathrm{KL}$ with one coefficient per setting and window, fitted through the origin as detailed in Appendix G. On Qwen, the 260 soft-minus-hard configuration differences correlate at 0.965, with $b = 1.04$ and 1.14, and a pooled coefficient fitted without one depth or one family predicts its seed-averaged gaps with mean absolute error (MAE) 0.017–0.044 nats. On Llama, a coefficient fitted to three control targets predicts the fourth target's gap with MAE 0.002–0.036 nats. These are retrospective holdouts; coefficients frozen before the tail-control fits then predicted all 96 of their cell gaps with MAE 0.005–0.017 nats, and all six registered tests in Appendix H pass. b is 0.85–0.86 in the first 128 tokens and 0.98 in tokens 385–512. The account takes the measured KL gaps as given and predicts from them the size of the NLL gap; it also places full affine first under both endpoints in all eight Qwen model–depth cells.

5 Discussion and conclusion

A predictive lens should be assessed as a fitted distribution. After matched validation scalar selection, full-distribution fitting lowers observed NLL on two Qwen checkpoints and on Llama/WikiText, where it is stable across training seeds and holds without refitting on untouched articles in two windows. Entropy-matched and tail-permuted targets do not reproduce it; reassigning whole teacher distributions within same-argmax blocks removes 74–81% of it, and swapping only the tail at a fixed argmax probability 61–71%. Its size tracks the teacher-KL gap within each setting, prospectively for the tail controls. Under our fitting budget full affine maps gain most, and fitted maps beat every cheap calibrated readout, the validation-selected one by 2.4–2.7 nats on untouched articles. For the observed NLL of frozen-head readouts, our results favor the full teacher distribution over its argmax, calibrated afterwards; the evidence covers these two teacher-derived targets on English text and 8–9B models, not observed-token training or target mixtures.

AI USE STATEMENT

In this work, we used generative AI tools to design and improve experiments and to implement methods. We have not used generative AI tools to support qualitative and thematic data analysis, and survey or interview design and the transcription of recordings are not applicable to this work. Additionally, we used generative AI tools to edit the paper to improve readability and to create or edit software code. We have reviewed all AI-assisted work: AI-assisted code was tested before use, and every reported number is computed from recorded experimental outputs by analysis scripts. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies readouts of intermediate states of the publicly released language models Qwen3-8B, Qwen3.5-9B and Llama-3.1-8B-Instruct on the public English text corpora FineWeb-Edu and WikiText. It involves no human subjects, no user data and no newly collected personal data. Web text can incidentally contain personal information; we report only aggregate loss statistics, and the supplement contains per-document statistics, the identifiers and roles of the public documents and articles used, and code that rebuilds their texts from the public corpora. The fitted readouts are diagnostic tools for measuring what intermediate states predict; they do not add capabilities to the models, and we see no direct harmful or dual-use application beyond those of the public models themselves. Our conclusions concern English text and models of 8–9B parameters, and readouts evaluated this way may behave differently on other languages or text sources.

REPRODUCIBILITY STATEMENT

The readout families, targets, calibration and scoring are defined in Section 3, the confirmation and Llama protocols in Sections 3.5–3.7, and the data sources, document separation and statistical procedures in Appendices A, B and J. The target controls are specified in Appendices E and H, the assumptions and complete proof of the counterexample in Appendix L, and the registered contrasts in Appendix M. Table 33 lists the evidence supplied for every result. The anonymous supplementary material contains the document and article identifiers with their roles, code that rebuilds the texts from the public corpora, reference implementations of every step from corpus preparation to scoring, and the recorded statistics listed in Appendix K.

REFERENCES

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7/.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the Tuned Lens. *arXiv preprint arXiv:2303.08112*, 2023. URL https://arxiv.org/abs/2303.08112. v1 2023; v6 (Nov. 2025) adds Muon training.

Manish Dhakal, Uthman Jinadu, Anjila Budathoki, Rajshekhar Sunderraman, and Yi Ding. DistillLens: Symmetric knowledge distillation through logit lens. *arXiv preprint arXiv:2602.13567*, 2026.

Wen-Shu Fan, Su Lu, Xin-Chun Li, De-Chuan Zhan, and Le Gan. Revisit the essence of distilling knowledge through calibration. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12882–12894, 2024. URL https://proceedings.mlr.press/v235/fan24d.html.

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

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. Patchscopes: A unifying framework for inspecting hidden representations of language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 15466–15490, 2024. URL https://proceedings.mlr.press/v235/ghandeharioun24a.html.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, 2019. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

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. NIPS 2014 Deep Learning Workshop.

Abhinav Joshi, Areeb Ahmad, and Ashutosh Modi. Calibration across layers: Understanding calibration evolution in LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 14686–14714, 2025. doi: 10.18653/v1/2025.emnlp-main.742. URL https://aclanthology.org/2025.emnlp-main.742/.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. Transformer language models handle word frequency in prediction head. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 4523–4535, 2023. doi: 10.18653/v1/2023.findings-acl.276. URL https://aclanthology.org/2023.findings-acl.276/.

Gianfranco Lombardo, Giuseppe Trimigno, and Stefano Cagnoni. A geometric perspective on next-token prediction in large language models: Three emerging phases. *arXiv preprint arXiv:2605.09011v1*, 2026. URL https://arxiv.org/abs/2605.09011.

Aditya Krishna Menon, Ankit Singh Rawat, Sashank J. Reddi, Seungyeon Kim, and Sanjiv Kumar. A statistical perspective on distillation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7632–7642, 2021. URL https://proceedings.mlr.press/v139/menon21a.html.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *The Fifth International Conference on Learning Representations (ICLR)*, 2017. URL https://openreview.net/forum?id=Byj72udxe.

Rafael Müller, Simon Kornblith, and Geoffrey E. Hinton. When does label smoothing help? In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/f1748d6b0fd9d439f71450117eba2725-Abstract.html.

nostalgebraist. interpreting GPT: the logit lens. LessWrong blog post, 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The FineWeb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems*, volume 37, pp. 30811–30849, 2024. doi: 10.52202/079017-0970. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/370df50ccfdf8bde18f8f9c2d9151bda-Abstract-Datasets_and_Benchmarks_Track.html.

Jordan Pettyjohn, Mansi Sakarvadia, Nathaniel Hudson, Daniel McKenzie, Kyle Chard, and Ian Foster. Interpreting language model hidden states at scale. *arXiv preprint arXiv:2608.10260*, 2026. URL https://arxiv.org/abs/2608.10260.

Mary Phuong and Christoph H. Lampert. Distillation-based training for multi-exit architectures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1355–1364, 2019. doi: 10.1109/ICCV.2019.00144. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Phuong_Distillation-Based_Training_for_Multi-Exit_Architectures_ICCV_2019_paper.html.

Qwen Team. Qwen3.5-9B. https://huggingface.co/Qwen/Qwen3.5-9B, 2026. Model card.

Amrit Diggavi Seshadri. Normalized narrow jump to conclusions: Normalized narrow shortcuts for parameter efficient early exit transformer prediction. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 5187–5192, 2024. doi: 10.18653/v1/2024.findings-emnlp.298. URL https://aclanthology.org/2024.findings-emnlp.298/.

Alessandro Stolfo, Ben Wu, Wes Gurnee, Yonatan Belinkov, Xingyi Song, Mrinmaya Sachan, and Neel Nanda. Confidence regulation neurons in language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 125019–125049, 2024. doi: 10.52202/079017-3970. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/e21955c93dede886af1d0d362c756757-Abstract-Conference.html.

Jiaxi Tang, Rakesh Shivanna, Zhe Zhao, Dong Lin, Anima Singh, Ed H. Chi, and Sagar Jain. Understanding and improving knowledge distillation. *arXiv preprint arXiv:2002.03532*, 2020. URL https://arxiv.org/abs/2002.03532.

Daniil Tiapkin, Daniele Calandriello, Johan Ferret, Sarah Perrin, Nino Vieillard, Alexandre Ramé, and Mathieu Blondel. On teacher hacking in language model distillation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 59552–59569. PMLR, 2025. URL https://proceedings.mlr.press/v267/tiapkin25a.html.

Giuseppe Trimigno, Gianfranco Lombardo, and Stefano Cagnoni. Low-rank lens for scalable LLMs interpretability. In *Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2026)*, pp. 35–40, 2026. doi: 10.14428/esann/2026.ES2026-221. URL https://www.esann.org/sites/default/files/proceedings/2026/ES2026-221.pdf.

Elena Voita and Ivan Titov. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 183–196, 2020. doi: 10.18653/v1/2020.emnlp-main.14. URL https://aclanthology.org/2020.emnlp-main.14/.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Alexander Yom Din, Taelin Karidi, Leshem Choshen, and Mor Geva. Jump to conclusions: Short-cutting transformers with linear transformations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 9615–9625, 2024. URL https://aclanthology.org/2024.lrec-main.840/.

Li Yuan, Francis E. H. Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3902–3910, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Yuan_Revisiting_Knowledge_Distillation_via_Label_Smoothing_Regularization_CVPR_2020_paper.html.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=Hf17y6u9BC.

Ej Zhou, Caiqi Zhang, Tiancheng Hu, Chengzu Li, Nigel Collier, Ivan Vulić, and Anna Korhonen. Beyond the final layer: Intermediate representations for better multilingual calibration in large language models. *arXiv preprint arXiv:2510.03136*, 2025. URL https://arxiv.org/abs/2510.03136.

A COMPLETE CONFIRMATION EVIDENCE

The confirmation contains every prespecified configuration: 480 original readouts plus 40 rank-32 readouts, each with identity, teacher-KL-selected and observed-NLL-selected scalars. All 520 weights and 1,560 scalar evaluations are retained. The corresponding 780 soft-minus-hard contrasts contain no outcome-dependent exclusions. The six-family primary averages cover 120 contrasts per model under observed-NLL calibration; rank 32 remains secondary. Table 6 reports every calibrated family mean, including the separately designated rank-32 bridge. Figure 2 summarizes the separation of fitting, calibration and confirmation.

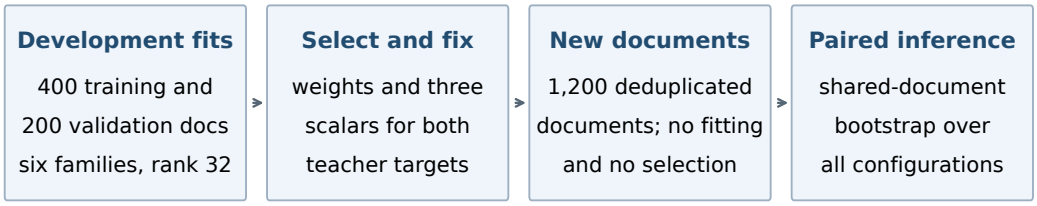


Figure 2: Qwen confirmation: separation of fitting, calibration and confirmation. Original training/validation data determine each readout and all scalar choices. All of them are fixed before new-document evaluation; no confirmation outcome selects a family, target, rank or temperature.

Document separation. The source is the first shard of a public, shuffled 100B-token release of FineWeb-Edu (Penedo et al., 2024). The complete development pool contains exactly 1,000 unique normalized documents. Confirmation selection excludes this union, exact duplicates and tokenizer-prefix duplicates, and applies the fixed five-word Jaccard threshold of 0.8. An independent whole-document audit checks all 1,919,400 development–confirmation and confirmation–confirmation pairs; among 8,676 pairs with positive shingle overlap, the maximum Jaccard similarity is 0.13853. All 1,200 new documents have nonempty shingle sets. These are experimental-exposure checks, not a pretraining contamination claim. The 128-token prefix and full-text duplicate rules are both necessary because truncation can otherwise yield identical evaluation inputs from different documents.

Selection before confirmation. The 480 original selected weights are fixed before document selection. Rank 32 uses only the original 400 training and 200 validation documents per split; all ten rate histories, thirty epochs and 33 scalar candidates are retained. Its 40 selected weights and three selected scalars per weight are jointly fixed before evaluating any new-document outcome. No confirmation label, loss, accuracy, temperature sweep or family selection feeds back into fitting. These fixed readouts are then scored on the same intermediate states, with the same model, input data and scoring procedure for all 40 configurations.

Paired bootstrap. Let S_{cj}^{a} and n_j be the observed NLL sum and valid observed-token count for readout a in configuration c and document j. For document multiplicities $w_j^{(b)}$ in bootstrap draw b, the loss contrast is

$$\Delta_m^{(b)} = \frac{1}{120}\sum_{c=1}^{120} \frac{\sum_j w_j^{(b)}(S_{cj}^{\mathrm{soft}} - S_{cj}^{\mathrm{hard}})}{\sum_j w_j^{(b)} n_j}.$$

The same multiplicities apply to all configurations within the model. The reported endpoints are the 0.0125 and 0.9875 quantiles of 5,000 draws. Bonferroni adjustment yields nominal simultaneous 95% intervals for the two fixed model effects without requiring independence of the models. Resampling individual token positions, treating splits as independent corpora or treating 240 configurations as independent observations would answer a different question.

Endpoint completeness. Qwen3-8B and Qwen3.5-9B have 152,907 and 152,956 teacher positions and 151,707 and 151,756 observed positions, respectively. All 1,560 selected positive-scalar evaluations preserve the argmax. Full teacher agreement, teacher cross-entropy, KL, entropy, maximum probability, observed accuracy, observed NLL and 15-bin expected calibration error (ECE) are

Table 6: Complete calibrated family means on confirmation documents: hard/soft observed NLL and soft-minus-hard teacher KL, written Δ KL, all in nats. Original families average 20 configurations per model; rank 32 averages ten early-depth configurations and is separated from the primary six-family mean.

Model	Family	Hard NLL	Soft NLL	Δ KL
Qwen3-8B	Scale	5.894	5.844	−0.062
	Rot. scale	5.699	5.640	−0.061
	Diagonal	5.275	5.171	−0.085
	Rot. diagonal	5.331	5.225	−0.089
	Rank 4	5.191	5.090	−0.065
	Full affine	4.842	4.518	−0.256
	Rank 32, layers 8/16	5.747	5.571	−0.100
Qwen3.5-9B	Scale	5.546	5.444	−0.101
	Rot. scale	5.309	5.208	−0.099
	Diagonal	4.987	4.794	−0.183
	Rot. diagonal	5.046	4.869	−0.168
	Rank 4	4.709	4.464	−0.236
	Full affine	4.399	3.871	−0.507
	Rank 32, layers 8/16	5.445	4.998	−0.440

Table 7: Capacity differences of soft fits on the confirmation documents after matched observed-NLL calibration. Full-affine minus restricted-family NLL in nats, averaged over five frozen fits at each depth. Negative favors full affine. Rank 32 was specified only at depths 8/16; dashes denote untested conditions. All individual contrasts remain in the evidence.

Model	Layer	Diagonal	Rank 4	Rank 32
Qwen3-8B	8	−0.990	−0.775	−0.350
	16	−0.881	−0.593	−0.286
	24	−0.590	−0.544	—
	31	−0.154	−0.377	—
Qwen3.5-9B	8	−1.423	−0.815	−0.291
	16	−1.365	−0.794	−0.283
	24	−0.812	−0.577	—
	31	−0.093	−0.186	—

retained for every scalar. The supplied tables also provide all depth and family strata. ECE uses the complete corpus sufficient sums, not an average of document-level calibration errors. The intervals in the main table are the prespecified analysis.

B Complete Llama/WikiText replication

Fixed experiment and exposure. The public corpus is the raw release of WikiText-103 (Merity et al., 2017), first training shard. The fixed article list has 400 fit, 200 calibration and 400 confirmation entries. Article identifiers, normalized titles/text and processed prefixes are disjoint across roles. Each confirmation article has 128 teacher positions and 127 observed successors, giving 51,200 and 50,800 positions respectively. Confirmation is not the official corpus test split. The Llama-3.1-8B-Instruct checkpoint and corpus were selected before outcomes. This is a joint model-family/corpus replication.

What changed between seeds. All three runs use exactly the same model, prompts, training and calibration states, teacher targets, 10 rates, 30 epochs, batch size 256 and Adam defaults. Each rate fit starts independently; hard and soft targets share the family/depth/seed-derived initialization and shuffle seed. Rate/epoch selection uses the corresponding teacher objective. Post-fit scalars are selected on the calibration articles, over the unchanged 33-point grid. Seed 42 fixes its 12 selected weights, covering three families, two depths and two targets, together with their histories

Table 8: Every predeclared Llama family–depth pair after observed-NLL calibration. Hard and soft NLLs are those of the first seed, 42; the last three columns retain all seed-specific soft-minus-hard effects. All rows use the same 400 articles.

Layer	Family	Hard NLL	Soft NLL	Δ_{first}	Δ_{second}	Δ_{third}
8	Diagonal	6.724	6.610	−0.115	−0.112	−0.109
8	Rank 32	5.442	5.259	−0.183	−0.255	−0.233
8	Full affine	5.194	4.742	−0.452	−0.440	−0.456
16	Diagonal	6.076	5.957	−0.118	−0.106	−0.103
16	Rank 32	4.844	4.557	−0.287	−0.261	−0.279
16	Full affine	4.556	4.116	−0.441	−0.434	−0.443

and selections before the confirmation states are extracted. The second and third training seeds then change only training randomness, fix their own 12 selected weights, histories and selections before scoring, and are scored on the same confirmation articles. All 360 candidate fits, 10,800 epoch records, 36 selected weights, 12 per seed, and 108 readout/scalar evaluations, three scalar rules per weight, are retained. The two additional seeds were fixed before their outcomes and are sensitivity analyses on exposed articles.

Targets. Native model logits supply complete-vocabulary targets at teacher temperature one, with a softmax over the full vocabulary. The selected weights are evaluated with their own fixed scalars.

Inference. The primary bootstrap jointly resamples the same 400 article indices for all six pairs, using 10,000 draws and a fixed generator seed listed in the supplementary README. The last two runs reuse those articles and are reported per seed rather than pooled; the three primary effects span 0.004628 nats. No selected rate or scalar reaches a grid boundary. Continuous scalar optimization and a common cross-objective checkpoint selector are examined on these articles in Appendices D and E. The supplementary material includes the complete reported statistics, training histories and selected configurations, and identifies the model and corpus used.

All six pairs and all scalar rules. Tables 9–11 report every paired comparison. Identity, teacher-KL and observed-NLL denote the three fixed scalar rules. NLL and teacher KL decrease in all 54 rows. Positive scalars produce zero recorded argmax changes. The native full teacher, identical across depths and seeds, has observed NLL 2.039372, accuracy 58.0945%, teacher agreement 100% and teacher KL zero. The saved sufficient statistics additionally retain each arm's accuracy, agreement, entropy, confidence, teacher cross-entropy and complete ECE bins.

B.1 Absolute prediction anchors

The following reference results use the same confirmation experiments. Qwen raw-lens rows are identity-scaled and use the same documents/positions as the fitted readouts; the five repeated fit-seed copies agree exactly. All 80 cell-level raw and full-teacher results are retained. For Llama, the table gives the Llama full-teacher reference; identity-scaled raw-lens anchors come from the matched and four-target diagnostics on the same 400 articles and are given in the paragraph after the table. Different tokenizers and corpora preclude interpreting absolute NLL differences across rows as model rankings.

Identity-scaled raw-lens anchors on the 400 WikiText articles give NLL 13.080072 and accuracy 0.151575% at layer 8, and NLL 11.391977 and accuracy 1.356299% at layer 16, each over 50,800 observed positions. Calibrated raw and bias-only comparisons are in Table 4.

Table 9: Complete Llama paired endpoints, training seed 42. NLL and teacher KL are nats/token. Δ is soft minus hard.

Layer	Family	Scalar	Hard NLL	Soft NLL	Δ NLL	Δ KL
8	Diagonal	Identity	6.724	6.610	−0.115	−0.156
8	Diagonal	Teacher-KL	6.746	6.610	−0.137	−0.145
8	Diagonal	Observed-NLL	6.724	6.610	−0.115	−0.156
8	Rank 32	Identity	5.983	5.429	−0.553	−0.702
8	Rank 32	Teacher-KL	5.442	5.259	−0.183	−0.284
8	Rank 32	Observed-NLL	5.442	5.259	−0.183	−0.284
8	Full affine	Identity	5.451	4.846	−0.605	−0.753
8	Full affine	Teacher-KL	5.194	4.742	−0.452	−0.577
8	Full affine	Observed-NLL	5.194	4.742	−0.452	−0.577
16	Diagonal	Identity	6.080	5.957	−0.122	−0.164
16	Diagonal	Teacher-KL	6.076	5.957	−0.118	−0.130
16	Diagonal	Observed-NLL	6.076	5.957	−0.118	−0.130
16	Rank 32	Identity	5.166	4.652	−0.514	−0.656
16	Rank 32	Teacher-KL	4.844	4.557	−0.287	−0.388
16	Rank 32	Observed-NLL	4.844	4.557	−0.287	−0.388
16	Full affine	Identity	4.782	4.166	−0.615	−0.753
16	Full affine	Teacher-KL	4.556	4.116	−0.441	−0.557
16	Full affine	Observed-NLL	4.556	4.116	−0.441	−0.557

Table 10: Complete Llama paired endpoints, second training seed. NLL and teacher KL are nats/token. Δ is soft minus hard.

Layer	Family	Scalar	Hard NLL	Soft NLL	Δ NLL	Δ KL
8	Diagonal	Identity	6.720	6.608	−0.112	−0.159
8	Diagonal	Teacher-KL	6.737	6.608	−0.129	−0.141
8	Diagonal	Observed-NLL	6.720	6.608	−0.112	−0.159
8	Rank 32	Identity	5.903	5.427	−0.477	−0.623
8	Rank 32	Teacher-KL	5.521	5.266	−0.255	−0.352
8	Rank 32	Observed-NLL	5.521	5.266	−0.255	−0.352
8	Full affine	Identity	5.437	4.845	−0.592	−0.738
8	Full affine	Teacher-KL	5.194	4.753	−0.440	−0.564
8	Full affine	Observed-NLL	5.194	4.753	−0.440	−0.564
16	Diagonal	Identity	6.086	5.962	−0.124	−0.168
16	Diagonal	Teacher-KL	6.068	5.962	−0.106	−0.120
16	Diagonal	Observed-NLL	6.068	5.962	−0.106	−0.120
16	Rank 32	Identity	5.139	4.663	−0.476	−0.653
16	Rank 32	Teacher-KL	4.841	4.581	−0.261	−0.388
16	Rank 32	Observed-NLL	4.841	4.581	−0.261	−0.388
16	Full affine	Identity	4.773	4.173	−0.600	−0.744
16	Full affine	Teacher-KL	4.557	4.123	−0.434	−0.556
16	Full affine	Observed-NLL	4.557	4.123	−0.434	−0.556

Table 11: Complete Llama paired endpoints, third training seed. NLL and teacher KL are nats/token. Δ is soft minus hard.

Layer	Family	Scalar	Hard NLL	Soft NLL	Δ NLL	Δ KL
8	Diagonal	Identity	6.720	6.611	−0.109	−0.155
8	Diagonal	Teacher-KL	6.753	6.611	−0.142	−0.154
8	Diagonal	Observed-NLL	6.720	6.611	−0.109	−0.155
8	Rank 32	Identity	5.918	5.411	−0.507	−0.632
8	Rank 32	Teacher-KL	5.498	5.265	−0.233	−0.314
8	Rank 32	Observed-NLL	5.498	5.265	−0.233	−0.314
8	Full affine	Identity	5.437	4.851	−0.587	−0.736
8	Full affine	Teacher-KL	5.194	4.739	−0.456	−0.581
8	Full affine	Observed-NLL	5.194	4.739	−0.456	−0.581
16	Diagonal	Identity	6.091	5.960	−0.131	−0.179
16	Diagonal	Teacher-KL	6.063	5.960	−0.103	−0.121
16	Diagonal	Observed-NLL	6.063	5.960	−0.103	−0.121
16	Rank 32	Identity	5.151	4.661	−0.490	−0.652
16	Rank 32	Teacher-KL	4.848	4.568	−0.279	−0.396
16	Rank 32	Observed-NLL	4.848	4.568	−0.279	−0.396
16	Full affine	Identity	4.779	4.163	−0.616	−0.760
16	Full affine	Teacher-KL	4.557	4.114	−0.443	−0.565
16	Full affine	Observed-NLL	4.557	4.114	−0.443	−0.565

Table 12: Rank-32 minus full-affine observed NLL after each fit receives its own observed-NLL-selected scalar. Positive values favor full affine; no rank sweep is implied.

Seed	Layer 8 hard	Layer 8 soft	Layer 16 hard	Layer 16 soft
First (42)	0.248	0.518	0.287	0.441
Second	0.327	0.513	0.284	0.458
Third	0.304	0.526	0.290	0.454

Table 13: Untrained raw-lens and full-teacher anchors. Accuracy and teacher agreement are percentages. These rows use identity scaling.

Model / corpus	Readout	NLL	Accuracy	Teacher agree.	Teacher KL
Qwen3-8B / FineWeb-Edu	Raw layer 8	12.557	0.694	0.812	10.670
	Raw layer 16	11.832	1.706	2.153	9.816
	Raw layer 24	9.886	7.045	9.673	7.555
	Raw layer 31	8.908	21.859	30.891	6.137
	Full teacher	2.600	47.312	100.000	0.000
Qwen3.5-9B / FineWeb-Edu	Raw layer 8	9.006	3.661	4.753	6.688
	Raw layer 16	8.754	4.951	6.975	6.422
	Raw layer 24	7.615	13.272	18.879	5.258
	Raw layer 31	3.207	43.941	68.975	0.765
	Full teacher	2.400	48.775	100.000	0.000
Llama-3.1-8B-Instruct / WikiText	Full teacher	2.039	58.094	100.000	0.000

C FITTED READOUTS RESCORED ON UNTOUCHED ARTICLES

Cohort and windows. The Llama/WikiText article selection is a fixed, output-independent pseudorandom ordering of the public training shard. Re-running it reproduces the 1,000 earlier articles exactly; continuing in the same order, and still excluding normalized duplicates and five-word-shingle Jaccard similarity of at least 0.8 against all earlier and newly passed articles, yields 700 candidates with at least 450 normalized words. The first 400 of them whose tokenization reaches 512 tokens form the cohort. No readout, temperature, checkpoint or constant predictor is refitted: every scored object was fixed on the earlier fitting, temperature and checkpoint-selection articles. The cohort was extracted after all four-target, bias-only, raw-lens and constant-predictor readouts had been fixed. The stratified-shuffle and extended-budget fits use only the fitting, scalar-fitting, checkpoint-selection and diagnostic roles of the earlier 1,000 articles; each was fixed before its untouched-article scoring. One forward pass per article supplies the teacher and the layer-8/16 states. The first window is the first 128 tokens, identical to the prefixes used elsewhere; the second is tokens 385–512 of the same pass. Each window has 51,200 teacher positions and 50,800 observed positions.

Inference. Per article, the four fit seeds are averaged before a 20,000-draw article bootstrap. Each family of four declared contrasts receives 98.75% percentile intervals; a family is either full minus argmax within one window, at two depths and two calibrated selectors, or policy minus cheap readout at two depths and two windows. The policy and the cheap comparator are those ranked first on checkpoint-selection NLL in Table 24; the policy is determined solely from validation scores on the earlier articles. Control contrasts, including the stratified shuffle, and the extended-budget rank-256 contrasts are rescored from the same fits and use the same draws.

Table 14: Untouched articles: observed NLL in nats of the full-target and argmax-target readouts, biased full affine before normalization with four fits averaged, and the full-target readout minus the argmax-, entropy-matched-, tail-permuted- and stratified-shuffle-target readouts in the columns −Argmax, −Ent. m., −Tail p. and −Shuf., respectively. Intervals are 98.75% paired-article intervals for full minus argmax; the control contrasts have intervals of half-width at most 0.030 nats.

Tokens	Layer/temp.	Full	Argmax	−Argmax	Interval	−Ent. m.	−Tail p.	−Shuf.
1–128	8 / global	4.646	5.302	−0.656	[−0.677, −0.634]	−0.811	−1.115	−0.502
1–128	8 / cond.	4.632	5.273	−0.640	[−0.662, −0.619]	−0.819	−1.119	−0.511
1–128	16 / global	4.044	4.639	−0.594	[−0.614, −0.574]	−0.729	−0.996	−0.438
1–128	16 / cond.	4.028	4.585	−0.557	[−0.576, −0.537]	−0.729	−0.992	−0.450
385–512	8 / global	5.015	5.737	−0.722	[−0.747, −0.698]	−0.930	−1.304	−0.564
385–512	8 / cond.	5.000	5.715	−0.716	[−0.740, −0.692]	−0.936	−1.308	−0.581
385–512	16 / global	4.264	4.919	−0.655	[−0.677, −0.633]	−0.836	−1.174	−0.504
385–512	16 / cond.	4.260	4.871	−0.612	[−0.632, −0.590]	−0.817	−1.152	−0.509

Table 15: Untouched articles: observed NLL in nats / accuracy in percent of the anchors, cheap readouts and full affine readouts, with conditional temperature where applicable; seeded rows average four fits. Constant predictors do not depend on the layer.

	Tokens 1–128		Tokens 385–512	
Readout	Layer 8	Layer 16	Layer 8	Layer 16
Native teacher, final layer	2.077 / 57.69	2.077 / 57.69	1.879 / 57.10	1.879 / 57.10
Raw lens, global temperature	11.691 / 0.15	10.980 / 1.41	11.658 / 0.24	10.893 / 1.73
Observed unigram, constant	7.706 / 4.62	7.706 / 4.62	7.912 / 4.49	7.912 / 4.49
Teacher marginal, constant	7.526 / 4.62	7.526 / 4.62	7.707 / 4.49	7.707 / 4.49
Cheap readout: bias only, full target	7.343 / 7.66	6.754 / 12.13	7.427 / 7.17	6.791 / 12.62
Policy: full affine, full target	4.632 / 27.57	4.028 / 33.18	5.000 / 23.63	4.260 / 30.22

The validation-selected policy minus the cheap readout is -2.711 $[-2.772, -2.652]$ and -2.428 $[-2.475, -2.381]$ nats at layer 8 in the first and second window, and -2.726 $[-2.785, -2.670]$ and -2.531 $[-2.580, -2.482]$ at layer 16, with accuracy gains of 19.92, 16.46, 21.05 and 17.60 points. The secondary policy, constrained to validation accuracy at least that of the cheap readout, selects

the same readout. On untouched articles the bias-only readout again improves on the constant teacher marginal by only 0.18 and 0.28 nats at layer 8 in the two windows, and the ordering of the raw, constant, bias-only and full affine readouts in Table 15 matches their ordering on the exposed evaluation articles of Table 24 in every window and depth. The extended-budget fits of Appendix D keep the same ordering on the untouched articles: rank 256, from either start, is 0.144–0.159 nats behind full affine in both windows and at both depths, with every 98.75% interval above 0.132.

D MATCHED NORMALIZATION, SELECTION AND CAPACITY CONTROLS

These diagnostics reuse the exposed 400 WikiText articles and one paired fit seed. The training/calibration/diagnostic split is 400/200/400; each article supplies its first 128 tokens, hence 50,800 observed-token positions in the diagnostic. The 200 calibration articles jointly determine checkpoints and post-fit scalars in this experiment. The following four-target experiment separates those two uses.

A common fitting comparison. At layers 8 and 16, a residual translator is inserted either before the frozen final normalization or after it. Each site fits biased full affine under hard and soft supervision, and soft rank-32 without/with bias and biased rank-128/256. All 24 training conditions receive ten learning rates, Adam without weight decay, batch size 256 and 60 epochs. Full affine starts at the identity; low-rank factors have independent 0.01-scale initialization with an optional zero bias. Initialization and minibatch order are paired across corresponding targets, sites and learning rates. All 240 trajectories complete their fixed budget.

The same trajectory supports maximum horizons of 30 or 60 epochs and two checkpoint rules: the fitted target objective, or observed NLL at identity temperature. A common rule need not choose the same learning rate or epoch for the two targets. Each selected checkpoint receives identity, the original 33-node scalar grid, or continuous observed-NLL calibration. For logits z and observed label y, the scalar objective is

$$\begin{aligned} g(\alpha) &= \mathbb{E}\Big[\log \sum_j \exp(\alpha z_j) - \alpha z_y\Big], \\ g''(\alpha) &= \mathbb{E}\big[\mathrm{Var}_{\mathrm{softmax}(\alpha z)}(z)\big] \geq 0. \end{aligned} \tag{3}$$

The second line makes g convex, so its minimizer is located by the sign of g'. Derivative bisection on $\alpha \in [1/16, 16]$ uses at most 40 iterations, mean derivative tolerance 10^{-7} and relative bracket tolerance 10^{-6}; endpoint derivatives identify boundary minima. Both grid and continuous objectives use the same log-sum-exp computation on identical logits. This optimizes the declared scalar family, not the affine fitting objective globally.

Across the full 48 target contrasts, which cross two depths, two sites, two horizons, two selectors and three scalar rules, soft NLL is lower throughout. The equal-depth, 60-epoch, continuously calibrated pre-minus-post soft NLL is -0.060672, with paired-article interval $[-0.066222, -0.055166]$. All intervals use 10,000 article resamples and are unadjusted comparisons for these fits.

Bias and larger ranks. Table 17 retains every soft family at both sites and depths under the common 60-epoch horizon and continuous scalar. The prescribed calibration-only capacity rule chooses the fewest parameters with NLL within 0.05 nats of the best full affine and accuracy within 0.005 in absolute probability. Across both depths and all three scalar rules, all six choices retain full affine itself. A selected-minus-full value of zero therefore compares identical predictors; it is not evidence of a cheaper equivalent model. Rank 256 narrows the loss gap without eliminating it.

Registered extended budget and identity-preserving start. A registered follow-up refits the soft target before normalization at layers 8 and 16 with one fit seed, twelve learning rates from 0.3 to 10^{-6}, 120 epochs and checkpoints every ten epochs: biased full affine from the identity, and biased rank 256 either from the original small random factors or with the output factor set to zero, so that the map starts at the identity (Pettyjohn et al., 2026). Selection uses global-calibrated checkpoint-selection NLL. Table 18 reports the diagnostic-article NLL; rescoring the same fits on the untouched articles of Appendix C gives rank-256-minus-full differences of 0.144–0.159 nats in both windows and at both depths.

Table 16: Matched full-affine comparison with common observed-NLL checkpoint selection and continuous scalar fitting. Differences are soft minus hard. Nominal 95% paired-article intervals are exploratory and condition on these fixed fits.

Layer/site	Max. epochs	Soft NLL	Hard NLL	ΔNLL [interval]
8/pre	30	4.663	5.280	-0.617 [-0.636, -0.597]
8/pre	60	4.612	5.280	-0.668 [-0.688, -0.647]
8/post	30	4.739	5.224	-0.485 [-0.501, -0.469]
8/post	60	4.695	5.224	-0.529 [-0.548, -0.511]
16/pre	30	4.055	4.640	-0.585 [-0.602, -0.568]
16/pre	60	4.028	4.640	-0.612 [-0.630, -0.594]
16/post	30	4.104	4.573	-0.470 [-0.484, -0.455]
16/post	60	4.067	4.576	-0.510 [-0.527, -0.493]

Table 17: Complete soft capacity comparison, with common observed-NLL checkpoint selection, 60-epoch maximum and continuous temperature. Entries are observed NLL in nats / accuracy in percent. Parameter counts: full 16,781,312; rank 32 without bias 262,144; with bias 266,240; rank 128 1,052,672; rank 256 2,101,248.

Family	L8/pre	L8/post	L16/pre	L16/post
Full affine	4.612 / 28.35	4.695 / 28.58	4.028 / 33.52	4.067 / 33.52
Rank 32, no bias	4.976 / 26.92	5.245 / 25.15	4.417 / 31.80	4.556 / 30.53
Rank 32, bias	4.963 / 26.55	5.244 / 25.12	4.417 / 31.77	4.556 / 30.54
Rank 128, bias	4.882 / 28.21	5.106 / 26.70	4.260 / 33.10	4.424 / 31.60
Rank 256, bias	4.746 / 28.29	5.048 / 26.98	4.174 / 33.36	4.354 / 32.44

Readout cost. Every map feeds the same frozen head, whose $Vd = 5.25 \times 10^8$ multiply-adds per position, with $V = 128{,}256$ and $d = 4096$, dominate the readout. Full affine adds $d^2 \approx 1.68 \times 10^7$ and rank 256 adds $2rd \approx 2.10 \times 10^6$, so per-position readout compute differs by 2.7% of the full-affine count although the adapter parameters differ eightfold, as Table 19 shows. The capacity comparison therefore concerns observed NLL under a fixed training recipe; the table gives arithmetic counts.

Scope. The evidence includes every candidate history, selected calibration, article-level endpoint and comparison. The data are exposed diagnostics.

Table 18: Registered extended-budget soft fits with one seed on 400 diagnostic articles: NLL and rank-256 minus full-affine NLL with 99.17% intervals.

Layer	Full affine NLL	Rank 256, random start − full	Rank 256, zero start − full
8	4.597	$0.145\ [0.132, 0.158]$	$0.155\ [0.141, 0.169]$
16	4.024	$0.145\ [0.132, 0.158]$	$0.144\ [0.131, 0.157]$

Table 19: Readout cost by arithmetic count for Llama-3.1-8B-Instruct with $d = 4096$ and $V = 128{,}256$: adapter parameters and multiply-adds per position including the shared frozen head.

Readout	Adapter parameters	Multiply-adds per position	Relative to full affine
Full affine with bias	16,781,312	5.421×10^8	1.000
Rank 256 with bias	2,101,248	5.274×10^8	0.973
Frozen head alone	0	5.253×10^8	0.969

E CLASS CORRESPONDENCE VERSUS ENTROPY AND PROBABILITY MULTISET

Llama states at layers 8 and 16 are separately translated before final normalization by a biased full affine map. Five teacher-only target distributions are compared. For vocabulary size V, teacher probabilities p and a maximum-probability class j^*, the entropy-matched control is

$$\begin{aligned} q_{j^*} &= 1 - \epsilon, \\ q_j &= \frac{\epsilon}{V-1} \qquad (j \neq j^*), \\ H(q) &= H(p), \qquad 0 \leq \epsilon \leq 1 - 1/V. \end{aligned} \tag{4}$$

Bisection determines ϵ with an entropy residual below 10^{-8}. Full supervision retains p; hard supervision uses its argmax. The probability-permutation control independently permutes probabilities strictly below the maximum at each document position, with a deterministic position/fit-seed-specific permutation. Every maximum tie is fixed. This preserves the full probability multiset and entropy while disrupting correspondence between nonmaximum probabilities and vocabulary classes. It is the DKPP control of Furlanello et al. (2018); it also changes class-marginal frequencies and cannot isolate correspondence from that change.

Registered stratified whole-distribution shuffle. The fifth target reassigns complete teacher distributions among fitting positions. Positions are grouped by teacher argmax, sorted by teacher entropy within each class, and cut into consecutive blocks of eight, where a short trailing block joins the preceding one and a class smaller than eight forms one block; a seeded derangement within each block of at least two positions gives every member another member's full teacher distribution. Of the 51,200 fitting positions, 47,239, or 92.3%, are assigned another position's distribution and 3,961 in single-member classes keep their own. Every target keeps its position's argmax and is a genuine teacher softmax that the frozen normalization and head produce exactly; the multiset of target rows, and therefore the class marginal, is unchanged up to an L1 difference of 4.6×10^{-16}. Only the pairing between states and teacher distributions changes. Entropy is ordered, not matched numerically: on a fixed 2,048-position sample, the absolute difference between a position's own and its assigned teacher entropy has median 0.030, mean 0.499, 90th percentile 1.72 and maximum 8.30 nats, and exceeds one nat at 339 positions. The target, its four fit seeds, learning rates, epochs, checkpoints and selectors are those of the other arms; the decision rule and endpoints were registered before fitting.

Registered frozen-head attainability diagnostic. The same registration fixed a diagnostic of how closely each target can be produced by the frozen head, on a fixed keyed sample of 2,048 fitting positions; seeded targets use the first of the four fit seeds. For a target q with vocabulary-centered log-probabilities ℓ, it reports the fraction of $\|\ell\|^2$ outside the span of the centered head, the KL at the least-squares projection of ℓ onto that span, the KL of $\mathrm{softmax}(W_{\mathrm{lm}}u)$ reached by limited-memory Broyden–Fletcher–Goldfarb–Shanno optimization, L-BFGS, over an unconstrained $u \in \mathbb{R}^d$ started at that projection, and the KL reached by L-BFGS over a free input passed through the frozen final

normalization and head. The last two are achieved values after a finite budget of 150 iterations, hence upper bounds on the attainable minimum. Table 20 gives the means. The full and stratified-shuffle targets are genuine teacher distributions and are reproduced to within 4×10^{-5} nats at every sampled position, with means below 5×10^{-6}, so the cost of the stratified shuffle does not come from targets the frozen head cannot produce; it arises from fitting one shared map to the reassigned pairs. The entropy-matched and DKPP targets lie mostly outside the head's span in log space, with 96–97% of the squared centered norm outside it, and the optimization through the frozen normalization stops at 0.15 and 0.21 nats.

Table 20: Frozen-head attainability of the targets as means over 2,048 fitting positions: residual fraction outside the head's span, KL at the log-space least-squares projection, KL reached over unconstrained head inputs and through the frozen normalization and head, both achieved values and hence upper bounds on the minimum, and total variation to the position's own teacher distribution.

Target	Residual	KL, projection	KL, unconstrained	KL, through norm.	TV to own
Full distribution	3.3×10^{-6}	2.0×10^{-4}	4.5×10^{-6}	1.7×10^{-6}	1.1×10^{-7}
Stratified shuffle	3.3×10^{-6}	1.9×10^{-4}	4.0×10^{-6}	1.7×10^{-6}	0.337
Entropy matched	0.961	9.55	0.0041	0.154	0.365
DKPP tail permutation	0.968	9.44	0.157	0.213	0.363

Disjoint scalar fitting and selection. The four original targets share four paired fit seeds, ten learning rates, Adam, batch size 256 and 60 epochs. Checkpoints are evaluated at epochs 5, 10, …, 60, giving 160 trajectories, 1,920 checkpoint evaluations and 9,600 training epochs per depth; the shuffle arm adds 40 trajectories per depth with the same seeds, rates and checkpoints. Both depths completed these budgets. Initialization and minibatch permutations agree across corresponding arms and rates. Of the 200 calibration articles, 100 fit continuous inverse temperatures and 100 select checkpoints; their identities are disjoint from the 400 training and 400 exposed diagnostic articles.

Three selectors are retained: identity-NLL checkpoint selection followed by a global temperature; global-calibrated NLL checkpoint selection; and confidence-conditional calibrated NLL checkpoint selection. The latter two fit temperatures separately for every candidate before selecting its rate/epoch on the other 100 articles. The conditional family has nine fixed bins of each uncalibrated student's maximum probability, with boundaries 0, 0.05, 0.1, 0.2, 0.4, 0.6, 0.8, 0.9, 0.95 and 1. Group assignment uses no observed label or teacher output. Each bin fits its own positive inverse temperature in $[1/16, 16]$; empty bins use one. Ties select the first listed rate then earliest epoch. This tests one finite conditional family, not all input-dependent calibration functions.

At each depth, all 48 selected four-target variants and 12 shuffle variants are scored on the same 400 articles, with 50,800 observed positions and 51,200 teacher positions per variant. The four declared comparisons are soft minus entropy-matched or permutation, under global or conditional calibrated selection. Each article's loss difference is first averaged over the four fits. A paired 400-article bootstrap uses 20,000 draws; 98.75% intervals apply Bonferroni adjustment to the four comparisons within each depth; they are not simultaneous intervals for the combined eight-comparison family. The article is the resampling unit; Appendix F adds fitting-seed uncertainty.

At layer 8, each of the four fitted seeds gives a negative NLL contrast for every declared comparison. Table 22 shows that the full-distribution target improves observed NLL over every control target. The result supports the importance of teacher probability assignment within this fitting design. For the stratified shuffle, every one of the 16 seed-level contrasts is negative, ranging from -0.422 to -0.530 nats, and the shuffle retains 19.0–25.7% of the soft-over-hard gain, measured as

$$\frac{\mathrm{NLL}_{\mathrm{argmax}} - \mathrm{NLL}_{\mathrm{shuffle}}}{\mathrm{NLL}_{\mathrm{argmax}} - \mathrm{NLL}_{\mathrm{full}}}.$$

Appendix C repeats every contrast with the same fits on untouched articles. Complete article statistics, all three selectors and all candidate choices accompany the scientific supplement.

Second-depth replication. The layer-16 experiment holds fixed the four targets, split, paired fit seeds, optimizer, checkpoint budget and calibration families. Its complete outcomes are in Table 23,

Table 21: Layer-8 calibrated observed NLL of every target, averaged over the four paired fits. These Llama results are distinct from the Qwen confirmation.

Target	Global NLL	Conditional NLL
Full distribution	4.6102	4.5933
Hard argmax	5.2570	5.2310
Entropy matched	5.4114	5.4062
Tail permutation	5.7167	5.7083
Stratified shuffle	5.1116	5.1067

Table 22: Full-distribution target minus each control target on the exposed Llama/WikiText articles, with biased full affine before normalization, under global and nine-bin conditional calibration. Intervals are 98.75% paired-article intervals over 400 articles, with four fit seeds averaged within article, Bonferroni-adjusted within each registered family of four contrasts.

		Global calibration		Conditional calibration	
Layer	Control	ΔNLL	Interval	ΔNLL	Interval
8	Entropy matched	-0.801	$[-0.827, -0.775]$	-0.813	$[-0.839, -0.787]$
8	Tail permutation	-1.106	$[-1.137, -1.076]$	-1.115	$[-1.145, -1.085]$
16	Entropy matched	-0.715	$[-0.738, -0.690]$	-0.715	$[-0.739, -0.691]$
16	Tail permutation	-0.983	$[-1.011, -0.954]$	-0.980	$[-1.007, -0.951]$
8	Stratified shuffle	-0.501	$[-0.522, -0.482]$	-0.513	$[-0.533, -0.494]$
16	Stratified shuffle	-0.432	$[-0.451, -0.414]$	-0.443	$[-0.461, -0.426]$

and its four contrasts with their intervals are the layer-16 rows of Table 22. This repeats the observed NLL pattern at a second depth without treating shared articles as a second independent population. Four fits are averaged within each article before the paired bootstrap. All 48 four-target and 12 shuffle variants produce all 400 article endpoints, and all are reported. At layer 16, full-target fitting has lower observed NLL than every target control under each calibrated selector. Zero-parameter and bias-only references with the same roles are reported in Table 4.

Selected configurations. All 96 selected four-target records and the 24 stratified-shuffle selections, with their learning rates and epochs, accompany the evidence. No stratified-shuffle selection lies at a learning-rate grid edge or at the last checkpoint.

Constant predictors and a validation-selected policy. Two constant predictors use no hidden state: the mean teacher distribution over the 51,200 fitting positions, and the observed unigram distribution of the fitting articles with add-0.5 smoothing. Each receives a temperature fitted on the scalar-fitting articles and is scored by the same procedure on the 400 evaluation articles. Paired 20,000-draw article bootstraps with Bonferroni adjustment over eight contrasts, which gives 99.4% intervals, yield, relative to the teacher marginal: calibrated raw lens $+4.17$ $[4.11, 4.24]$ and $+3.48$ $[3.40, 3.55]$ nats at layers 8 and 16; full-target bias-only -0.18 $[-0.20, -0.16]$ and -0.78 $[-0.81, -0.74]$; full-target full affine -2.91 $[-2.97, -2.85]$ and -3.49 $[-3.54, -3.43]$ with global temperature, and -2.93 $[-2.99, -2.87]$ and -3.50 $[-3.55, -3.45]$ with conditional temperature. Table 24 lists nine representative readouts with their seed-mean NLL on the 100 checkpoint-selection articles. The policy pool is every readout fixed before the untouched cohort was extracted: the raw lens, bias-only and four-target full affine readouts under all selectors, and the constant predictors. The registration also listed the stratified-shuffle and extended-budget fits for this pool; they were completed after the cohort extraction and are reported separately in Appendices C and D. Validation NLL alone selects the full-target affine readout with conditional temperature as the policy and the full-target bias-only map as the best cheap alternative, at both depths; a secondary policy constrained to validation accuracy at least that of the cheap readout selects the same readout. At layer 16 the global-temperature full-target bias-only readout has slightly lower evaluation NLL than the validation-selected conditional one, 6.742 versus 6.748. Because the 100 checkpoint-selection articles also selected each fitted readout's checkpoint, the evaluation column is the comparison of record.

Table 23: Layer-16 calibrated observed NLL of every target, averaged over four fitted seeds.

Target	Global NLL	Conditional NLL
Full distribution	4.0338	4.0185
Hard argmax	4.6155	4.5656
Entropy matched	4.7484	4.7337
Tail permutation	5.0164	4.9981
Stratified shuffle	4.4661	4.4619

Table 24: Validation-selected readout policy on Llama/WikiText, with conditional temperature where applicable; seeded rows average four fits. Val.: 100 checkpoint-selection articles; Eval.: 400 evaluation articles. NLL in nats.

		Layer 8		Layer 16	
Readout, target	Parameters	Val. NLL	Eval. NLL	Val. NLL	Eval. NLL
Full affine, full distribution	$d^2 + d$	4.703	4.593	4.109	4.019
Full affine, argmax	$d^2 + d$	5.340	5.231	4.647	4.566
Full affine, entropy matched	$d^2 + d$	5.523	5.406	4.821	4.734
Full affine, tail permutation	$d^2 + d$	5.832	5.708	5.092	4.998
Bias only, full distribution	d	7.417	7.335	6.824	6.748
Bias only, argmax	d	7.465	7.385	6.860	6.774
Teacher marginal, constant	V	7.611	7.519	7.611	7.519
Observed unigram, constant	V	7.799	7.696	7.799	7.696
Raw lens	9	11.686	11.692	10.975	10.996

The complete list of registered contrasts is in Appendix M.

F UNCERTAINTY OVER FITTING SEEDS

The original article/document-bootstrap intervals condition on the fitted readouts. This appendix adds fitting-seed variation to them; Sections 4.1 and 4.6 quote the resulting intervals. This post hoc analysis supplements the registered analyses. It fits nothing new: it re-uses the released article and document statistics, and every point estimate equals the corresponding estimate of the original analyses.

Procedure. For Llama/WikiText, the four paired fit seeds of the four-target and stratified-shuffle experiments share one initial state: the full affine translator starts at the identity, with zero residual weights and bias. The seeds differ in minibatch order and, for control targets that are built with random draws, in that random construction. Each two-stage draw resamples the four seeds with replacement, averages the drawn seeds within article, and then resamples the 400 articles, with 20,000 draws, a fixed generator seed and 98.75% percentile intervals as in the registered families. For Qwen, each of the five seeds defines its own 400/200 training/validation partition and its own training randomness. Each draw resamples the five seeds, which all families and depths share, and the 1,200 confirmation documents, with 5,000 draws, a fixed generator seed and 95% intervals; both generator seeds are listed in the supplementary README. A bootstrap over four or five seeds describes variation among the observed seeds, so the tables also give the standard deviation of the seed-level estimates.

Results. In Table 25, every two-stage interval excludes zero, on the exposed articles and in both untouched-article windows. Each of the 192 seed-level estimates, from 48 contrasts each computed for the same four fitting seeds, is negative. The largest seed-level standard deviation is 0.0223 nats, for the stratified shuffle at layer 8 in the later untouched window, tokens 385–512, under conditional temperature. Compared with the article-only bootstrap interval computed alongside it in the same analysis, adding seed resampling widens an interval by at most 0.0220 nats. In Table 26, the two-stage interval excludes zero for every family on both models, and all five seed-level estimates are negative in every family; these family intervals are pointwise, not simultaneous across families. Over the five seed-level estimates, the t-based 95% intervals for the primary effects, conditional on the evaluation documents, are $[-0.126, -0.122]$ and $[-0.226, -0.222]$. The supplementary statistics reproduce both tables; the Qwen estimates weight each model by its own observed-token counts per document.

Table 25: Llama-3.1-8B-Instruct/WikiText, full target minus each control target with biased full affine before normalization, with fitting-seed uncertainty. Δ: point estimate of the original analysis in nats; SD: standard deviation of the four seed-level estimates; Ent. m.: entropy matched; Tail perm.: tail permutation; two-stage: 98.75% interval resampling fit seeds and articles. Exposed: 400 evaluation articles; W1/W2: tokens 1–128 and 385–512 of the 400 untouched articles, whose point estimates are in Table 14.

		Exposed articles			Untouched, W1	Untouched, W2
Layer / temp.	Control	Δ	SD	Two-stage	Two-stage	Two-stage
8 / global	Argmax	-0.647	0.009	$[-0.672, -0.622]$	$[-0.680, -0.631]$	$[-0.748, -0.695]$
8 / global	Ent. m.	-0.801	0.006	$[-0.829, -0.774]$	$[-0.837, -0.786]$	$[-0.957, -0.903]$
8 / global	Tail perm.	-1.106	0.005	$[-1.136, -1.076]$	$[-1.143, -1.087]$	$[-1.341, -1.270]$
8 / global	Strat. shuffle	-0.501	0.013	$[-0.527, -0.477]$	$[-0.526, -0.477]$	$[-0.595, -0.534]$
8 / cond.	Argmax	-0.638	0.010	$[-0.664, -0.612]$	$[-0.669, -0.611]$	$[-0.740, -0.691]$
8 / cond.	Ent. m.	-0.813	0.007	$[-0.840, -0.786]$	$[-0.847, -0.791]$	$[-0.963, -0.910]$
8 / cond.	Tail perm.	-1.115	0.004	$[-1.146, -1.084]$	$[-1.149, -1.088]$	$[-1.343, -1.274]$
8 / cond.	Strat. shuffle	-0.513	0.014	$[-0.539, -0.489]$	$[-0.539, -0.479]$	$[-0.612, -0.550]$
16 / global	Argmax	-0.582	0.006	$[-0.604, -0.560]$	$[-0.616, -0.573]$	$[-0.680, -0.630]$
16 / global	Ent. m.	-0.715	0.004	$[-0.739, -0.690]$	$[-0.753, -0.706]$	$[-0.864, -0.808]$
16 / global	Tail perm.	-0.983	0.007	$[-1.012, -0.952]$	$[-1.023, -0.968]$	$[-1.205, -1.143]$
16 / global	Strat. shuffle	-0.432	0.008	$[-0.453, -0.412]$	$[-0.456, -0.419]$	$[-0.525, -0.484]$
16 / cond.	Argmax	-0.547	0.006	$[-0.569, -0.525]$	$[-0.578, -0.536]$	$[-0.635, -0.587]$
16 / cond.	Ent. m.	-0.715	0.005	$[-0.740, -0.690]$	$[-0.753, -0.705]$	$[-0.846, -0.789]$
16 / cond.	Tail perm.	-0.980	0.008	$[-1.008, -0.950]$	$[-1.019, -0.965]$	$[-1.181, -1.122]$
16 / cond.	Strat. shuffle	-0.443	0.008	$[-0.463, -0.423]$	$[-0.469, -0.431]$	$[-0.529, -0.489]$

Table 26: Qwen confirmation, soft minus hard observed NLL by readout family under observed-NLL scalars, with 20 configurations per family and 120 in the primary mean, and with fitting-seed uncertainty. Doc.: 95% interval resampling documents with the fits fixed; two-stage: 95% interval resampling the five fitting seeds and the documents; SD: standard deviation of the five seed-level estimates; last column: negative seed-level estimates.

Model	Family	Δ	Doc.	Two-stage	SD	< 0
Qwen3-8B	Scale	-0.050	$[-0.052, -0.047]$	$[-0.052, -0.047]$	0.0016	5/5
	Rot. scale	-0.059	$[-0.061, -0.057]$	$[-0.062, -0.056]$	0.0014	5/5
	Diagonal	-0.104	$[-0.107, -0.101]$	$[-0.108, -0.101]$	0.0017	5/5
	Rot. diagonal	-0.106	$[-0.109, -0.103]$	$[-0.110, -0.102]$	0.0032	5/5
	Rank 4	-0.101	$[-0.104, -0.098]$	$[-0.107, -0.096]$	0.0054	5/5
	Full affine	-0.324	$[-0.329, -0.319]$	$[-0.331, -0.318]$	0.0056	5/5
	Primary, all six	-0.124	$[-0.126, -0.122]$	$[-0.126, -0.122]$	0.0017	5/5
Qwen3.5-9B	Scale	-0.102	$[-0.104, -0.099]$	$[-0.104, -0.099]$	0.0012	5/5
	Rot. scale	-0.101	$[-0.103, -0.099]$	$[-0.103, -0.099]$	0.0005	5/5
	Diagonal	-0.193	$[-0.196, -0.189]$	$[-0.197, -0.188]$	0.0024	5/5
	Rot. diagonal	-0.177	$[-0.181, -0.174]$	$[-0.181, -0.173]$	0.0018	5/5
	Rank 4	-0.245	$[-0.249, -0.241]$	$[-0.256, -0.235]$	0.0127	5/5
	Full affine	-0.528	$[-0.534, -0.522]$	$[-0.535, -0.521]$	0.0048	5/5
	Primary, all six	-0.224	$[-0.227, -0.222]$	$[-0.227, -0.221]$	0.0014	5/5

G A TEACHER-PROXY ACCOUNT OF THE GAP SIZE

This post hoc re-analysis supplements the registered analyses. It fits and rescales no readout: it uses the released Qwen document statistics and Llama article statistics from which the main-text contrasts are computed, and compares every pair of readouts on the same documents or articles.

Account. For readouts A and B and teacher p_L, let $\Delta\mathrm{NLL}$ be the difference of their observed-token NLL and

$$\Delta\mathrm{KL} = \mathrm{KL}(p_L \,\|\, A) - \mathrm{KL}(p_L \,\|\, B)$$

the difference of their teacher KL, each a token-weighted mean over the positions of its endpoint.

Motivation. On a common set of positions with common weights, for fixed readouts,

$$\mathbb{E}_{y_i \sim p_{L,i}}[\Delta\mathrm{NLL}] = \Delta\mathrm{KL},$$

because both sides average $\log B - \log A$, once over the observed token and once over a draw from p_L.

Tested relation. The two endpoints use different masks, since observed NLL omits the final position of each document, which teacher KL includes as specified in Section 3, and the observed tokens come from the text; we therefore test the proportionality empirically, as the approximation

$$\Delta\mathrm{NLL} \approx b\,\Delta\mathrm{KL},$$

whose residual contains the mask difference, the teacher–data discrepancy and sampling variation. b is one coefficient shared by the specified cells of a setting and evaluation window, fitted by least squares through the origin; $b = 1$ is the zero-parameter version. The predictor $b\,\Delta\mathrm{KL}$ needs no observed tokens once the readouts and b are given, but estimating b uses observed-NLL differences. A cell is one pair of calibrated readouts: for Qwen, soft minus hard in one model–depth–family–seed configuration under the primary observed-NLL scalar, giving 260 cells including rank 32; for Llama, each of the four control targets, namely argmax, entropy-matched, tail-permuted and stratified shuffle, minus the full target, at two depths, four fit seeds and the two calibrated temperature rules, giving 64 cells per window.

Held-out prediction. Each held-out test refits b without the held-out cells and predicts them. These are retrospective cell holdouts: training and held-out cells share the evaluation documents or articles and, on Llama, the full-target reference readouts; the prospective test with frozen coefficients is in

Appendix H. On Qwen the holdouts use cells averaged over the five seeds and one coefficient pooled over both checkpoints; the per-model slopes in Table 27 are separate descriptive fits. Leaving out one depth gives MAE of 0.017–0.036 nats, leaving out one family 0.021–0.044 and leaving out one model 0.028 and 0.046. On Llama, leaving out one control target gives MAE 0.002–0.036 nats across the three windows; for the held-out stratified shuffle the MAE is 0.009–0.013 and the mean gap is predicted within 0.004–0.013 nats. With b fitted on all cells of a window, the Llama MAE is 0.005–0.016 nats. In all eight Qwen model–depth cells, full affine has both the largest teacher-KL reduction and the largest observed-NLL reduction among the six original families. Document-bootstrap intervals for b, with 2,000 draws and a fixed generator seed listed in the supplementary README, resample the 1,200 documents jointly for all cells of a model. These intervals are conditional on the fitted readouts, with the fitting seeds held fixed. The coefficient is estimated and applied within each setting.

Reproduction. The supplementary statistics reproduce Table 27 and every held-out prediction.

Table 27: Teacher-proxy account: observed-NLL differences against teacher-KL differences between calibrated readouts. b: least-squares slope through the origin, with 95% document-bootstrap interval for the Qwen models; Id. MAE: mean absolute error of $\Delta\mathrm{NLL} = \Delta\mathrm{KL}$ in nats; sign: cells in which $\Delta\mathrm{KL}$ and $\Delta\mathrm{NLL}$ agree in sign. Llama evaluations: exposed or untouched articles, token window.

Setting	Contrast	Evaluation, tokens	Cells	Corr.	b	Id. MAE	Sign
Qwen3.5-9B	soft − hard	1,200 documents	130	0.998	1.036 $[1.028, 1.044]$	0.010	130/130
Qwen3-8B	soft − hard	1,200 documents	130	0.840	1.145 $[1.133, 1.157]$	0.047	124/130
Qwen, both	soft − hard	1,200 documents	260	0.965	1.055	0.029	254/260
Llama	control − full	exposed, 1–128	64	0.997	0.854	0.126	64/64
Llama	control − full	untouched, 1–128	64	0.998	0.858	0.123	64/64
Llama	control − full	untouched, 385–512	64	1.000	0.983	0.015	64/64

H CONFIDENCE-PRESERVING TAIL CONTROLS

The stratified shuffle of Appendix E keeps each target's argmax but not its own confidence or entropy. Two registered targets hold both the argmax and the argmax probability of every fitting position exactly and change only the rest of the distribution. For fitting position i with teacher distribution p_i, let

$$a_i = \arg\max_k p_i[k], \qquad m_i = p_i[a_i],$$
$$t_i[k] = \frac{p_i[k]}{1 - m_i} \qquad (k \neq a_i), \tag{5}$$

so that t_i is the normalized tail. Both targets set $q_i[a_i] = m_i$. The *tail swap* sets

$$q_i[k] = (1 - m_i)\, t_{d(i)}[k] \qquad (k \neq a_i), \tag{6}$$

where the donor $d(i)$ is another fitting position with the same argmax and adjacent tail entropy: each argmax class is sorted by the entropy of t, cut into consecutive blocks of eight, where a short trailing block joins the preceding one, and a seeded derangement is applied within each block; single-member classes keep their own tail. The entropy of the target then changes by $(1 - m_i)\,(H(t_{d(i)}) - H(t_i))$ only. The *class-mean tail* replaces $t_{d(i)}$ by the mean tail T_{a_i} of all fitting positions with argmax a_i. If a new tail would give some other class a probability of at least m_i, the argmax guard keeps the position's own teacher distribution. At every other position the class-mean target is therefore a deterministic function of the argmax and its probability; at the guarded positions it still carries the position's own tail. Everything else is identical to the other targets: the biased full affine map before normalization, layers 8 and 16, the four paired fit seeds, ten learning rates, 60 epochs, twelve checkpoints, roles and selectors. The primary family is full target minus tail swap at both depths and both calibrated selectors, four contrasts with 98.75% intervals from 20,000 article draws; the class-mean tail forms a second family of four. Both families were registered before any of these fits existed; every selected readout was then rescored without refitting on the untouched articles.

Targets. Per fit seed, the tail swap changes 42,634–42,685 of the 51,200 fitting positions; 4,556–4,607 keep their own tail under the argmax guard and 3,961 are single-member classes. The absolute entropy change has mean 0.114–0.116, median 0.0036–0.0037, 90th percentile 0.36–0.37 nats, and exceeds one nat at 1,413–1,466 positions. The class-mean tail changes 45,416 positions; 1,825, or 3.6%, keep their own distribution under the guard and the 3,961 positions of single-member classes keep their own tail. It raises mean target entropy from 1.769 to 2.342 nats, with an absolute change of mean 0.603 and median 0.359. Averaged over positions, the class marginal of the target moves by an L_1 distance of 0.040 under the tail swap and 0.039 under the class-mean tail; the stratified shuffle leaves it unchanged but does not keep each position's confidence. Each control thus holds some properties of the full target fixed and changes others, so the fractions below describe the effect of each replacement under this procedure, not the share of the gain carried by one property of the target. Of the 32 selected readouts, none selects an edge learning rate and 7 select the last checkpoint.

Table 28: Registered confidence-preserving tail controls on the 400 diagnostic articles: full-target minus control observed NLL in nats, 98.75% article intervals and the fraction $(\mathrm{argmax} - \mathrm{control})/(\mathrm{argmax} - \mathrm{full})$ of the full target's gain that the control retains.

	Full − tail swap, primary			Full − class-mean tail		
Layer / selector	ΔNLL	Interval	Kept	ΔNLL	Interval	Kept
8 / global	-0.437	$[-0.454, -0.421]$	0.32	-0.305	$[-0.318, -0.291]$	0.53
8 / conditional	-0.454	$[-0.469, -0.439]$	0.29	-0.313	$[-0.326, -0.300]$	0.51
16 / global	-0.354	$[-0.370, -0.337]$	0.39	-0.306	$[-0.321, -0.292]$	0.47
16 / conditional	-0.365	$[-0.381, -0.349]$	0.33	-0.316	$[-0.330, -0.301]$	0.42

Results. All eight contrasts are negative with intervals excluding zero. Intervals that also resample the four fit seeds with the procedure of Appendix F widen by at most 0.009 nats and exclude zero for all eight, and each of the 32 seed-level estimates is negative. The tail swap retains 0.29–0.39 of the full target's gain over the argmax target and the class-mean tail 0.42–0.53. Rescored without refitting on the untouched articles in Table 29, all 16 contrasts over two windows, depths, selectors and targets remain negative with 98.75% intervals excluding zero, and so do all 64 seed-level estimates.

Table 29: Tail controls rescored on the 400 untouched articles without refitting: full-target minus control observed NLL in nats, four fits averaged within article; every 98.75% article interval excludes zero.

		Full − tail swap		Full − class-mean tail	
Tokens	Layer	Global	Conditional	Global	Conditional
1–128	8	-0.439	-0.451	-0.301	-0.305
1–128	16	-0.361	-0.373	-0.313	-0.322
385–512	8	-0.492	-0.497	-0.337	-0.339
385–512	16	-0.418	-0.423	-0.352	-0.360

Fitting difficulty. The training record of every epoch holds the running mean of the target cross-entropy over the 51,200 fitting positions, accumulated while the map is being updated. This running loss minus the mean target entropy is an in-training excess KL, not the KL of a fixed readout. In the epoch of the checkpoint selected under global temperature it is 0.28 and 0.20 nats for the full target at layers 8 and 16, 1.06 and 1.10 for the tail swap, 0.42 and 0.54 for the class-mean tail and 1.51 and 1.40 for the stratified shuffle. In the last epoch of the learning rate with the lowest final value it is 0.26 and 0.18, 0.49 and 0.44, 0.40 and 0.34, and 0.54 and 0.49. Every control is thus fitted less closely than the full target during training, and the observed-NLL differences above include this difference in fit.

A prospective teacher-proxy test. Before any tail-control fit existed, we fixed the coefficients b of Appendix G from the earlier Llama control targets and registered the prediction $\Delta\mathrm{NLL} = b\,\Delta\mathrm{KL}$

for every cell of both new targets over two depths, four fit seeds and two calibrated selectors, with pass criteria on sign, a mean absolute error of at most 0.04 nats, mean gap and the ratio of means. Table 30 reports all six registered tests; all pass.

Table 30: Registered prospective test of the teacher-proxy account on the 16 cells of each tail control and window: frozen coefficient b, mean observed and predicted control-minus-full NLL gap, mean and maximum absolute error in nats, sign agreement and the observed ratio of mean NLL to mean KL gap.

Target	Window	b	Observed	Predicted	MAE	Max.	Signs	Ratio
Tail swap	Diagnostic, 1–128	0.8539	0.403	0.388	0.017	0.053	16/16	0.886
Tail swap	Untouched, 1–128	0.8581	0.406	0.392	0.015	0.031	16/16	0.887
Tail swap	Untouched, 385–512	0.9831	0.457	0.447	0.010	0.016	16/16	1.006
Class-mean tail	Diagnostic, 1–128	0.8539	0.310	0.302	0.009	0.018	16/16	0.877
Class-mean tail	Untouched, 1–128	0.8581	0.310	0.304	0.007	0.015	16/16	0.875
Class-mean tail	Untouched, 385–512	0.9831	0.347	0.342	0.005	0.006	16/16	0.997

I Development comparisons

The development evidence contains all 2,400 metric rows of 40 model–depth–split cells and is supplied as supplementary numerical data. These documents were previously examined and are not an untouched confirmation set; they motivate the confirmation experiment rather than supplying its inference.

Each of the 480 selected fits minimizes the corresponding validation objective over ten learning rates and thirty epochs; all 4,800 rate histories are retained. All rate histories use the identical recipe, seeds and data. Two hard full-affine Qwen3.5-9B layer-31 fits select the lowest rate, and 107 fits select epoch 30. Every selected scalar preserves the argmax.

For the development-only conditional intervals, define the document influence

$$u_j = \frac{A_j - \hat{\mu}_A n_{A,j}}{\bar{n}_A} - \frac{B_j - \hat{\mu}_B n_{B,j}}{\bar{n}_B}.$$

The standard error is its sample standard deviation divided by $\sqrt{400}$, with a normal 95% interval. The evidence retains all 7,680 metric contrasts. These are per-fit intervals without refitting or multiplicity correction; only the separately declared document bootstrap supports the joint confirmation claim.

I.1 Capacity comparisons after matched confidence calibration

Each fitted family uses its own scalar selected on the same validation observed-NLL criterion. Table 31 reports full-affine minus comparator test NLL for soft fits on the development documents, in nats, averaged over five overlapping document partitions; negative values favor full affine. These means are descriptive over this exposed corpus, not independent replications. All per-partition family contrasts under identity, teacher-KL and observed-NLL scaling are retained in the evidence.

Table 31: Calibrated full-minus-restricted-family observed NLL of soft fits on the development documents. Every contrast uses the same test documents, with separately validation-selected scalars.

Model	Layer	Diagonal	Rank 4
Qwen3-8B	8	−0.985	−0.764
Qwen3-8B	16	−0.874	−0.583
Qwen3-8B	24	−0.586	−0.539
Qwen3-8B	31	−0.152	−0.373
Qwen3.5-9B	8	−1.401	−0.801
Qwen3.5-9B	16	−1.343	−0.778
Qwen3.5-9B	24	−0.805	−0.572
Qwen3.5-9B	31	−0.094	−0.187

For soft fits, both calibration selectors give negative full-minus-diagonal and full-minus-rank-four NLL in all 40 cells, and all corresponding conditional document intervals have negative upper endpoints; under observed-NLL calibration the mean gains are 0.780 and 0.575 nats. On the confirmation documents of Table 7, soft full affine has lower calibrated NLL than both diagonal and rank four in all 40 configurations, and at the two early depths it also beats rank 32 in all 20 configurations. Rank 32 uses eight times the trainable parameters of rank four, and its parameter ratio to full affine is $64/(d+1)$; all three maps share the same frozen vocabulary head, so adapter parameter counts do not by themselves determine scoring cost.

J Experimental protocols and reproducibility

Table 32: Protocols reported in this paper. Split repeats reuse a document pool.

ID	Data/model scope	Fitting and scoring
N2	1,000 FineWeb-Edu documents; Qwen3-8B/Qwen3.5-9B; four depths, five splits	Six families, ten rates, thirty epochs, hard versus full-distribution targets; all teacher/observed endpoints and three post-scalings; development only.
N3	Additional 1,200 deduplicated FineWeb-Edu documents; both Qwen checkpoints	Frozen N2 readouts and rank-32 bridge; three fixed scalars, no fitting or selection. Shared-document bootstrap for the two declared effects.
N4	400/200/400 WikiText article roles; Llama-3.1-8B-Instruct, layers 8/16	Three families, two targets, ten rates, thirty epochs. Seed 42 is fixed before confirmation; two further training seeds use the same articles. All three seeds and scalar rules retained.
M	Same Llama articles; layers 8/16, two insertion sites	Matched full-affine hard/soft and soft low-rank fits, ten rates, 60 epochs, common selectors, continuous scalars; Appendix D.
T	Same Llama articles, 400/100/100/400 roles	Full, argmax, entropy-matched, DKPP and stratified-shuffle targets, four paired seeds, ten rates, 60 epochs; cheap and constant references in Appendix E; extended-budget rank-256/full-affine comparison in Appendix D.
H	400 untouched WikiText articles, two windows	Every T readout rescored without refitting; Appendix C.

N2 data. The fixed FineWeb-Edu pool (Penedo et al., 2024) was sampled from the first shard of a public, shuffled 100B-token release, using a fixed reservoir seed listed in the supplementary README after Unicode compatibility normalization, NFKC, and whitespace normalization, uniqueness filtering and a 256-character minimum. Each document is independently truncated at 128 tokens. Normalized intermediate states and aligned observed labels are kept for all 1,000 documents. Teacher distributions are recomputed from the frozen final states, never from readout predictions. The positive scalar multiplies the logits after the head, so it tests probability calibration without changing the mapped state. Selected-fit boundaries, all calibration settings and all 2,400 test metric rows are provided as supplementary numerical data; Table 33 lists the evidence level of every reported result. N2's fixed-logit scalar check has zero argmax mismatches and is the evidence for the confidence claim.

K Supplementary data and field definitions

Table 33 states, for each result, the level of evidence supplied. The results fall into four tiers: the Qwen confirmation on new documents with every fit fixed beforehand; diagnostics on the exposed Llama articles; rescoring of fixed Llama readouts on untouched articles; and registered prospective tests of the tail controls. The uploaded supplement contains the code, every analysis entry point with its output, the registrations, the Qwen endpoints, selections and paired document differences, the per-seed Llama replication results, and the Llama article statistics of the fitting-seed, tail-control, fitting-difficulty and teacher-proxy analyses; the document and article texts are rebuilt from the public corpora by the included code. Evidence marked "complete package" belongs to the complete supplementary package, whose additional files the supplement lists.

Table 33: Evidence level of each reported result in the supplement. "Article/document statistics": per-article or per-document sufficient statistics from which every reported mean and interval is recomputed. "Histories": complete candidate learning-rate/epoch histories and selections.

Result	Supplied evidence
Qwen confirmation: Tables 1, 2 and 6, Figure 1	Configuration-level endpoints of all 1,560 evaluations under the three scalars; fixed scalars, selected learning rates and epochs; per-configuration, per-document soft-minus-hard differences, from which the primary intervals and Table 26 are recomputed exactly. Complete package: document statistics of all 520 readouts under the primary observed-NLL scalar, from which the primary means and intervals are recomputed
Llama/WikiText replication: Table 3, Appendix B	Per-seed analysis outputs with the primary effect and interval of each seed. Complete package: article statistics of every selected readout for all three seeds and scalar rules; learning-rate grids and selections
Matched Llama controls: Appendix D	Complete package: histories, selections and article statistics of all 96 variants
Target controls at both depths: Tables 22, 21 and 23	Article statistics of the four-target and stratified-shuffle readouts used for the fitting-seed analysis. Complete package: histories of all 320 four-target trajectories, 96 selected variants and article statistics; the 24 stratified-shuffle selections with their article statistics and boundary records
Frozen-head attainability: Table 20	Summary values of the four targets over the 2,048 sampled positions, with their computation. Complete package: per-position values
Extended-budget capacity and registered contrasts: Tables 18 and 34	The recomputed registered contrasts. Complete package: selections, boundary records and article statistics of the six extended-budget readouts, from which Table 34 is recomputed
Cheap and constant references, readout policy: Tables 4 and 24	Complete package: article statistics of all 78 scored variants and 6 constant predictors; validation-role scores
Untouched articles: Tables 5, 14 and 15	Article statistics of the untouched-article readouts used for the fitting-seed analysis; the cohort selection, window definitions and the bootstrap, with the recomputed contrasts. Complete package: article statistics of all 200 rescored readouts in both windows, namely 158 first-stage readouts, 24 shuffle and 18 extended-budget readouts, and the final cohort list, from which every untouched-article contrast is recomputed, averaging four fits within article, while extended-budget contrasts condition on one fit
Fitting-seed uncertainty: Tables 25 and 26	The article statistics and paired document differences above, from which both tables are recomputed
Teacher-proxy account: Table 27	The article statistics above and the recomputed table and held-out predictions. Complete package: the Qwen document statistics, from which the table and every held-out prediction are recomputed
Tail controls and prospective teacher-proxy test: Tables 28, 29 and 30	Selections, boundary records, target statistics and article statistics of all 32 selected readouts on the diagnostic and untouched articles, from which the three tables are recomputed; the target construction with its invariants checked on synthetic distributions; the per-epoch running training losses behind the fitting-difficulty values
Qwen scalar-grid bound: Table 35	Validation observed-NLL values of all 480 original Qwen readouts at the 33 grid scalars
Development comparisons: Appendix I	2,400 metric rows. Complete package: the document statistics under all three scalars

For the Qwen confirmation, the supplement contains configuration-level endpoints of all 1,560 scalar evaluations and, in the complete package, document statistics of all 520 selected readouts under the primary scalar; the primary analysis uses token-weighted document sums, 5,000 shared document-resampling draws with a fixed generator seed listed in the supplementary README and per-model 97.5% percentile intervals. For the Llama experiments the complete package contains

candidate histories, selections and article-level statistics of the three-seed replication, the matched controls, both four-target depths and the cheap baselines. Backbone weights, fitted readouts and intermediate states are regenerated from the public Qwen3-8B, Qwen3.5-9B and Llama-3.1-8B-Instruct checkpoints and the texts. The supplement gives the models and corpora, the document and article identifiers with their roles, code that rebuilds the texts from the public corpora, and reference implementations of corpus preparation, role assignment, state extraction, which sets the layer, insertion site, 128-token windows and label shift, target construction, the Qwen and Llama readout families, fitting, calibration and scoring, together with checks of teacher-head consistency, layer indexing, the normalization site, the label shift, the window denominators and the document selection of both untouched-article windows. Per-document records hold the summed observed-token loss, the number of correct observed predictions and the number of observed tokens; the summed full-distribution forward KL and the number of teacher positions; and the number of predictions equal to the teacher argmax. The recorded teacher cross-entropy sum is taken against the teacher's argmax class, not against the full teacher distribution, and the recorded entropy sum is the entropy of the readout's own predictive distribution.

L PREDICTIVE APPROXIMATION NEED NOT CONTROL INTERVENTION RESPONSE

The ladder measures how well a readout predicts the final argmax on unperturbed states. This appendix records why such a measurement, however good, does not by itself bound how the readout's prediction would track the model's response to a perturbation of the state. A single token's vector also does not determine the remaining computation, because later attention reads prefix states. A tokenwise readout therefore estimates predictive regularities conditional on its input vector; its success measures that fitted predictor, not a completed decision at the probed token.

Construction. Consider width two, identity output head, and $\mathrm{RMS}(h) = \sqrt{2}h/\|h\|$. Let

$$f_0(h) = \mathrm{RMS}(h),$$
$$f_c(h) = (I + c\, e_2 e_2^\top)\, \mathrm{RMS}(h).$$

Both belong to the identity-plus-rank-one post-normalization family, which includes rank zero. Let the input distribution have positive density on the full-dimensional box $[1, 2] \times [-\sigma, \sigma]$, $\sigma > 0$.

Step 1: the predictive error is small. The logit difference has norm at most $\sqrt{2}|c|\sigma$. The log-sum-exp Hessian has operator norm at most $1/2$, so its Bregman divergence satisfies, pointwise,

$$\mathrm{KL}\big(\mathrm{softmax}\, f_0 \,\|\, \mathrm{softmax}\, f_c\big) \le \frac{c^2\sigma^2}{2}.$$

Step 2: the response discrepancy is large. At $h = (1, 0)$, the logit-map directional Jacobian discrepancy along e_2 is $\sqrt{2}|c|$.

Step 3: the two can separate arbitrarily. Taking $\sigma = c^{-2}$ in Step 1 gives

$$\frac{c^2\sigma^2}{2} = \frac{1}{2c^2} \;\longrightarrow\; 0,$$
$$\sqrt{2}|c| \;\longrightarrow\; \infty \qquad (|c| \to \infty),$$

so the predictive KL bound tends to zero while the response discrepancy of Step 2 diverges. Thus an arbitrarily small value-only error has no distribution-uniform response bound without directional-coverage or conditioning assumptions on the calibration distribution. Exact equality of differentiable output values on an open neighborhood would determine their derivatives there; equality of argmax labels would not. The construction concerns finite predictive error over a family of increasingly thin distributions; it makes no claim about the support geometry of real hidden states, and the bounds were checked numerically for $c \in \{4, 16, 64\}$.

A sufficient condition under directional assumptions. For a scalar token margin m and its readout estimate $\hat{m}$, twice continuously differentiable along an intervention segment $h + t\delta$, $t \in [0, \varepsilon]$, let

$$g(t) = \hat{m}(h + t\delta) - m(h + t\delta).$$

Assume explicitly that $|g'(0)| \le \eta$ and that the second derivative of g is bounded by $B\|\delta\|^2$ on the segment. By Taylor's theorem,

$$\begin{aligned}\left|(\hat{m}(h + \varepsilon\delta) - \hat{m}(h)) - (m(h + \varepsilon\delta) - m(h))\right| &= |g(\varepsilon) - g(0)| \\ &\le \varepsilon|g'(0)| + \tfrac{1}{2}\varepsilon^2 \sup_{t\in[0,\varepsilon]} |g''(t)| \\ &\le \varepsilon\eta + \tfrac{1}{2}B\varepsilon^2\|\delta\|^2.\end{aligned}$$

Estimating η and B on a real model requires running the remaining network on perturbed states.

M REGISTERED CONTRASTS

Table 34 lists every registered endpoint of the stratified-shuffle, extended-budget and tail-control studies of Appendices D, E and H on the diagnostic articles; the six registered teacher-proxy tests are in Table 30.

Table 34: All registered stratified-shuffle and tail-control contrasts, with 98.75% intervals, and extended-budget contrasts, with 99.17% intervals, on the 400 diagnostic articles; observed NLL in nats.

Contrast	Layer / selector	Estimate	Interval
Full − stratified shuffle	8 / global	-0.501	$[-0.522, -0.482]$
Full − stratified shuffle	8 / conditional	-0.513	$[-0.533, -0.494]$
Full − stratified shuffle	16 / global	-0.432	$[-0.451, -0.414]$
Full − stratified shuffle	16 / conditional	-0.443	$[-0.461, -0.426]$
Full − tail swap	8 / global	-0.437	$[-0.454, -0.421]$
Full − tail swap	8 / conditional	-0.454	$[-0.469, -0.439]$
Full − tail swap	16 / global	-0.354	$[-0.370, -0.337]$
Full − tail swap	16 / conditional	-0.365	$[-0.381, -0.349]$
Full − class-mean tail	8 / global	-0.305	$[-0.318, -0.291]$
Full − class-mean tail	8 / conditional	-0.313	$[-0.326, -0.300]$
Full − class-mean tail	16 / global	-0.306	$[-0.321, -0.292]$
Full − class-mean tail	16 / conditional	-0.316	$[-0.330, -0.301]$
Rank 256 zero start − full	8 / global	0.155	$[0.141, 0.169]$
Rank 256 zero start − full	16 / global	0.144	$[0.131, 0.157]$
Rank 256 random start − full	8 / global	0.145	$[0.132, 0.158]$
Rank 256 random start − full	16 / global	0.145	$[0.132, 0.158]$
Rank 256 zero − random start	8 / global	0.010	$[0.002, 0.018]$
Rank 256 zero − random start	16 / global	-0.001	$[-0.009, 0.007]$

A registered descriptive analysis bounds, from the recorded 33-point validation curves of all 480 original Qwen readouts, the validation observed-NLL decrease that any continuous scalar could reach relative to the selected grid point, using the convexity of the scalar objective through the secant lines of the neighboring grid intervals. Table 35 reports the bound. Because each arm can gain at most its own bound, the configuration-mean validation soft-minus-hard difference can move by at most the mean bound of the argmax arm in one direction and of the full arm in the other.

Table 35: Registered descriptive bound on the grid discretization of the Qwen post-fit scalar: largest validation observed-NLL decrease in nats that a continuous scalar in $[1/16, 16]$ could reach relative to the selected grid point, over the 120 readouts of each model and target.

Model	Target	Mean bound	99th percentile	Maximum bound
Qwen3-8B	Argmax	0.099	0.130	0.137
Qwen3-8B	Full	0.115	0.142	0.142
Qwen3.5-9B	Argmax	0.115	0.164	0.165
Qwen3.5-9B	Full	0.089	0.116	0.120