TRAINING-SUFFICIENT SUPERVISION: WHAT DISTILLATION CACHES MUST PRESERVE

Anonymous authors
Paper under double-blind review

ABSTRACT

A distillation cache can be built to reconstruct the teacher's prediction, or to reproduce the training a student would have performed. These are different requirements. For cross-entropy under a frozen affine head the teacher enters the update only through $W_0^\top q$, so targets differing elsewhere on the simplex induce identical gradients. We call a representation *training-sufficient* when it preserves the gradient field over the whole declared parameter domain rather than at one iterate, and characterize it: a fixed linear cache serves every target exactly when the reachable space of centered logit changes lies in its column space, and the least number of coordinates it can use is the dimension of that space. Two constructions follow, one for a frozen head and one for output directions opened on a schedule, the second identifying what must be appended when training reaches directions the cache does not cover. Controlled interventions confirm the predicted separation and its repair. Trained at a matched byte budget against renormalized Top-k and a low-rank reconstruction of the teacher distribution, the two exact representations hold gradient error four to five orders of magnitude lower along the whole trajectory, and store the supervision in under 5% of a full target's bytes, which on one pair yields a measured $1.22\times$ per-step rate. Sufficiency remains a property of the declared loss and training family, not of the teacher alone: reversing the divergence, with the student untouched, raises the requirement to $V - 1$ coordinates.

1 SUPERVISION THROUGH THE STUDENT

Offline distillation separates two stages. A frozen teacher produces a target distribution for each training prefix, those targets are written once, and a student then reads them repeatedly over many updates. Whoever writes the cache has to decide what it should contain. One answer is to store enough to reconstruct the teacher's prediction, which is what a cache of teacher hidden states does (Pai et al., 2026). Another is to store enough to reproduce the training the student would have performed. These are different requirements, and the second is the one an offline cache is written for. The question this paper asks is therefore: given a student, a loss and a declared domain of trainable parameters, which information in the teacher's target determines the gradient field over that whole domain, and when the training setup changes, when does that information remain enough and when must it be extended?

A single gradient makes the gap visible. For cross-entropy against a fixed target q under a linear output head W_0 that is not trained,

$$\nabla_h L = W_0^\top p_\theta - W_0^\top q, \tag{1}$$

where p_θ is the student's current prediction. The student's own state determines the first term; the teacher enters only through the second, as $W_0^\top q$. Two targets that agree in that product therefore induce the same internal gradient, whatever they do elsewhere on the simplex. The dimensions involved are far apart in practice: in the Qwen3 pair studied below the teacher assigns probability to $151{,}936$ vocabulary entries, while the student's frozen head transmits the target through a $1{,}024$-dimensional hidden state. We write d_S for that width and treat it as a sufficient size, distinguishing it throughout from the smallest sufficient number of coordinates, which is a separate quantity defined in Section 2.

Equation (1) is an identity, and for a frozen affine head it already exhibits a target map that does not depend on the current parameters. What it does not settle is the general question, because a

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

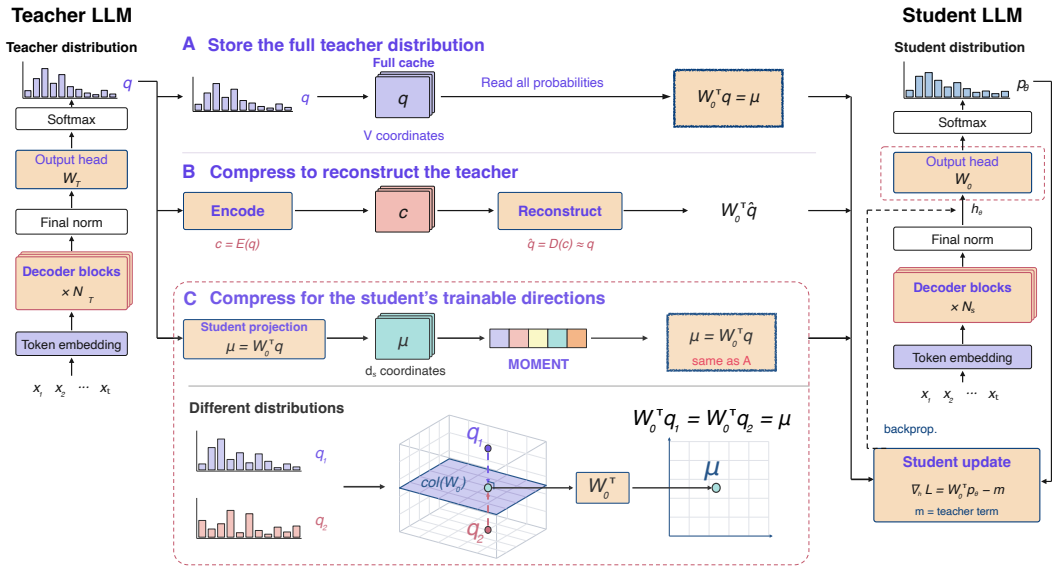


Figure 1: **Three ways to store a teacher target, and what the student's update actually reads.** A frozen teacher supplies a distribution q over V vocabulary entries. The student trains its backbone under a frozen output head W_0, so the target reaches its parameters only through $\nabla_h L = W_0^\top p_\theta - W_0^\top q$. **(A)** Storing q in full keeps V coordinates and the update then reads the projection $W_0^\top q$ anyway. **(B)** Storing a code from which $\hat{q} \approx q$ can be reconstructed keeps the route back to the teacher's distribution, and so serves any head compatible with that teacher, at a size the teacher sets. **(C)** MOMENT stores $\mu = W_0^\top q$ directly: d_S coordinates set by the student, and the same teacher term route (A) eventually forms. The lower row of (C) shows why this is a statement about training and not about compression. Two distinct teacher distributions whose difference is orthogonal to $\mathrm{col}(W_0)$ project to the same μ, so they share one cached target and induce the same backbone gradient at every parameter point in the declared family. Section 2 makes the condition exact, and identifies what must be appended when training opens output directions the cache does not cover.

representation chosen by matching gradients at one parameter point need not match them at the next, and a cache is written before training begins. What a cache must preserve is the dependence of the update on the target across the whole domain the training procedure is allowed to visit: there must exist a fixed $c(q)$, not rewritten as the student moves, and a map G with

$$\nabla_\theta L(\theta; q) = G\big(\theta, c(q)\big) \qquad \text{for every } \theta \in \Theta. \tag{2}$$

Calling a representation *training-sufficient* when it satisfies Eq. (2) turns the observation about one output head into a property of a declared training family, and makes it possible to ask when such a c exists, how few coordinates it needs, and what happens to it when the family changes.

Several lines of work approach this question. The geometry of the output layer is understood: a low-rank head limits what the model can express (Yang et al., 2018), can leave tokens unreachable (Demeter et al., 2020; Grivas et al., 2022), and compresses the feedback reaching earlier parameters, which Godey & Artzi (2026) measure as a loss of logit-gradient norm. Those statements are pointwise at an iterate and read as a pathology; the same projection, stated as equality of gradient *fields* over a declared domain, is instead what licenses writing a cache in advance. That training moves within a small subspace is a recurring finding (Gur-Ari et al., 2018; Li et al., 2018; Aghajanyan et al., 2021), and adapters impose one by construction (Houlsby et al., 2019; Hu et al., 2022; Zhao et al., 2024; Chen et al., 2016). Distillation from soft targets goes back to Buciǎ et al. (2006) and Hinton et al. (2015) and is standard for language models (Romero et al., 2015; Sanh et al., 2019; Kim & Rush, 2016; Furlanello et al., 2018; Xu et al., 2024); work on the objective changes the divergence or the sampling distribution (Agarwal et al., 2024; Gu et al., 2024; Wen et al., 2023) and studies the matching target directly (Zheng & Yang, 2024; Müller et al., 2019), while the fixed-head form of Eq. (1) instantiates the exponential-family split into a log-partition term and expected sufficient statistics (Wainwright & Jordan, 2008). What is missing from both lines is a whole-domain criterion: which target coordinates an offline cache must retain to reproduce training,

as opposed to approximating the teacher. Storage work supplies the contrast. Truncation is biased as an estimator of the teacher distribution and costs accuracy at small k (Anshumann et al., 2025; Muralidharan et al., 2024; Dasgupta et al., 2026), while Nitrobrew caches teacher hidden states and evaluates divergences with fused vocabulary-tiled kernels (Pai et al., 2026), in the lineage of fused large-vocabulary losses (Dao et al., 2022; Wijmans et al., 2025; Hsu et al., 2024; Sheng et al., 2025). A teacher-state cache retains the ability to reconstruct the teacher's prediction, and so serves many student heads and any divergence that head can express; the representations below retain only what one declared student family actually reads, are built from the student's own head, and need no calibration data.

We give a criterion for training-sufficiency together with the minimum number of linear coordinates it requires, two constructions that realize it for a frozen affine head and for a scheduled expansion of the output space, and experiments that test the predicted mechanism, the numerical fidelity of the substitution in real models, and its effect on learning.

2 Training-sufficient representations

2.1 A criterion over the whole parameter domain

Fix a training state s and let $z_\theta(s) \in \mathbb{R}^V$ be the student logits, $p_\theta(s) = \mathrm{softmax}(z_\theta(s))$. The trainable parameters range over a connected open set $\Theta \subseteq \mathbb{R}^m$, the map z_θ is continuously differentiable, and the normalized teacher target $q \in \Delta^{V-1}$ is fixed over Θ. This section treats cross-entropy,

$$L_s(\theta; q) = \mathrm{LSE}(z_\theta(s)) - q^\top z_\theta(s), \tag{3}$$

so the target enters linearly. What changes when it does not is taken up in the scope remarks at the end of this section and proved in Appendix A.

A common shift of all logits leaves the prediction unchanged, so define the centering operator $P = I - \mathbf{1}\mathbf{1}^\top/V$ and the space of logit changes the family can reach,

$$\mathcal{V}_s = \mathrm{span}\big\{P[z_\theta(s) - z_{\theta_0}(s)] : \theta \in \Theta\big\}, \tag{4}$$

whose definition does not depend on the reference point $\theta_0 \in \Theta$.

Theorem 1 (Training-sufficiency and its minimum size)**.** *Let $B \in \mathbb{R}^{V\times k}$ be fixed and known to the learner, and consider the cache $q \mapsto B^\top q$.*

1. *For two fixed targets $q_1, q_2 \in \Delta^{V-1}$,*

$$\nabla_\theta L_s(\theta; q_1) = \nabla_\theta L_s(\theta; q_2) \text{ for every } \theta \in \Theta \iff q_1 - q_2 \perp \mathcal{V}_s. \tag{5}$$

2. *$B^\top q$ determines the gradient field for every target in Δ^{V-1} if and only if*

$$\mathcal{V}_s \subseteq \mathrm{col}(PB). \tag{6}$$

3. *The least number of independent linear coordinates admitting that guarantee is*

$$k_{\min} = \dim \mathcal{V}_s. \tag{7}$$

The reading is direct. Two targets give objectives differing by $-\delta^\top z_\theta$ with $\delta = q_1 - q_2$, so their gradient fields agree on Θ exactly when that quantity is constant there, which for connected open Θ means δ annihilates every reachable centered logit change: a target difference that always points where training cannot look changes no update. A fixed cache is sufficient exactly when no difference it discards can move that quantity, so its centered columns must cover every visible direction, and a cache smaller than $\dim \mathcal{V}_s$ leaves one uncovered. The necessity construction and the rank argument are in Appendix A, together with a controlled instance in which six coordinates spanning $\mathcal{V}_s$ suffice while six arbitrary ones do not, so that size and span are seen to be separate conditions.

Eq. (7) counts fixed linear coordinates of the target that are universally sufficient, not bytes under an arbitrary encoding: a nonlinear code for q lies outside part 3, which is why a teacher-state cache is not bounded by it.

In practice one basis is shared by every prefix in a corpus $\mathcal{S}$, not rebuilt for each. The condition then applies to all of them at once, so what such a basis must cover is the joint space

$$\mathcal{V}_{\mathcal{S}} = \operatorname{span}_{s\in\mathcal{S}} \mathcal{V}_s, \tag{8}$$

whose dimension can exceed that of any individual $\mathcal{V}_s$. Nothing else in the statement changes; the quantity to be covered does.

2.2 MOMENT: THE FROZEN AFFINE HEAD

Let the loss be the forward cross-entropy of Eq. (3) and let the head be affine and untrained, $z_\theta = W_0 h_\theta$ with $W_0 \in \mathbb{R}^{V\times d_S}$ fixed and zero bias; the guarantee below is tied to both of these and to the declared internal family. Then $\mathcal{V}_s \subseteq \operatorname{col}(PW_0)$, Eq. (6) holds with $B = W_0$, and

$$\mu = W_0^\top q, \qquad L_{\mathrm{M}}(\theta;\mu) = \mathrm{LSE}(W_0 h_\theta) - h_\theta^\top \mu \tag{9}$$

reproduces the objective and every gradient with respect to the declared internal parameters. We call this representation MOMENT. A fixed nonzero bias and the target entropy of forward KL affect the recovery of the loss *value* but not its gradients, and are handled in Appendix B. Here d_S is a sufficient size that the student's own head supplies directly; it is not automatically the minimum, since $\dim \mathcal{V}_s$ may be strictly smaller for a given family, and the adapter rank does not bound it (Appendix A.7).

2.3 MOMENT+: SCHEDULED DIRECTIONS AND REUSE

Training often opens output directions later in a run. Let

$$W = W_0 + UA, \qquad \mu = W_0^\top q, \qquad \nu = U^\top q, \tag{10}$$

with W_0 and $U \in \mathbb{R}^{V\times r}$ both fixed and known before the cache is written, and only the coefficients $A \in \mathbb{R}^{r\times d_S}$ trainable. This is the scope of the construction: U is a declared set of output directions, not a second trainable factor, and the guarantee does not extend to a head in which both factors are free. The pair (μ, ν), which we call MOMENT+, preserves the objective and all trainable gradients through

$$L_{\mathrm{M}+}(\theta, A; \mu, \nu) = \mathrm{LSE}\big((W_0 + UA)h_\theta\big) - h_\theta^\top \mu - (Ah_\theta)^\top \nu. \tag{11}$$

New coordinates are needed only for directions the cache does not already cover. If $U = W_0 R + \mathbf{1}c^\top$ then

$$\nu = R^\top \mu + c, \tag{12}$$

so an opened direction already inside the cached span costs no teacher query, and the same identity lets one basis serve a planned schedule of output spaces; covering that schedule is also necessary, in the sense made precise in Appendix A.

The uncovered case yields a testable prediction. Take two legal targets whose difference $\delta = q_+ - q_-$ satisfies $\mathbf{1}^\top\delta = 0$ and $W_0^\top\delta = 0$, so their MOMENT caches coincide. Under the expanded head, at $A = 0$,

$$\Delta\nabla_A L = -(U^\top\delta)h^\top, \qquad \|\Delta\nabla_A L\|_F = \|U^\top\delta\|_2\,\|h\|_2, \tag{13}$$

while the internal gradient at $A = 0$ is still identical for the two targets. Opening a direction with $U^\top\delta \neq 0$ therefore separates the newly writable block by a predicted amount, and appending ν restores each target's *own* full-target update; it does not make the two targets agree again, since they are genuinely different targets once that direction is trainable. Section 3.1 measures both the separation and the recovery, and Algorithm 1 in Appendix B states the resulting procedure.

Scope. These constructions are stated for the forward cross-entropy of Eq. (3), and so also for the teacher-to-student divergence $\mathrm{KL}(q\,\|\,p_\theta)$, which differs from it by a term constant in θ; they are stated at a fixed prefix, for a fixed target definition, and they depend on the declared output map and the declared set of trainable directions. Changing the training setup does not automatically invalidate a cache: what matters is whether the information the new setup makes visible is already covered, which is why opening a direction inside the cached span is free while opening one outside it is not. The requirement does move under two changes. A non-affine output map breaks the construction even with every output parameter frozen, because the operative hypothesis is affineness, not frozenness; Appendix A.8 proves this for a softcapped head of the form now deployed in practice. And reversing the divergence to $\mathrm{KL}(p_\theta\,\|\,q)$ raises the requirement for a fixed linear cache that must be universally sufficient over strictly positive targets to $V-1$ coordinates, with the student, its head and its trainable set untouched, under the reachability condition stated in Appendix A.6.

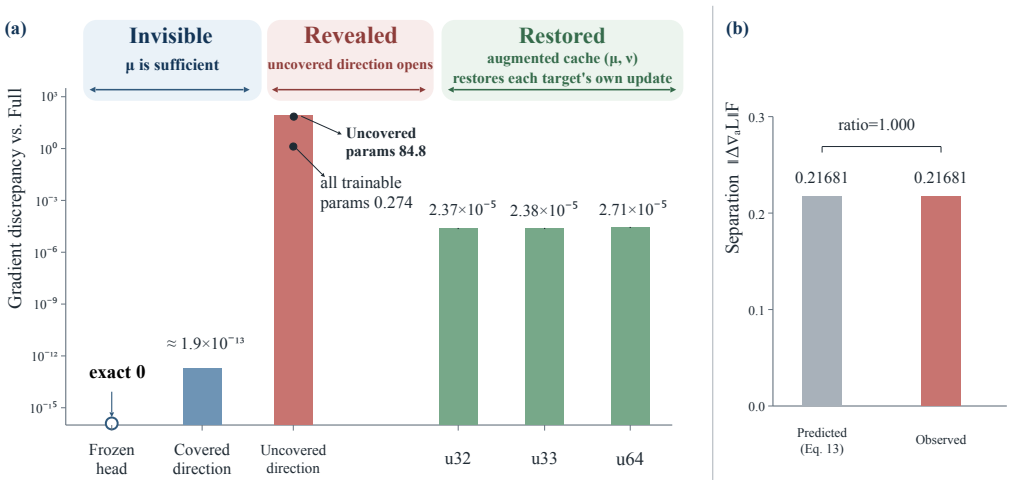


Figure 2: **Supervision is invisible, revealed, then restored.** (a) Gradient discrepancy across the mechanism chain. The first three bars are one constructed target pair at a fixed parameter point carrying the MOMENT cache μ alone: the first two read the separation between the two targets' gradients, the third the relative error of the newly writable block against its own full-target gradient. The last three are a separate staged run with the augmented cache (μ, ν), written once before training, whose added block becomes trainable at update 33. (b) Measured separation of the uncovered intervention against the prediction of Eq. (13). Per-seed values are in Appendix C.

3 MECHANISM AND LEARNING

Every comparison below holds fixed the teacher target, the parameter point or initialization, the supervision mask, the loss reduction and the set of trainable parameters, so that the target representation is the only treatment. Same-point checks compare the gradient obtained from a cached target against the gradient obtained from the full target, over the whole trainable-parameter vector,

$$e_{\text{grad}} = \frac{\|g_{\text{cache}} - g_{\text{Full}}\|_2}{\|g_{\text{Full}}\|_2}, \qquad \|g_{\text{Full}}\|_2 > 0. \tag{14}$$

Exact arithmetic predicts $e_{\text{grad}} = 0$; the measurement reports how far a particular implementation departs from it. This is a gradient comparison, not a comparison of optimizer increments.

Eight teacher–student pairs are used, drawn from three families so that tokenizers match within each pair: from Qwen3 (Yang et al., 2025), **Q1** (4B → 0.6B), **Q2** (4B → 1.7B) and **Q3** (8B → 0.6B); from SmolLM2 (Allal et al., 2025), **S1** (1.7B → 135M), **S2** (1.7B → 360M) and **S3** (360M → 135M); from Pythia (Biderman et al., 2023), **P1** (2.8B → 410M) and **P2** (2.8B → 1.4B). These labels are used throughout. Q1 and Q3 share a student, as do S1 and S3, so quantities that depend on the student alone are reported per student. Model revisions, data identities, cache precisions, adapter and optimizer configurations are in Appendix B, and per-seed records in Appendix C.

3.1 MECHANISM AND STAGED REUSE

We first construct the situation the criterion describes, then open one direction at a time (Figure 2). Two legal targets are built with $\mathbf{1}^\top \delta = 0$ and $W_0^\top \delta = 0$, so their MOMENT caches agree; the construction holds $W_0^\top \delta$ to 3.0×10^{-15}, and the pair differs by 3.94×10^{-5} in L_1 with cross-entropy values equal to sixteen digits. Under the frozen head the two targets produce a bitwise identical internal gradient, as part 1 of Theorem 1 requires. Opening a direction already inside $\mathrm{col}(W_0)$ leaves the new block unchanged, and its coordinate is recoverable from μ alone, so enlarging the trainable set does not by itself invalidate a cache. Opening a direction outside that span separates the new block by the amount Eq. (13) predicts, to 1.3×10^{-17}, while the internal gradient stays bitwise identical. The failure is localized in the newly opened direction, which is what the criterion predicts and what a generic loss of accuracy would not produce.

Table 1: **Measured dimension of the joint reachable space.** $\dim \mathcal{V}_S = \dim \operatorname{span}_s \mathcal{V}_s$ over the measurement set, the space a cache basis shared across prefixes must cover. It depends on the student and its trainable family, not on the teacher, so students shared between pairs appear once. Separation is the smallest resolved singular value in units of the noise level of an otherwise identical run at perturbation scale 10^{-12}; these are empirical ranks at FP32 resolution. The two Pythia students, whose final direction is unresolved, are in Appendix F.3, and the procedure is in Appendix F.

Student	Used by	d_S	$\dim \mathcal{V}_S$	Separation
SmolLM2-135M	S1, S3	576	576	6.62×10^4
SmolLM2-360M	S2	960	960	1.35×10^4
Qwen3-0.6B	Q1, Q3	1024	1024	1.07×10^3
Qwen3-1.7B	Q2	2048	2048	2.22×10^2

The staged run repeats the last step inside training. Here U is a fixed seeded random subspace of rank 64 with 99.66% of its norm outside the student's prior column space, so the directions it opens are genuinely new (Appendix C.1). The added block is held at zero through update 32 and becomes trainable at update 33, the freeze confirmed from the recorded trainable support and optimizer state, and the augmented cache is written once before training and never rebuilt. Across nine same-point checks over three seeds the relative error stays at the 10^{-5} level, worst case 2.79×10^{-5}, and does not step up at the transition. One cache written before training therefore serves a training family that changes during the run, which is the operational content of Eq. (10).

3.2 Reachable dimension in real students

Theorem 1(3) sets the minimum size of a fixed linear cache at the dimension of the space it must cover, and MOMENT uses d_S coordinates. For a frozen affine head Eq. (6) gives $\dim \mathcal{V}_S \le d_S$, so the construction is sufficient; whether it is also smallest depends on whether that inequality is tight, which is a property of the student and its trainable family. Since one basis is shared across prefixes, the deciding quantity is the joint space of Eq. (8). We perturb the adapter isotropically at the training configuration, pool the resulting hidden-state differences over 256 prefixes and ten draws, and read off the numerical rank (Appendix F.1); every column is a legal reachable difference, so that rank is a lower bound on $\dim \mathcal{V}_S$ against the upper bound d_S.

On four of the six students the measured empirical rank of $\mathcal{V}_S$ equals d_S, meeting the upper bound of Eq. (6) (Table 1). The measurement is a finite-precision one, read against the resolution of an FP32 forward pass, so it establishes that the bound is attained at that resolution, short of an identity in exact arithmetic; the margins are wide, the narrowest being a factor of 2.22×10^2 on Qwen3-1.7B, and we separately confirm that the all-ones logit shift is unreachable, so centering removes no dimension. With Theorem 1(3), a joint reachable space that fills the student width means no fixed linear cache below d_S coordinates can serve every target at every prefix, so on these four students MOMENT is consistent with being the smallest cache of its class. This concerns the joint space; an individual $\mathcal{V}_s$ may be smaller, and a per-prefix basis smaller still at the cost of storing one basis per prefix.

On the two Pythia students every direction but the last is resolved: the final singular value falls inside the band that pure numerical noise occupies at that index and stays there when the perturbation is enlarged, so what is left open is a true rank of $d_S - 1$ or a true rank of d_S whose last direction is below FP32 resolution. The spectra and the diagnosis are in Appendix F.3. A cache of d_S coordinates is sufficient either way; only minimality is left open.

3.3 Numerical fidelity in real models

Exact algebra need not translate into exact floating-point arithmetic, so we measure same-point fidelity in real models, over 2,560 probes per pair drawn from ten seeds at a fixed parameter point. Median discrepancies are consistently small with right-skewed tails, and variation across pairs exceeds variation across seeds within a pair (Figure 6(a)). Separating the absolute discrepancy from the reference-gradient scale that e_{grad} divides by (Figure 6(b)) explains part of that spread: within a pair the upper tail tracks small reference norms, not growth in the absolute discrepancy. The Pythia pairs, however, stay above the others at comparable reference norms, so their offset survives the metric and

Table 2: **Four substituted representations at a matched byte budget.** Three pairs, one protocol, 128 updates. MOMENT, Top-k and the SVD cache are given the same bytes per position; the teacher state costs what its width costs. e_{grad} is the median relative gradient error over the training trajectory, and ΔNLL is held-out WikiText NLL minus the Full-target run's, in units of 10^{-3} nats per token; for scale, full supervision moves this readout by 19 to 91 in the same units. Full-target NLL is 2.7219, 2.9100 and 2.7053. Shaded rows are the two representations that are exact by construction. Per-seed values are in Appendix E.2.

Pair	Representation	Bytes/pos.	% of Full	e_{grad}	ΔNLL (10^{-3})
Q1	Full target	607,744	100	—	—
	MOMENT	4,096	0.67	$3.9{\cdot}10^{-6}$	+10.79
	Teacher state	5,120	0.84	0	−0.01
	Top-k	4,096	0.67	$1.3{\cdot}10^{-1}$	+12.84
	SVD cache	4,096	0.67	$9.5{\cdot}10^{-1}$	+182.37
S1	Full target	196,608	100	—	—
	MOMENT	2,304	1.17	$4.8{\cdot}10^{-6}$	−0.11
	Teacher state	4,096	2.08	0	+0.10
	Top-k	2,304	1.17	$1.2{\cdot}10^{-1}$	+9.32
	SVD cache	2,304	1.17	$1.1{\cdot}10^{0}$	+347.10
P2	Full target	201,216	100	—	—
	MOMENT	8,192	4.07	$2.2{\cdot}10^{-5}$	+0.20
	Teacher state	5,120	2.54	$1.5{\cdot}10^{-5}$	+0.16
	Top-k	8,192	4.07	$1.3{\cdot}10^{-1}$	−1.36
	SVD cache	8,192	4.07	$6.0{\cdot}10^{-1}$	+23.44

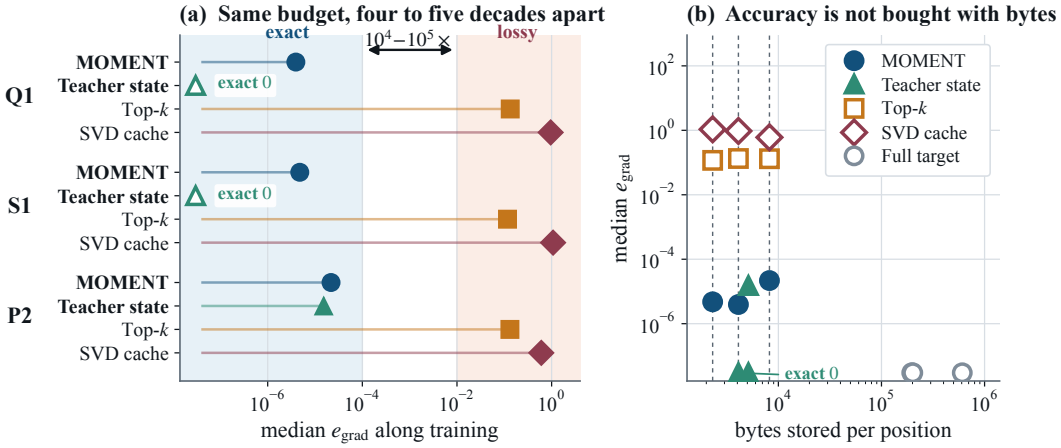


Figure 3: **At a matched budget, the exact representations are four to five decades more faithful.** Median relative gradient error e_{grad} against the Full-target gradient, evaluated along each arm's own trajectory over 128 updates. **(a)** By pair. MOMENT and the teacher state hold 10^{-6} to 10^{-5}; renormalized Top-k and the low-rank SVD cache hold 10^{-1} to 1 while storing the same bytes. On Q1 and S1 the teacher state reproduces the update bitwise, drawn as an exact zero on the floor of the axis. **(b)** The same values against bytes stored. Dashed lines mark the three matched budgets: along each one, the four substituted representations are separated by five decades of error at identical cost, which is the separation Theorem 1 predicts. Values are in Table 2; per-checkpoint curves are in Appendix E.2.

points to the arithmetic path realizing an identity that holds exactly in exact arithmetic; Appendix A.9 bounds what a perturbation of the stored target can do to an update, which is the part of this that the theory does reach. Figure 6, the full quantiles and the variance decomposition are in Appendix C.5.

3.4 LEARNING UNDER THE SUBSTITUTED TARGET

Same-point fidelity says the substituted target produces the same update; it does not say the update was worth making, and it says nothing about what happens once the two runs no longer share a parameter point. We therefore train, at a byte budget matched across representations, and compare three quantities: the gradient error along each arm's own trajectory, how far its parameters travel from the Full-target run, and where its held-out likelihood ends up. Four representations are compared on three pairs for 128 updates. Held-out likelihood is per-token NLL on WikiText (Merity et al., 2017). For scale, full supervision itself moves this readout by 0.019 to 0.091 nats per token on these pairs, so differences of order 10^{-4} are small against what training achieves and differences of order 10^{-1} are not.

Two of the four are exact by construction. MOMENT stores the projection $\mu = W_0^\top q$; the *teacher state* stores the teacher's last hidden vector h_T and recomputes q from it through the teacher's own output head at every visit, which is the representation a fused teacher-state system caches (Pai et al., 2026). It pays teacher-side computation during training and keeps a route back to the full distribution, but it reproduces the same update. The other two are lossy: renormalized Top-k keeps the k largest probabilities, and the SVD cache stores a low-rank reconstruction of the teacher distribution. At matched bytes the two groups separate by four to five orders of magnitude in trajectory gradient error, the exact routes holding 10^{-6} to 10^{-5} while the lossy ones sit between 0.1 and 1 (Figure 3). The measurement is taken at the parameters each arm has itself reached, so it is a property of the whole run: over the four checkpoints of Appendix E.2 the exact routes stay flat, and the parameter divergence from the Full-target run keeps the same ordering as the gradient error. A budget spent on reconstructing the teacher's distribution buys accuracy the optimizer cannot read, which is what Theorem 1 predicts and what these arms make measurable.

The ordering carries into learning, with a compressed spread. On S1 and P2 both exact routes land within 2×10^{-4} nats per token of the Full-target run, Top-k within 9×10^{-3}, and the SVD cache 0.023 to 0.347 away, the last comparable to the gain full supervision produces in the first place. Q1 separates the two readouts. There the projection ends 0.0108 above Full and Top-k 0.0128 above it, so the two are close on likelihood while their gradient errors differ by more than four decades. Exactness of the target map, which is what Theorem 1 delivers, does not by itself pin a trajectory: optimizer state and reduction order compound small differences over a hundred updates, and on Q1 that compounding is visible at the scale of the readout. An earlier and broader survey, eight pairs at 64 updates, is in Appendix E.3; there the largest substitution difference is 0.0011.

3.5 FREE-RUNNING GENERATION

Held-out likelihood is a teacher-forced readout, so we also ask what the substituted models do when they generate. In the same-tokenizer setting, substituted arms are compared against their own Full reference by greedy decoding on the GSM8K test set (Cobbe et al., 2021) and on SVAMP (Patel et al., 2021), under two prompting protocols and for two head structures. Differences are accuracy in percentage points over three training seeds. Of the 72 evaluated units, 28 differ from their Full reference by less than a percentage point, which is itself below the seed-to-seed spread of the Full arms; the other 44 differ by more, in both directions. The 24 main units are in Appendix E.4. Accuracy is a coarser readout than likelihood at this scale, and equivalence on one does not follow from equivalence on the other. A second, exploratory experiment carries the construction across tokenizers, where the target must be realigned before projection and the realignment is itself lossy (Appendix E.5).

4 COST AND REUSE

Training-sufficiency decides what a cache must contain; it does not by itself decide whether that cache is the smaller one. With b_M bytes per projected coordinate and b_H per teacher coordinate,

$$B_M = b_M d_S, \qquad B_H = b_H d_T, \qquad \frac{B_H}{B_M} = \frac{b_H d_T}{b_M d_S}, \tag{15}$$

so the comparison turns on the two widths and the two storage precisions, not on the choice of method. At FP32 projections against natively BF16 activations the boundary sits at $d_T = 2d_S$. Across the

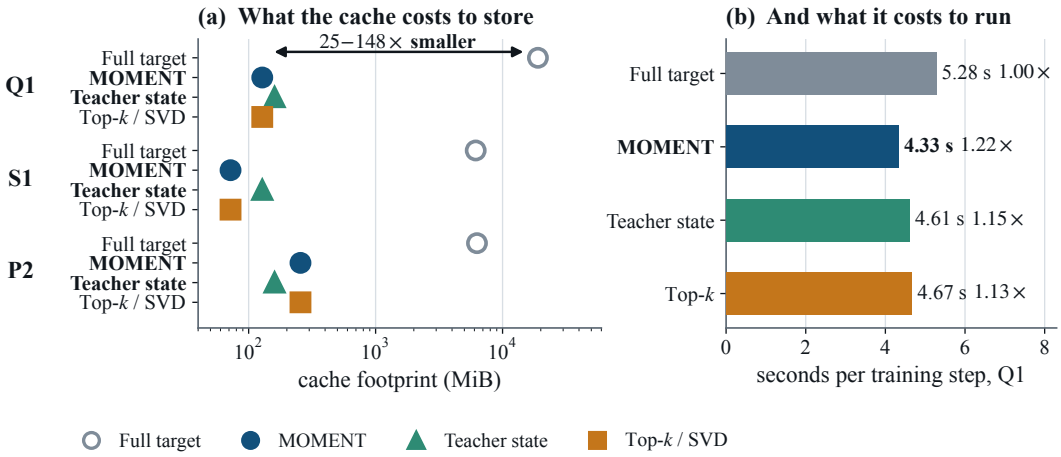


Figure 4: **The projection is smaller to store and faster to train with. (a)** Cache footprint over the 32,768 supervised positions of the run. Top-k and the SVD cache are budget-matched to MOMENT and so occupy the same footprint. **(b)** Measured wall-clock per training step on Q1, with the rate relative to the Full-target run; the step is instrumented and carries probe overhead a production run would not. A per-step suite over two students and both head structures, decomposed into read, forward–backward and optimizer time, is in Appendix E.6.

eight pairs the ratio ranges from 2.000 to 0.625: five favour the projection and three do not (Table 9). That the ratio crosses one inside our own model set is part of the result: Eq. (15) says in advance which regime a configuration is in, which is more useful than a claim that one representation is always smaller.

The two representations also differ in what they can be reused for. MOMENT specializes to a declared student family, whereas a teacher-state cache keeps a route back to the teacher's prediction and with it broader reuse across students and, by Theorem 2, across divergences. Multi-student amortization, the cost of preparation under on-policy refresh, and a cross-tokenizer setting in which the byte ordering reverses are worked out in Appendix C.6.

Smaller caches also run faster, by about what the bytes predict. On Q1 a measured training step falls from 5.28 to 4.33 seconds when the full target is replaced by the projection, a rate of 1.22×, with the teacher state at 1.15× and Top-k at 1.13× (Figure 4). The saving is concentrated in getting supervision onto the device; the Transformer step itself is untouched. One pair on one machine, with probe overhead a production run would not carry, so we read it as confirmation that the byte reduction is realizable in time, not as a benchmark; Appendices E.6 and D give a second per-step suite and the limits of both.

Read as advice, the criterion is narrower than a compression result and more actionable. Before writing an offline cache, ask which output directions the next stage of training will be allowed to move: that question, not the size of the teacher's vocabulary, fixes what the cache must contain, and directions a schedule will open later are worth $b_M r$ extra bytes at write time. Its less comfortable form is that a cache is only ever sufficient for a declared loss, so swapping the divergence or softcapping the head invalidates one without touching a student parameter. The loss belongs in the cache's metadata alongside the student's width.

The Q1 result marks where we would go next, since it is the one outcome our theory does not reach: a companion to Theorem 1 that bounds realized trajectory divergence, through the conditioning of W_0 and the summation order, would turn it into a prediction, and would give the tolerance version of sufficiency that lossy representations already need.

AI USE STATEMENT

Large language models were used to polish prose, to assist with LaTeX formatting, and to help check the proofs of the mathematical claims: verifying derivations in the appendices and flagging steps that did not follow from the stated assumptions. The theorems, their proofs and the experimental findings are the authors' own, all numerical values were produced by the authors' code and verified against the recorded run outputs, and no AI tool was used to generate data or results.

REPRODUCIBILITY STATEMENT

The theoretical claims are self-contained: Appendix A states every assumption and proves Theorem 1, Theorem 2 and the accompanying propositions. Two constructions in it can be re-run without any trained model, and we recommend them as the entry point: the small-dimensional witness of Appendix B.4, which exhibits two targets that a frozen head cannot distinguish and the update that separates them once a direction opens, and the coverage check of Appendix A.3, which needs only a 64×6 matrix and a pseudoinverse.

For the model experiments, Algorithm 1 states the cache construction, and Appendix B defines each target representation, the matched-comparison protocol, and the trainable support. Appendix C gives the per-seed and per-pair values behind every main table, including the widths and payloads of the eight pairs, the staged-expansion checks, and the replay conditions under which a stored teacher activation meets or misses the criterion. Appendix E.2 records the matched-budget protocol and its per-arm seed counts, Appendix E.3 the eight-pair survey and all seventeen endpoint comparisons, and Appendix E.4 the twenty-four generation units. Model identities, adapter and optimizer configurations, and evaluation protocols are specified with the measurements they belong to. All code, configuration files and the original run manifests are included in the supplementary materials.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2306.13649.

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, pp. 7319–7328. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.568. URL https://aclanthology.org/2021.acl-long.568/.

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

Anshumann, Mohd Abbas Zaidi, Akhil Kedia, Jinwoo Ahn, Taehwak Kwon, Kangwook Lee, Haejun Lee, and Joohyung Lee. Sparse logit sampling: Accelerating knowledge distillation in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 18105–18124. Association for Computational Linguistics, 2025. URL https://aclanthology.org/2025.acl-long.885/.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430. PMLR, 2023. URL https://proceedings.mlr.press/v202/biderman23a.html.

Cristian Buciǎ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*,

pp. 535–541, 2006. doi: 10.1145/1150402.1150464. URL https://dl.acm.org/doi/10.1145/1150402.1150464.

Tianqi Chen, Ian Goodfellow, and Jonathon Shlens. Net2net: Accelerating learning via knowledge transfer. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1511.05641.

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

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems*, 2022. URL https://arxiv.org/abs/2205.14135.

Sayantan Dasgupta, Trevor Cohn, and Timothy Baldwin. Don't ignore the tail: Decoupling top-K probabilities for efficient language model distillation. *arXiv preprint arXiv:2602.20816*, 2026. URL https://arxiv.org/abs/2602.20816.

David Demeter, Gregory Kimmel, and Doug Downey. Stolen probability: A structural weakness of neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2191–2197. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.198. URL https://aclanthology.org/2020.acl-main.198/.

Tommaso Furlanello, Zachary C. 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. PMLR, 2018. URL https://proceedings.mlr.press/v80/furlanello18a.html.

Gemma Team. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024. URL https://arxiv.org/abs/2408.00118. Version 3, October 2, 2024.

Nathan Godey and Yoav Artzi. Lost in backpropagation: The LM head is a gradient bottleneck. *arXiv preprint arXiv:2603.10145*, 2026. URL https://arxiv.org/abs/2603.10145. Version 2, July 10, 2026; accepted to COLM 2026.

Andreas Grivas, Nikolay Bogoychev, and Adam Lopez. Low-rank softmax can have unargmaxable classes in theory but rarely in practice. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pp. 6738–6758. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.465. URL https://aclanthology.org/2022.acl-long.465/.

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

Guy Gur-Ari, Daniel A. Roberts, and Ethan Dyer. Gradient descent happens in a tiny subspace. *arXiv preprint arXiv:1812.04754*, 2018. URL https://arxiv.org/abs/1812.04754.

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.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2790–2799. PMLR, 2019. URL https://proceedings.mlr.press/v97/houlsby19a.html.

Pin-Lun Hsu, Yun Dai, Vignesh Kothapalli, Qingquan Song, Shao Tang, Siyu Zhu, Steven Shimizu, Shivam Sahni, Haowen Ning, and Yanning Chen. Liger kernel: Efficient triton kernels for LLM training. *arXiv preprint arXiv:2410.10989*, 2024. URL https://arxiv.org/abs/2410.10989.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2106.09685.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1317–1327. Association for Computational Linguistics, 2016. doi: 10.18653/v1/D16-1139. URL https://aclanthology.org/D16-1139/.

Chunyuan Li, Heerad Farkhoor, Rosanne Liu, and Jason Yosinski. Measuring the intrinsic dimension of objective landscapes. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1804.08838.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017. URL https://arxiv.org/abs/1609.07843.

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

Saurav Muralidharan, Sharath Turuvekere Sreenivas, Raviraj Joshi, Marcin Chochowski, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro, Jan Kautz, and Pavlo Molchanov. Compact language models via pruning and knowledge distillation. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2407.14679.

Dhruv Pai, Timor Averbuch, Alec Dewulf, Ben Keigwin, and Ashley Zhang. Nitrobrew: Fast, lossless distillation for free. Tilde Research technical blog, April 2026. URL https://blog.tilderesearch.com/blog/nitrobrew. Published April 28, 2026.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. Are NLP models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2080–2094. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.naacl-main.168. URL https://aclanthology.org/2021.naacl-main.168/.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for thin deep nets. In *International Conference on Learning Representations*, 2015. URL https://arxiv.org/abs/1412.6550.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. URL https://arxiv.org/abs/1910.01108.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. HybridFlow: A flexible and efficient RLHF framework. In *Proceedings of the Twentieth European Conference on Computer Systems (EuroSys)*, 2025. doi: 10.1145/3689031.3696075. URL https://arxiv.org/abs/2409.19256.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1–2):1–305, 2008. doi: 10.1561/2200000001. URL https://people.eecs.berkeley.edu/~jordan/papers/wainwright-jordan-fnt.pdf.

Yuqiao Wen, Zichao Li, Wenyu Du, and Lili Mou. f-divergence minimization for sequence-level knowledge distillation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 10817–10834. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.605. URL https://arxiv.org/abs/2307.15190.

Erik Wijmans, Brody Huval, Alexander Hertzberg, Vladlen Koltun, and Philipp Krähenbühl. Cut your losses in large-vocabulary language models. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2411.09009.

Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. A survey on knowledge distillation of large language models. *arXiv preprint arXiv:2402.13116*, 2024. URL https://arxiv.org/abs/2402.13116.

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

Zhilin Yang, Zihang Dai, Ruslan Salakhutdinov, and William W. Cohen. Breaking the softmax bottleneck: A high-rank RNN language model. In *International Conference on Learning Representations*, 2018. URL https://arxiv.org/abs/1711.03953.

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. GaLore: Memory-efficient LLM training by gradient low-rank projection. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*. PMLR, 2024. URL https://proceedings.mlr.press/v235/zhao24s.html.

Kaixiang Zheng and En-Hui Yang. Knowledge distillation based on transformed teacher matching. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2402.11148.

CONTENTS

A Proofs and consequences of supervision sufficiency **15**

B Representations and Experimental Definitions **21**

C Numerical Records **22**

D Scope **26**

E Same-budget baselines and learning intervals **27**

F Measuring $\dim \mathcal{V}_s$: procedure, numerical floor, and spectra **32**

A PROOFS AND CONSEQUENCES OF SUPERVISION SUFFICIENCY

A.1 GRADIENT-FIELD EQUIVALENCE

We work at one fixed state and omit its subscript. Let $\delta = q_1 - q_2$. Normalization gives $\mathbf{1}^\top \delta = 0$, hence $P\delta = \delta$. By cancellation of the log-partition terms,

$$L(\theta; q_1) - L(\theta; q_2) = -\delta^\top z_\theta, \qquad \nabla_\theta L(\theta; q_1) - \nabla_\theta L(\theta; q_2) = -J_z(\theta)^\top \delta, \tag{16}$$

where $J_z(\theta)$ is the logit Jacobian.

Proof of Theorem 1, part 1. Suppose the gradients agree everywhere in Θ. Then the continuously differentiable scalar function $f(\theta) = \delta^\top z_\theta$ has zero gradient everywhere. A connected open subset of Euclidean space is polygonally connected. Integrating the zero directional derivative along a polygonal path between θ_0 and θ gives

$$\delta^\top (z_\theta - z_{\theta_0}) = 0 \quad \text{for all } \theta \in \Theta. \tag{17}$$

Since $P\delta = \delta$, the same statement holds with $P(z_\theta - z_{\theta_0})$ in place of the logit difference. It is equivalent to $\delta \perp \mathcal{V}_s$.

Conversely, if $\delta \perp \mathcal{V}_s$, then $\delta^\top z_\theta = \delta^\top z_{\theta_0}$ throughout Θ. Differentiating this equality and using Eq. (16) yields equality of the two gradient fields. □

Changing the reference point does not change $\mathcal{V}_s$: every difference from a second reference is the difference of two vectors relative to the first reference, and the reverse inclusion follows symmetrically. Connectedness is what upgrades a zero local derivative to a single constant across the whole parameter domain. Targets are fixed functions of the state in this argument. Merely stopping autograd through a target that is recomputed as a function of θ does not establish that it is fixed across the domain.

A.2 WHICH LINEAR CACHES ARE SUFFICIENT?

Let $\mathcal{H} = \mathbf{1}^\perp$ and $\mathcal{C} = \mathrm{col}(PB) \subseteq \mathcal{H}$. For any $\delta \in \mathcal{H}$,

$$B^\top \delta = 0 \quad \Longleftrightarrow \quad (PB)^\top \delta = 0 \quad \Longleftrightarrow \quad \delta \in \mathcal{C}^\perp \cap \mathcal{H}. \tag{18}$$

Proof of Theorem 1, part 2. First suppose $\mathcal{V}_s \subseteq \mathcal{C}$. Two normalized targets with the same cache have difference $\delta \in \mathcal{H}$ and $B^\top \delta = 0$. Equation (18) implies $\delta \perp \mathcal{C}$ and therefore $\delta \perp \mathcal{V}_s$. Their gradient fields agree by Theorem 1(1).

For necessity, suppose $\mathcal{V}_s \not\subseteq \mathcal{C}$. Choose $v \in \mathcal{V}_s$ with nonzero orthogonal residual

$$\delta = v - \mathrm{Proj}_{\mathcal{C}}\, v. \tag{19}$$

Both v and its projection belong to $\mathcal{H}$, so $\delta \in \mathcal{H}$. Moreover, $B^\top \delta = 0$, while $\delta^\top v = \|\delta\|_2^2 > 0$. Let $q_0 = \mathbf{1}/V$ and choose

$$0 < \epsilon < \frac{1}{V\|\delta\|_\infty}, \qquad q_\pm = q_0 \pm \epsilon\delta. \tag{20}$$

These are strictly positive normalized targets with the same cache. Their difference is not orthogonal to $\mathcal{V}_s$, so their gradient fields differ somewhere. Thus the cache is not sufficient for all simplex targets.

Finally, inclusion in Eq. (6) implies $\mathrm{rank}(PB) \geq \dim \mathcal{V}_s$. A matrix whose columns form a basis of $\mathcal{V}_s$ attains equality. This proves the independent-coordinate statement. □

This result counts real linear coordinates, not finite-precision bits, and it is a statement about caches of the form $B^\top q$ only. It permits a different basis for each state when that basis is known during reconstruction; constructing, storing or communicating such bases is an additional cost. The universal simplex quantifier is equally essential, and it is what excludes nonlinear codes from the bound: a teacher-state cache passed through the teacher's own nonlinear decoder can reproduce the distributions that teacher actually emits without representing every target in the simplex, so neither this count nor the $V - 1$ bound of Theorem 2 applies to it. Appendix A.3 gives a controlled FP64 construction in which this minimum is known exactly, and checks the coordinate-count and subspace-coverage halves of the statement separately.

Table 3: **Controlled validation of the minimum linear cache size.** A synthetic FP64 construction with $V = 64$ and $\dim \mathcal{V}_s = 6$. Coverage is the condition $\rho(B) < 10^{-8}$ of Eq. (24). Six coordinates suffice when they span the required space; five cannot suffice, and six arbitrary coordinates need not. The last row is what distinguishes cache size from cache span.

Cache coordinates	k	Covers $\mathcal{V}_s$?	Interpretation
Designed spanning basis	6	yes	sufficient
One required column removed	5	no	below the minimum
Independent random basis	5	no	below the minimum
Independent random basis	6	no	correct size, wrong subspace

Loss values and an anchor scalar. The theorem concerns gradients. For a basis B spanning $\mathcal{V}_s$, write

$$z_\theta - z_{\theta_0} = Ba_\theta + c_\theta \mathbf{1}. \tag{21}$$

Then

$$q^\top z_\theta = q^\top z_{\theta_0} + \beta^\top a_\theta + c_\theta. \tag{22}$$

The cache and one anchor scalar $q^\top z_{\theta_0}$ therefore recover the objective value as well. Some structures, including the zero-bias frozen head, already encode this scalar through the same moment.

A.3 A CONTROLLED NUMERICAL CHECK OF MINIMUM LINEAR SUFFICIENCY

Theorem 1(2)–(3) carries two requirements for a fixed linear cache B. It needs at least as many independent coordinates as the dimension of the space it must cover, and those coordinates must span a space containing it once the common-logit-shift direction is quotiented out. Having the right number of arbitrary coordinates is therefore not enough. We check the two requirements separately in a synthetic construction, in double precision, where the correct answer is known in advance.

The vocabulary size is $V = 64$ and the reachable centered logit space is built to have $\dim \mathcal{V}_s = 6$. Let $R \in \mathbb{R}^{V\times 6}$ be a matrix whose centered columns generate that space, so that

$$\mathcal{V}_s = \mathrm{col}(PR), \qquad P = I - \mathbf{1}\mathbf{1}^\top/V, \tag{23}$$

with P the centering operator of Eq. (4). For a candidate cache B the coverage condition of Eq. (6) reads $\mathrm{col}(PR) \subseteq \mathrm{col}(PB)$, which we test through the residual of projecting one space onto the other,

$$\rho(B) = \big\| PR - (PB)(PB)^\dagger PR \big\|_F, \tag{24}$$

where $(PB)^\dagger$ is the Moore–Penrose pseudoinverse. Coverage holds exactly when $\rho(B) = 0$, and is declared numerically when $\rho(B) < 10^{-8}$. Centering is applied to both sides before the comparison, so no cache is credited for the direction $\mathbf{1}$, which no prediction can observe.

Table 3 reports four caches. The designed six-column basis covers $\mathcal{V}_s$. Dropping one of its columns leaves five coordinates and cannot cover a six-dimensional space. An independently drawn five-column cache fails for the same reason. The informative case is the last: an independently drawn *six*-column cache also fails, despite having exactly $\dim \mathcal{V}_s$ columns, because its centered column space does not contain $\mathcal{V}_s$. The conclusion is not that the particular matrix R is irreplaceable, since any basis of $\mathcal{V}_s$ serves equally well; it is that the columns cannot be replaced by arbitrary directions that fail to cover the required space.

The five-column failures also admit a certificate that needs no numerical tolerance. For any B with five columns,

$$\mathrm{rank}(PB) \le 5 < 6 = \dim \mathcal{V}_s, \tag{25}$$

so $\mathcal{V}_s \not\subseteq \mathrm{col}(PB)$ whichever five columns are chosen and however the projection is computed. For the five-column cache actually recorded here the observed deficiency is $\dim \mathcal{V}_s - \mathrm{rank}(PB) = 1$, which is one instance of that bound and not a general claim that every five-column cache attains rank exactly five.

This construction is a numerical witness for the linear minimum of Theorem 1(3) in a setting where $\dim \mathcal{V}_s$ is known exactly: fewer than $\dim \mathcal{V}_s$ coordinates cannot be universally sufficient, and $\dim \mathcal{V}_s$

Table 4: **Additional FP64 checks from the same controlled construction.** Residuals compare a numerical evaluation against the corresponding closed form. Every entry belongs to the synthetic setting of this appendix and none of them is a measurement on a trained model.

Check	Value
Gradient-field equivalence, residual	2.11×10^{-15}
Cache collision of the target pair, $\|B^\top q_+ - B^\top q_-\|_2$	1.19×10^{-16}
Gradient difference under the sufficient cache	1.78×10^{-15}
Necessity witness, gradient separation	1.2108×10^{-1}
Necessity witness against its closed form	1.41×10^{-15}
Centering identity, residual	6.94×10^{-18}
Expanded head, algebraic residual	6.66×10^{-16}
Augmented-cache consistency, residual	1.11×10^{-16}

coordinates suffice only when their centered span covers $\mathcal{V}_s$. It fixes the meaning of the minimum in a case where the answer is known by construction, at a single state and in exact arithmetic. The measurement on trained students is a different object in both respects: Section 3.2 and Appendix F estimate the dimension of the *joint* space $\mathcal{V}_S$ pooled over prefixes, at the resolution of an FP32 forward pass, and find it equal to d_S on four of the six students.

Other identities checked in the same construction. The same synthetic setting exercises several closed forms from this appendix, and Table 4 lists the residuals. They are diagnostics of the construction, in double precision at $V = 64$, and are separate objects from the finite-precision measurements on real models in Section 3.3: agreement here says the algebra was implemented correctly, not that a trained model reproduces it. The necessity witness is the one entry that is not a residual: it is the gradient separation this construction is built to exhibit, and it is a separate quantity from the separation reported for the model-scale intervention of Section 3.1.

One further check compares a measured two-factor boundary gap of 0.11559725468673518 against the closed-form prediction 0.11559725468673519, agreeing to the last stored digit.

A.4 Frozen and expanded output heads

Proof of Eq. (9). Substitution of $z_\theta = W_0 h_\theta + b_0$ into Eq. (3) gives Eq. (9). The target-dependent part of the internal gradient is fully determined by μ:

$$\nabla_h L = W_0^\top p - \mu, \qquad \nabla_\theta L = J_h(\theta)^\top (W_0^\top p - \mu). \tag{26}$$

The remaining target-dependent term, $-b_0^\top q$, is constant in every trainable parameter. It vanishes for a zero-bias head. □

The output matrix must remain fixed as an actual computation parameter. For architectures with tied input embeddings and output weights, this is achieved by freezing the shared tensor or separating the output copy before training. The sufficient size d_S follows from $\mathcal{V}_s \subseteq \mathrm{col}(PW_0)$; equality depends on which hidden-state changes the internal trainable family can realize.

Proof of Eq. (11). Expand the target inner product:

$$q^\top (W_0 + UA)h = h^\top W_0^\top q + (Ah)^\top U^\top q. \tag{27}$$

This proves Eq. (11). Its partial derivatives are

$$\nabla_h L = (W_0 + UA)^\top p - \mu_0 - A^\top \nu, \tag{28}$$

$$\nabla_A L = (U^\top p - \nu) h^\top. \tag{29}$$

They equal the derivatives of full cross-entropy, and the chain rule gives equality for all upstream parameters. If $U = W_0 R + \mathbf{1} c^\top$, normalization yields $U^\top q = R^\top \mu_0 + c$. □

This parameterization keeps the vocabulary-side basis U fixed. Making that basis trainable changes the required target information: for example, the target contribution to its gradient is $-q(Ah)^\top$, which is not generally determined by (μ_0, ν) cached with the old basis.

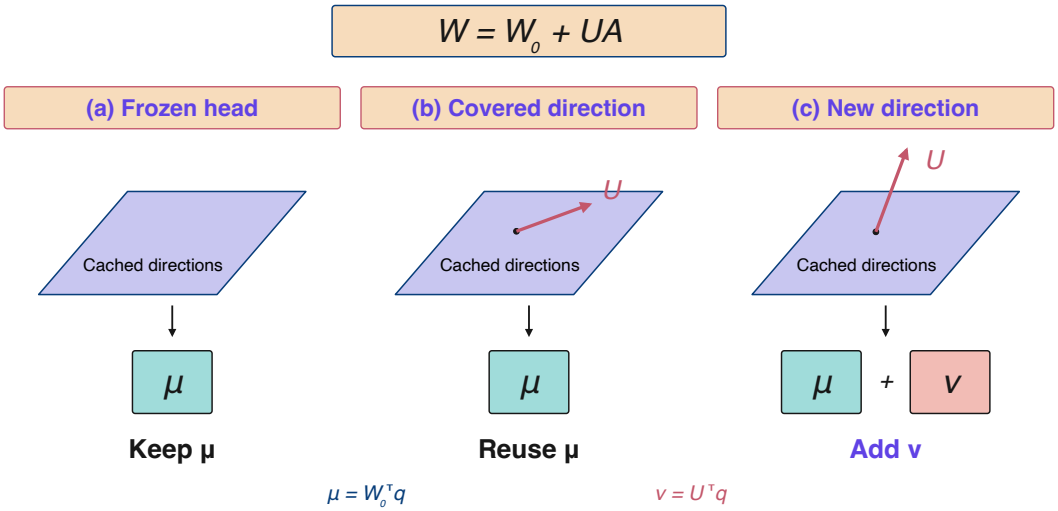


Figure 5: **When a cache stays valid, and what has to be appended when it does not.** The shaded plane is the set of output directions the cache already covers; U is the direction training opens next, and the arrow leaves the plane exactly when that direction is not covered. **(a)** Under the frozen head the cache $\mu = W_0^\top q$ is sufficient. **(b)** Opening a direction inside the cached span needs no new teacher information: if $U = W_0 R + \mathbf{1}c^\top$ then $\nu = R^\top \mu + c$, so the same μ is reused and enlarging the trainable set is not by itself a reason to rebuild. **(c)** Opening a direction outside the span makes teacher differences that were previously invisible move the update, and the missing information is identified rather than merely lost: appending $\nu = U^\top q$ restores exactness for the enlarged family. The three cases are the constructed intervention of Figure 2(a), where (a) and (b) are measured to agree bitwise and to 1.9×10^{-13} while (c) separates by the amount Eq. (13) predicts.

A separation and recovery witness. For two targets with $W_0^\top \delta = 0$, subtracting the gradients in Eq. (29) gives Eq. (13). The Frobenius norm of an outer product is the product of its vector norms, proving Eq. (13). If U lies in $\mathrm{span}\{W_0, \mathbf{1}\}$, the difference is zero. If $U^\top \delta \neq 0$ and $h \neq 0$, the new A gradients differ. A cache collision under $W_0^\top q$ consequently witnesses insufficient information for this larger trainable family. The valid augmented targets in Eq. (11) recover the respective full gradients for both distributions. At $A = 0$ the hidden gradients agree; they need not agree after A has changed.

A.5 Cache conversion and staged training

Proof of the reuse rule. If $W_1 = BR + \mathbf{1}c^\top$, then

$$W_1^\top q = R^\top B^\top q + c(\mathbf{1}^\top q) = R^\top \beta + c. \tag{30}$$

Conversely, universal recovery implies that any two simplex targets with equal $B^\top q$ also have equal $W_1^\top q$. Taking sufficiently small perturbations around the uniform target shows

$$\ker(B^\top) \cap \mathbf{1}^\perp \subseteq \ker(W_1^\top). \tag{31}$$

Orthogonal complementation within $\mathbf{1}^\perp$ yields $\mathrm{col}(PW_1) \subseteq \mathrm{col}(PB)$. This inclusion also implies a factorization $PW_1 = PBR$ for some R, and thus $W_1 - BR = \mathbf{1}c^\top$ for some c. □

For a planned collection of fixed output directions, a basis of their joint span together with the initial output head gives a cache valid throughout the schedule. Directions shared modulo $\mathbf{1}$ need be stored only once. Different teacher targets, prefixes, or token coordinate systems define different cache entries or different conversion problems; Eq. (10) applies to the same fixed normalized target.

A.6 The student-to-teacher direction

Every statement in the main text uses one structural fact: the target enters Eq. (3) linearly, through $q^\top z_\theta$. The teacher-to-student divergence $\mathrm{KL}(q \,\|\, p_\theta)$ differs from it by a term constant in θ and

therefore inherits all of it, with only the loss value requiring the target entropy in addition. The student-to-teacher direction does not: the target acts through $\log q$ rather than q, and it is paired against reachable changes in the *probabilities* rather than in the centered logits. Both the coordinates of the target and the space they are tested against have moved, so what follows is not a weaker version of Theorem 1 but a different regime.

Theorem 2 (One student, two losses)**.** *Fix a frozen affine head $z_\theta = W_0 h_\theta + b_0$ and suppose the internal family can move the hidden state along a segment $h_\theta = h_0 + tv$, with t ranging over a nonempty open interval, whose logit direction $w = W_0 v$ has pairwise distinct entries. Then over this same parameter family:*

1. *for forward cross-entropy, the d_S coordinates $W_0^\top q$ are training-sufficient for every $q \in \Delta^{V-1}$;*
2. *for $\mathrm{KL}(p_\theta \| q)$, equality of the gradient fields implies $q_1 = q_2$, so any fixed linear cache $B^\top q$ that is training-sufficient for all strictly positive targets needs at least $V - 1$ coordinates.*

Changing the loss therefore changes the cache requirement with the student, its head and its trainable set untouched. The hypothesis on w is generic and checkable, and enlarging Θ only enlarges the reachable set, so the conclusion persists for any family containing such a segment; if w takes G distinct values the bound reads $G - 1$. The statement is about fixed linear functionals of the target evaluated on all positive targets and leaves nonlinear codes untouched, which is why a teacher hidden state serves either direction.

The criterion. Write the student-to-teacher divergence as

$$\mathrm{KL}(p_\theta \,\|\, q) = \textstyle\sum_i p_{\theta i} \log p_{\theta i} - p_\theta^\top \log q, \tag{32}$$

and let $\mathcal{U}_s = \mathrm{span}\{p_\theta(s) - p_{\theta_0}(s) : \theta \in \Theta\}$ be the reachable set of probability changes. The criterion behind Theorem 2 is

$$\nabla_\theta \,\mathrm{KL}(p_\theta \| q_1) = \nabla_\theta \,\mathrm{KL}(p_\theta \| q_2) \;\; \forall \theta \in \Theta \quad \Longleftrightarrow \quad \log q_1 - \log q_2 \perp \mathcal{U}_s, \tag{33}$$

which follows by the same integration argument as Theorem 1(1). For strictly positive q, Eq. (32) gives $\mathrm{KL}(p_\theta \| q_1) - \mathrm{KL}(p_\theta \| q_2) = -p_\theta^\top r$ with $r = \log q_1 - \log q_2$, since the entropy term does not involve the target. The two gradient fields agree on Θ exactly when $\nabla_\theta (p_\theta^\top r) = 0$ throughout, and, Θ being connected and open and p_θ continuously differentiable, exactly when $p_\theta^\top r$ is constant on Θ, that is when $(p_\theta - p_{\theta_0})^\top r = 0$ for all θ. This is Eq. (33). Note that $\mathbf{1}^\top (p_\theta - p_{\theta_0}) = 0$, so r matters only modulo $\mathbf{1}$, which is the same freedom as rescaling q_1/q_2 before normalization. Two differences from Theorem 1(1) are worth stating explicitly: the target enters as $\log q$ rather than q, and it is tested against reachable probability differences rather than centered logit differences.

Proof of Theorem 2. Part 1 is Eq. (9) together with Theorem 1(2), since $\mathcal{V}_s \subseteq \mathrm{col}(PW_0)$ for a frozen affine head.

For part 2, restrict θ to the given segment and write $z(t) = W_0 h_0 + b_0 + t\, W_0 v = b + tw$ with $w = W_0 v$, so that

$$p_i(t) = \frac{e^{b_i + t w_i}}{\sum_j e^{b_j + t w_j}}, \qquad t \in I, \tag{34}$$

for a nonempty open interval I. Suppose the two gradient fields agree on the whole parameter domain; then in particular they agree along this segment, so by the criterion above $p(t)^\top r = c$ for some constant c and all $t \in I$, with $r = \log q_1 - \log q_2$. Multiplying by the positive partition function $\sum_j e^{b_j + t w_j}$,

$$\sum_{i=1}^{V} (r_i - c)\, e^{b_i}\, e^{w_i t} = 0 \qquad \text{for all } t \in I. \tag{35}$$

The functions $t \mapsto e^{w_i t}$ with pairwise distinct w_i are linearly independent on any nonempty open interval: a vanishing combination is real-analytic, so it vanishes on all of $\mathbb{R}$, and dividing by $e^{w_{\max} t}$ and letting $t \to +\infty$ kills every term but one, which forces its coefficient to vanish; iterating gives all coefficients zero. Since $e^{b_i} > 0$, Eq. (35) yields $r_i = c$ for every i, hence $q_1 = e^c q_2$, and normalization forces $c = 0$ and $q_1 = q_2$.

Now let $B \in \mathbb{R}^{V\times k}$ be a fixed cache that is training-sufficient for $\mathrm{KL}(p_\theta\|q)$ over all strictly positive targets, and suppose $\delta \in \ker(B^\top) \cap \mathbf{1}^\perp$ is nonzero. Choose a strictly positive q and $\epsilon > 0$ with $q \pm \epsilon\delta$ strictly positive. These two targets have the same cache, so sufficiency forces their gradient fields to agree, and the previous paragraph forces $q + \epsilon\delta = q - \epsilon\delta$, contradicting $\delta \neq 0$. Hence $\ker(B^\top) \cap \mathbf{1}^\perp = \{0\}$, which requires $\mathrm{rank}(B^\top) \geq V - 1$. □

The hypotheses are mild. Distinctness of the entries of w enters only through Eq. (35): if w takes G distinct values, the argument determines q only up to redistribution of mass within each tied group, giving the bound $G - 1$ in place of $V - 1$. Enlarging the parameter domain enlarges $\mathcal{U}_s$, so any family whose reachable set contains such a segment inherits the conclusion. The bound constrains fixed linear functionals of the target evaluated on all strictly positive targets; a teacher hidden state is a nonlinear code for q and is unaffected, as is the separate question of a cache tailored to the low-dimensional manifold of distributions one teacher actually produces.

A.7 Adapter rank does not bound the sufficient size

Eq. (9) gives d_S as a sufficient number of coordinates, and Theorem 1(3) gives $\dim \mathcal{V}_s$ as the minimum. It is tempting to read the rank of a low-rank adapter as a bound on the latter, and that reading is wrong. A rank-one update $h = h_0 + b\,(a^\top x)$ already sweeps a d_S-dimensional set of hidden states as b ranges over $\mathbb{R}^{d_S}$, so its reachable logit space can fill $\mathrm{col}(PW_0)$ even though each individual update has rank one. Three quantities that are easily conflated should be kept apart: the pointwise rank of the Jacobian, the span of reachable differences over the whole parameter domain, which is what Eq. (4) defines, and the directions actually visited along one optimization trajectory. Only the second enters our statements. How far $\dim \mathcal{V}_s$ falls below d_S for a given adapter family is an empirical question we do not settle; a smaller exact cache exists whenever it does, though Theorem 1(3) says nothing about the cost of constructing its basis.

A.8 A frozen head is not enough if it is not affine

Freezing the output parameters is not the operative hypothesis behind Eq. (9); affineness is, and deployed heads are not always affine. Gemma 2 applies $z \mapsto c\tanh(z/c)$ to the final logits with $c = 30$ (Gemma Team, 2024). For a general fixed output map $z_\theta = f(W_0 h_\theta)$ the target-dependent part of the hidden gradient is

$$-W_0^\top\big[f'(W_0 h_\theta) \odot q\big], \tag{36}$$

whose coefficient on q now varies with θ. Pointwise the reachable logit change still has rank at most d_S, as observed for arbitrary heads by Godey & Artzi (2026); the next statement shows that this pointwise fact does not upgrade to a single low-dimensional linear cache valid over a whole family. A scalar construction makes the difference visible: with an affine head two targets sharing $W_0^\top q$ stay indistinguishable along an entire ray, whereas replacing it by $c\tanh(\cdot/c)$ separates them.

Proposition 3 (Softcapped head)**.** *Let $z_i(t) = c\tanh\big((a_i + tw_i)/c\big)$ for $c > 0$, fixed $a \in \mathbb{R}^V$ and $w \in \mathbb{R}^V$ with pairwise distinct $|w_i| > 0$, and let t range over a nonempty open interval. For forward cross-entropy against a fixed target, equality of the gradient fields of q_1 and q_2 implies $q_1 = q_2$. Consequently a fixed linear cache that is training-sufficient for all simplex targets needs at least $V - 1$ coordinates, even though the hidden state is one-dimensional and the output parameters are frozen.*

Proof. With $\delta = q_1 - q_2$ the objectives differ by $-\delta^\top z(t)$, so the gradient fields agree exactly when $\frac{d}{dt}\,\delta^\top z(t) = 0$ on the interval, that is when

$$\psi(t) = \sum_{i=1}^{V} \delta_i\, w_i\, \mathrm{sech}^2\big((a_i + tw_i)/c\big) = 0. \tag{37}$$

Each term is real-analytic on $\mathbb{R}$, so ψ vanishes on the whole line. Assume $\delta \neq 0$ and let j index the smallest $|w_i|$ among the coordinates with $\delta_i \neq 0$; replacing t by $-t$ if necessary, assume $w_j > 0$. Using $\mathrm{sech}^2(u) = 4e^{-2u}\big(1 + O(e^{-2u})\big)$ as $u \to +\infty$, multiply Eq. (37) by $e^{2w_j t/c}$ and let $t \to +\infty$. Terms with $|w_i| > w_j$ decay strictly faster and vanish, the correction to term j is $O(e^{-2w_j t/c})$ and vanishes, and terms with $\delta_i = 0$ are absent, leaving $4\delta_j w_j e^{-2a_j/c} = 0$. Since $w_j \neq 0$ this gives $\delta_j = 0$, a contradiction; hence $\delta = 0$. The cache bound follows as in Theorem 2. □

The mechanism is that sech^2 attaches a different, target-independent but coordinate-dependent decay rate to each vocabulary entry, so the family of maps $q \mapsto \nabla_t L$ separates points of the simplex. The statement concerns exact sufficiency for a fixed linear cache; the corresponding question at a finite tolerance, where a softcapped head may well admit a much smaller approximate cache, remains open. Both boundaries of this appendix are stated for a fixed prefix and a fixed target definition: a newly generated prefix, or a change in how the target itself is defined, calls for its own cached coordinates rather than a conversion of existing ones.

A.9 Finite-precision target perturbations

Exact algebra separates representation sufficiency from arithmetic error. Suppose $\widehat{\mu} = \mu + e_\mu$ is used while the same student normalization term is evaluated. For a frozen head, the induced target-approximation error is exactly

$$\Delta\nabla_\theta L = -J_h(\theta)^\top e_\mu, \qquad \|\Delta\nabla_\theta L\|_2 \le \|J_h(\theta)\|_{\mathrm{op}}\, \|e_\mu\|_2. \tag{38}$$

For an expanded head with $\widehat{\nu} = \nu + e_\nu$, the corresponding partial-gradient errors are

$$\Delta\nabla_h L = -e_\mu - A^\top e_\nu, \qquad \Delta\nabla_A L = -e_\nu h^\top. \tag{39}$$

These formulas quantify the effect of target-coordinate errors at a common parameter point. Rounding in projection accumulation, the student normalization calculation, and the optimizer can introduce additional differences. Long-run trajectory agreement and downstream quality are therefore empirical questions even when the exact-arithmetic objectives coincide.

B Representations and Experimental Definitions

Algorithm 1 A target cache covering a planned training family

Require: Frozen teacher T, prefixes $\{s_i\}$, head W_0, fixed directions U

1: **for** each prefix s_i **do**
2: $\quad q_i \leftarrow \mathrm{stopgrad}(T(\cdot \mid s_i))$
3: $\quad$ Cache $\mu_i \leftarrow W_0^\top q_i$, $\nu_i \leftarrow U^\top q_i$
4: **end for**
5: **for** each training update with the declared active coordinates of A **do**
6: $\quad h_i \leftarrow h_\theta(s_i)$; $z_i \leftarrow (W_0 + UA)h_i$
7: $\quad \ell_i \leftarrow \mathrm{LSE}(z_i) - h_i^\top \mu_i - (Ah_i)^\top \nu_i$
8: $\quad$ Update active parameters using the original weights and optimizer
9: **end for**

B.1 Target construction and matched comparisons

All exact-representation comparisons refer to the same normalized target distribution at the same token prefix. The teacher is frozen, and both teacher targets and prefix sampling are outside the student's differentiation graph. Full evaluates $-\sum_v q_v \log p_v$; Moment evaluates Eq. (9) and Moment+ evaluates Eq. (11). The temperature is one throughout. At another temperature the logits, teacher target, loss multiplier and moments must be rescaled consistently. The target normalization, any effective-vocabulary mask, and the aggregation must match the Full arm exactly, since a mismatch in any of them changes the reference rather than the representation.

Hidden stores the teacher activation entering its final output computation, and reconstruction uses the teacher's output head and the same normalization path. For the representative teachers the native activation is BF16, so storing it in FP32 adds bytes without adding information. The replay comparison reconstructs the target from the saved activation and compares the target, the student gradient and a complete optimizer update against the full-target reference, restoring parameters and optimizer state before each update.

For Top-k, with $\mathcal{I}_k$ the indices of the k largest target probabilities, the renormalized approximation is

$$q_v^{(k)} = \frac{q_v\, \mathbb{I}[v \in \mathcal{I}_k]}{\sum_{u \in \mathcal{I}_k} q_u}, \tag{40}$$

whose training objective differs from the full-target objective, so a lower Top-k training loss is not comparable to a lower full-target loss. The randomized sparse-logit arm samples vocabulary entries instead of truncating to the largest, following Anshumann et al. (2025). Its guarantee holds in expectation over draws, while training-sufficiency is a property of the cache that is actually stored and read: the quantity we report is therefore the realized error of one written cache, which is what the definition of Section 2 requires. The reweighted-SVD arm is implemented from its published construction, and is reported as such.

B.2 Trainable support and parameter identities

Internal adaptation uses LoRA on the declared internal modules, with the output projection frozen for MOMENT. Where the output is tied to the input embedding, freezing or copying the output tensor must preserve the declared head throughout the run and after export. For the expanded family the vocabulary-side matrix U is fixed and the coefficient matrix A is trained, which has different trainable degrees of freedom from a two-factor output-head LoRA in which both factors move; the results of Section 3.1 do not transfer to the latter. Pairing fixes the initial trainable tensors, the input sequence, the selected supervision positions, the weighting and the optimization schedule, so that the target representation is the only treatment.

B.3 Same-point evaluation

Same-point checks compare two expressions at a shared parameter point rather than gradients from separately diverged trajectories. For a nonzero reference gradient g_F over the full trainable-parameter vector,

$$e_g = \frac{\|g_M - g_F\|_2}{\|g_F\|_2}, \tag{41}$$

and e_g is reported as measured. Per-block and total-gradient discrepancies are reported separately.

The staged-expansion checks of Section 3.1 evaluate the written cache against the full target at a common set of 256 positions, both by reading the cache from disk and by recomputing the same expression live in FP32. Both are reported in Table 5.

B.4 A reproducible small-dimensional witness

A standalone calculation instantiates the information boundary without a language model. Let

$$W_0 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad q_0 = \tfrac{1}{4}\mathbf{1}, \qquad \delta = (0.025, 0.025, -0.025, -0.025)^\top, \qquad q_\pm = q_0 \pm \tfrac{1}{2}\delta. \tag{42}$$

Both targets are strictly positive, $\mathrm{TV}(q_+, q_-) = 0.05$ and $W_0^\top \delta = 0$. With $u = \delta/\|\delta\|_2$ and any nonzero h, the fixed-head moments coincide while the newly available coefficient gradient differs by $-(u^\top \delta)h^\top$. The script checks these equalities and the augmented recovery in FP64, with residuals near 10^{-15}, confirming the algebra independently of the model-scale runs.

C Numerical Records

C.1 Staged expansion with a seeded random output subspace

C.2 Frozenness against affineness

A scalar construction separates the two hypotheses behind Eq. (9). With $W = (1, 2, 3)^\top$ and $q_\pm = \frac{1}{3} \pm 0.0125\,(1, -2, 1)$, so that $\mathrm{TV}(q_+, q_-) = 0.05$ and $W^\top \delta = 0$, the gradient separation along h is

h	0	0.1	0.5	1.0
affine head	0	0	0	2.22×10^{-16}
tanh head	0	2.71737×10^{-3}	8.78324×10^{-3}	4.17423×10^{-3}

Table 5: Per-seed same-point checks behind the staged run of Section 3.1, at a 256-position common set. Each cell is update 32 / 33 / 64; the added block is zero at update 32 and updates from 33. Figure 2(a) summarizes these columns as a mean over seeds ± half the observed range.

Seed	On-disk cache	Live FP32
01	$2.5073 / 2.4752 / 2.7239 \times 10^{-5}$	$1.24 / 1.26 / 1.35 \times 10^{-5}$
02	$2.1887 / 2.2297 / 2.6151 \times 10^{-5}$	$1.17 / 1.19 / 1.30 \times 10^{-5}$
03	$2.4248 / 2.4390 / 2.7907 \times 10^{-5}$	$1.25 / 1.24 / 1.32 \times 10^{-5}$

confirmed by autograd against the analytic expression. Agreement at a single point does not extend to the family. The construction says nothing about which points a real optimizer visits, and nothing about whether low-dimensional representations exist for non-affine heads at a finite tolerance.

C.3 Q3 common set

The Q3 same-byte checks are measured at a fixed parameter point on a common set of 256 positions, against the full-target reference gradient. They are not on the same set as the SVD measurements of Appendix E, which carry their own calibration and measurement identity. Two conditions of that measurement are worth separating, because they are often conflated.

First, the stored teacher activation and the arithmetic used to replay it are different objects. In the configurations of Figure 6 and of the cost measurements, reconstruction from the stored BF16 activation reproduces the full target with $\mathrm{TV} = 0$ and zero gradient difference. On this common set the same activation replayed through a BF16 head gives $\mathrm{TV} = 3.25 \times 10^{-5}$, and replayed through an FP64 head gives $\mathrm{TV} = 1.5149 \times 10^{-2}$. The FP64 figure is dominated by the target itself moving from q_{bf16} to q_{fp64} rather than by information lost in the stored bytes, which agreement with an FP64 reconstruction of the same hidden state confirms. The BF16-head residual is not explained that way.

Second, the SVD arm's byte count excludes shared state, which the other arms do not carry, so its compression ratio is not comparable to theirs without that qualification. A full-rank SVD arm stored in FP32 coordinates would occupy 16,384 bytes rather than 8,192; if BF16 values are merely cast to FP32 before use, that is a change of compute precision and not a separate FP32 cache.

C.4 Pythia P1 repeats

P1 is the pair with the largest discrepancies, with a median of 6.13×10^{-5}, a P99 of 4.31×10^{-4} and a maximum of 2.73×10^{-3}. Its per-seed medians span 5.84×10^{-5} to 6.94×10^{-5}, a factor of 1.19, so the position is stable across seeds and not the product of one unlucky run. Q1 has the same declared widths and a median $13\times$ smaller, and the audit subset shows the two separated by a comparable factor in unnormalized discrepancy at matched reference norm, so the difference is not a shape effect and not an artifact of the relative-error metric. We report its distribution as measured.

C.5 Same-point fidelity: full distribution statistics

Each pair contributes 2,560 probes drawn from ten training seeds, at a fixed parameter point. Figure 6(a) plots the median, P95 and P99 columns; the remaining quantiles are recorded here. P25, P75 and P90 were retrieved for four pairs only and the other rows are left blank rather than interpolated.

Pair	P25	Median	P75	P90	P95	P99	Max
Q1	$2.42{\cdot}10^{-6}$	$4.68{\cdot}10^{-6}$	$9.92{\cdot}10^{-6}$	$2.29{\cdot}10^{-5}$	$7.22{\cdot}10^{-5}$	$2.36{\cdot}10^{-4}$	$4.01{\cdot}10^{-4}$
Q2	—	$2.04{\cdot}10^{-5}$	—	—	$1.67{\cdot}10^{-4}$	$4.30{\cdot}10^{-4}$	$1.05{\cdot}10^{-3}$
Q3	$2.22{\cdot}10^{-6}$	$4.08{\cdot}10^{-6}$	$8.34{\cdot}10^{-6}$	$1.88{\cdot}10^{-5}$	$5.24{\cdot}10^{-5}$	$2.45{\cdot}10^{-4}$	$6.22{\cdot}10^{-4}$
S1	—	$6.93{\cdot}10^{-6}$	—	—	$7.28{\cdot}10^{-5}$	$3.42{\cdot}10^{-4}$	$8.69{\cdot}10^{-4}$
S2	$2.46{\cdot}10^{-6}$	$4.69{\cdot}10^{-6}$	$1.06{\cdot}10^{-5}$	$3.04{\cdot}10^{-5}$	$8.22{\cdot}10^{-5}$	$3.94{\cdot}10^{-4}$	$6.88{\cdot}10^{-4}$
S3	—	$7.28{\cdot}10^{-6}$	—	—	$7.66{\cdot}10^{-5}$	$3.97{\cdot}10^{-4}$	$2.31{\cdot}10^{-3}$
P1	$3.76{\cdot}10^{-5}$	$6.13{\cdot}10^{-5}$	$9.85{\cdot}10^{-5}$	$1.60{\cdot}10^{-4}$	$2.14{\cdot}10^{-4}$	$4.31{\cdot}10^{-4}$	$2.73{\cdot}10^{-3}$
P2	—	$1.57{\cdot}10^{-5}$	—	—	$8.22{\cdot}10^{-5}$	$4.27{\cdot}10^{-4}$	$1.45{\cdot}10^{-3}$

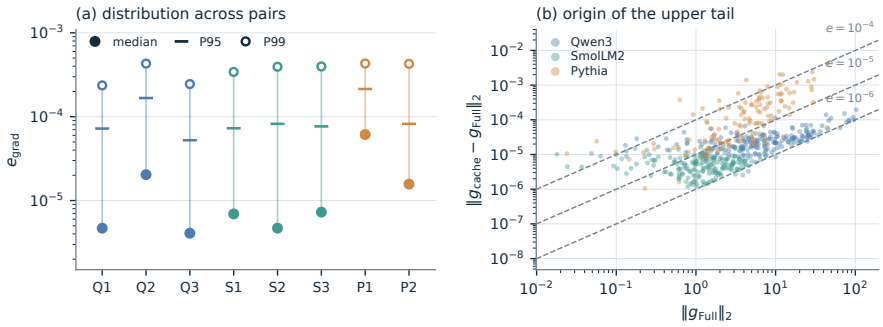


Figure 6: **Same-point gradient fidelity.** (a) Median, P95 and P99 of $e_{\text{grad}} = \|g_{\text{cache}} - g_{\text{Full}}\|_2 / \|g_{\text{Full}}\|_2$ over 2,560 probes per pair from ten seeds. (b) The same discrepancy before normalization, plotted against the reference norm that e_{grad} divides by; dashed diagonals are lines of constant relative error. Panel (b) uses a separate audit subset of 64 probes per pair, so its medians differ slightly from (a). Full quantiles and per-seed spread are in Appendix C.5.

Seed stability. Taking the median separately within each of the ten seeds gives the spread below. Across-pair variation in median discrepancy is about $15\times$, whereas within-pair seed variation ranges from $1.15\times$ to $2.32\times$, so the pair is the dominant factor and a single-seed measurement would not have misplaced any pair.

Pair	per-seed median range	factor
S3	6.63×10^{-6}–7.64×10^{-6}	$1.15\times$
P1	5.84×10^{-5}–6.94×10^{-5}	$1.19\times$
S1	5.88×10^{-6}–7.93×10^{-6}	$1.35\times$
S2	3.78×10^{-6}–5.21×10^{-6}	$1.38\times$
P2	1.41×10^{-5}–2.15×10^{-5}	$1.53\times$
Q3	2.46×10^{-6}–5.35×10^{-6}	$2.17\times$
Q1	2.68×10^{-6}–6.12×10^{-6}	$2.29\times$
Q2	1.16×10^{-5}–2.69×10^{-5}	$2.32\times$

The probe-level audit subset. Figure 6(b) uses a separate set of 64 audit probes per pair, 512 in total, for which the reference norm and the unnormalized discrepancy were both recorded. Its per-pair medians run between 1.1 and 5.5 times below the corresponding 2,560-probe medians, so it is a smaller and differently drawn sample rather than a subsample of the same measurement, and we use it only for the relationship between the two quantities, not for quantile estimates. Reference norms in it span 1.8×10^{-2} to 103 and unnormalized discrepancies 1.1×10^{-6} to 4.1×10^{-3}. Decomposing $\log e_{\text{grad}} = \log \|g_{\text{cache}} - g_{\text{Full}}\|_2 - \log \|g_{\text{Full}}\|_2$ within each pair attributes a mean 86% of the variance to the reference norm, with a range of 78% to 111%; values above 100% arise where the two terms are mildly anticorrelated. Pooling the pairs instead attributes only 42% to the reference norm, because pooling reintroduces the genuine difference between pairs that the within-pair decomposition removes. That difference is what remains after the normalization effect is accounted for: restricting to the band of reference norms where P1 and Q1 overlap still leaves P1 with an absolute discrepancy 13.6 times larger, although the two pairs have the same declared widths.

C.6 Cost, payload and system measurements

Reuse across students and over time. When several students share prefixes and targets, independent projected caches accumulate as $b_M \sum_j d_{S,j}$ while one teacher-state cache serves all of them, unless the students share a basis; scheduled output directions are covered in advance by the rule of Section 2 at $b_M r$ additional bytes. Whether preparation is amortized depends on how often each entry is revisited, $T_{\text{prepare}} + R\,T_{\text{visit}}$, and on-policy training refreshes prefixes and recovers little of it (Agarwal et al., 2024). The byte comparison also depends on which target is stored, sharply enough that the ordering can reverse; Appendix E.5 accounts one cross-tokenizer setting under two targets and two encodings.

Table 6: **Supervision-path cost under GPU-resident targets.** Forward and backward time of the loss path and peak memory, with all four targets resident on the device. All four rows are our own implementations of these representations, measured on one machine under a matched configuration.

	Forward (ms)	Backward (ms)	Peak memory (GiB)
Full-cached	0.096	0.447	4.568
MOMENT	0.090	0.367	4.197
Hidden-ref	0.200	0.378	4.568
Chunked-vocabulary route	0.973	29.672	3.468

Table 7: **Supervision transfer cost for CPU-resident offline caches.** Host-to-device transfer, loss computation, and total supervision-path time for a batch of N positions; S1 and Q3 are the pairs of Figure 6. The last column is the total relative to the cached-full-target row of the same block, and measures the supervision path alone rather than a full training step.

Pair	N	Representation	Batch payload	H2D (ms)	Loss (ms)	Total (ms)	vs FULL
S1	512	Full	96.00 MiB	5.020	0.551	5.571	1.0×
		Hidden	2.00 MiB	0.123	0.598	0.721	7.7×
		MOMENT	1.12 MiB	0.074	0.497	0.570	9.8×
	2048	Full	384.00 MiB	18.094	1.187	19.281	1.0×
		Hidden	8.00 MiB	0.421	1.079	1.500	12.9×
		MOMENT	4.50 MiB	0.226	0.687	0.913	21.1×
Q3	512	Full	296.75 MiB	14.574	0.859	15.433	1.0×
		Hidden	4.00 MiB	0.224	0.744	0.968	15.9×
		MOMENT	2.00 MiB	0.117	0.479	0.595	25.9×
	2048	Full	1,187.00 MiB	57.473	1.131	58.603	1.0×
		Hidden	16.00 MiB	0.893	1.195	2.088	28.1×
		MOMENT	8.00 MiB	0.429	0.806	1.235	47.5×

System measurements. Four different measurements are kept apart here: a device-resident loss-path microbenchmark, a host-to-device transfer decomposition, a head-only cost measurement, and a local per-step timing suite over two students, two head structures and four supervision routes. All are our own implementations measured on one machine. None of them establishes a training-time speedup, and they are not combined.

The device-resident microbenchmark isolates the loss path, where the projection saves backward time and peak memory against a cached full target, and a teacher-state route pays additional forward time to rebuild the V-length target at every visit. The transfer decomposition shows the payload reduction being realized when caches live in pinned host memory and the V-length target no longer moves, with the advantage growing in the number of positions transferred. Both are supervision-path measurements, so the ratios they give are not training-step ratios and the main text quotes none of them as a speedup. The rows labelled as a teacher-state route are our own implementation of that representation rather than the fused reference system of Pai et al. (2026).

Table 8: **Head-only cost at matched configuration.** Peak memory and time for the three target paths with only the output head resident. Head weights are listed separately to show that the memory difference is not simply one head.

Pair	Full GiB / ms	Moment GiB / ms	Hidden GiB / ms	Δpeak GiB	head weights GiB
Q2	10.76 / 173.65	10.75 / 171.75	11.78 / 173.32	1.03	0.725
Q3	5.03 / 123.35	5.01 / 119.82	6.48 / 122.39	1.47	1.160
S1	1.69 / 117.37	1.68 / 115.98	1.97 / 116.93	0.29	0.188

C.7 PAYLOAD ACCOUNTING

Table 9: Payload frontier across the eight pairs of Figure 6, ordered by ratio, with $b_M = 4$ for FP32 moments and $b_H = 2$ for native BF16 activations. Ratios above one favour the projected cache.

Pair	d_T	d_S	MOMENT B/pos	HIDDEN B/pos	HIDDEN/MOMENT
Q3	4,096	1,024	4,096	8,192	2.000
S1	2,048	576	2,304	4,096	1.778
Q1	2,560	1,024	4,096	5,120	1.250
P1	2,560	1,024	4,096	5,120	1.250
S2	2,048	960	3,840	4,096	1.067
S3	960	576	2,304	1,920	0.833
Q2	2,560	2,048	8,192	5,120	0.625
P2	2,560	2,048	8,192	5,120	0.625

For a Qwen student with $V = 151{,}936$ and $d_S = 1{,}024$, an FP32 full target costs $607{,}744$ bytes per position against $4{,}096$ for the projection, a ratio of 148.4. A BF16 full target costs $303{,}872$ bytes and gives 74.1875. We quote the FP32 figure in the main text, since that is the precision the substitution is measured at throughout.

Two students of widths $1{,}024$ and $2{,}048$ on common prefixes with identical teacher targets need $12{,}288$ bytes per position of separate FP32 moment caches against $5{,}120$ for one shared teacher-hidden cache. This assumes the same teacher, temperature and token coordinates, and independently generated trajectories do not satisfy it. A basis shared across student heads would reduce the projected total at the cost of its own accounting, and neither the naive sum nor the hidden width is an information-theoretic lower bound.

C.8 PRECISION DIAGNOSTICS

The FP64 fixed-head and expanded-head algebraic checks have residuals near 10^{-15}. Storing the projected cache in BF16 halves its payload and produces a relative error of 1.98×10^{-3}, three orders of magnitude above the FP32 medians of Figure 6, which is why the main comparison retains FP32 coordinates. That figure is not attributed here to any single cause: accumulating V terms, quantizing a freshly computed quantity, cancellation between $W_0^\top p - \mu$ and $W_0^\top (p - q)$, and the conditioning of the gradient are separable effects and we have isolated none of them. The native BF16 teacher activation is by contrast an intermediate of the teacher's own computation and undergoes no additional quantization step.

Export arithmetic is separate again. The FP32 save-and-merge check is exact in the reported test, whereas BF16 export changes the additional UA logit contribution by a relative amount 0.030, with cosine similarity 0.9996.

D SCOPE

The results hold at a shared token coordinate system, with a frozen teacher and a target fixed for each state. The eight pairs are within-family, so nothing here evaluates cross-tokenizer distillation; applying the construction on top of an alignment method's output would require the alignment target to be a fixed cross-entropy against a fixed affine output structure, and would otherwise need a fresh derivation.

Detaching a target that is recomputed from θ is weaker than holding it fixed across the domain, so the guarantees apply to the conditional update at an already-sampled prefix rather than to gradients taken through the sampling distribution.

E SAME-BUDGET BASELINES AND LEARNING INTERVALS

E.1 SAME-BUDGET BASELINES, EACH WITH ITS OWN PROTOCOL

The probability-reweighted SVD arm is implemented from the published construction, with $\bar{p} = \mathbb{E}[q_T]$ and $C = \mathbb{E}[h_T h_T^\top]$ estimated on a calibration split; it is the reweighted variant, not a naive unembedding SVD. Its measurements on the Q3 pair use a calibration set of $4{,}069$ rows and a measurement set of 256 positions, with the calibration split kept disjoint from the measured positions.

Budget	Coordinate dtype	Bytes/pos.	Rank used	TV to q	Rel. grad. error
0.5x	FP32	2,048	512	0.3858	0.2319
0.5x	BF16	2,048	1,024	0.2896	0.1719
1.0x	FP32	4,096	1,024	0.2897	0.1720
1.0x	BF16	4,096	2,048	0.1732	0.0992
2.0x	FP32	8,192	2,048	0.1733	0.0993
2.0x	BF16	8,192	4,096	0.0151	0.0037

The pattern across budgets is the informative part. Halving the budget barely changes the gradient error, and doubling it improves the error by more than an order of magnitude only once the rank reaches the full cap, where the reconstruction becomes near-exact. A low-rank reconstruction of the teacher distribution therefore does not degrade gracefully towards the projection's accuracy: it is either close to exact or off by a large factor, because the discarded spectrum is not aligned with $\mathcal{V}_s$. Renormalized Top-k and the teacher-state replay are discussed with their own conditions in Appendix C. These measurements use the SVD arm's own calibration and measurement sets, so they are not interchangeable with the same-point errors of Section 3.3; the comparison on one common footing is the matched-budget experiment of Appendix E.2.

Storage precision of the exact routes. Storing an exact representation in BF16 rather than FP32 halves it and raises the relative gradient error to the 10^{-3} level; the arithmetic behind that figure, and the export check, are in Appendix C.8.

E.2 THE MATCHED-BUDGET COMPARISON

Table 2 compares four substituted representations on Q1, S1 and P2 under one protocol: the same prefixes, the same optimizer configuration, the same supervision mask, 128 updates, and a byte budget per position held equal across MOMENT, Top-k and the SVD cache. The teacher state is not budget-matched, since its size is fixed by the teacher width; on these three pairs it happens to fall within a factor of two of the others. Renormalized Top-k keeps the k largest probabilities and renormalizes; the SVD cache stores a low-rank reconstruction of the teacher distribution at the same byte count, in the probability-reweighted form of Appendix E.

Three quantities are recorded per arm. The trajectory gradient error is e_{grad} of Eq. (14) evaluated at the parameters that arm has itself reached, so unlike the same-point measurement of Section 3.3 it compounds whatever divergence has already occurred. The divergence $D(t)$ is the normalized distance between the arm's trainable parameters and those of the Full-target run at the same update. Held-out NLL is WikiText per-token NLL as in Section 3.4.

Pair	Representation	NLL seeds	Grad. seeds	NLL at 128	$D(128)$
Q1	MOMENT	3	3	2.7327	0.0490
	Teacher state	3	3	2.7219	0.0180
	Top-k	3	3	2.7347	0.0691
	SVD cache	3	3	2.9042	0.3490
S1	MOMENT	2	3	2.9099	0.0000
	Teacher state	2	3	2.9101	0.0148
	Top-k	2	3	2.9193	0.0611
	SVD cache	2	3	3.2571	0.3804
P2	MOMENT	3	3	2.7055	0.0083
	Teacher state	3	3	2.7055	0.0261
	Top-k	3	3	2.7039	0.1712
	SVD cache	3	3	2.7287	0.3446

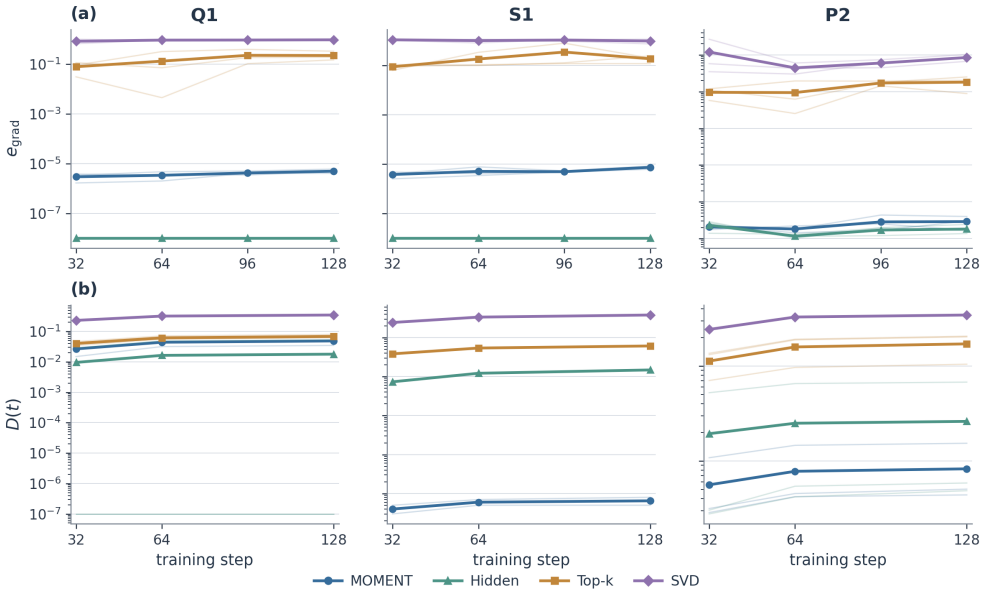


Figure 7: Per-checkpoint view of the same runs at 32, 64, 96 and 128 updates. (a) e_{grad} against the Full-target gradient; (b) divergence $D(t)$ of each arm's parameters from the Full-target run. Faint lines are individual seeds, markers their median. The ordering of Figure 3 holds at every checkpoint, and neither exact route trends upward over the horizon.

The S1 rows aggregate two NLL seeds rather than three, so its NLL differences carry correspondingly less resolution; gradient measurements use three seeds throughout. The $D(128)$ column is a mean over seeds, and on S1 and P2 the exact routes sit low enough that the per-seed spread in Figure 7(b) reaches below the mean by more than an order of magnitude; we quote the mean here and plot the seeds there. A generated-answer readout was collected alongside these runs but its scoring audit is unfinished, and we do not report it.

E.3 THE EIGHT-PAIR SURVEY AT 64 UPDATES

This is an earlier and broader protocol than the matched-budget comparison of Table 2: eight pairs rather than three, 64 updates rather than 128, and no lossy arms at matched bytes. Its purpose was coverage across teacher–student shapes. Full supervision improves held-out NLL by 0.0191 to 0.0908 nats per token across the eight pairs, and the largest substitution difference is 0.00110, two orders of magnitude below the smallest of those gains. The two experiments were run under different configurations and their per-pair values are not comparable term by term; Q1 in particular ends $+0.0011$ here and $+0.0108$ in Table 2.

Table 11 carries three of these pairs to 512 updates as checkpoints of single runs.

Table 10: **Language learning under supervision replacement, eight configurations at 64 updates.** Held-out WikiText NLL per token, computed from unrounded per-run values. $G_{\text{Full}} = \text{Warm} - \text{Full}$ is the gain of full supervision, so larger is better; $\Delta_{\text{replace}} = \text{MOMENT} - \text{Full}$ is the effect of substituting the representation, so nearer zero is better, in units of 10^{-3} nats per token. Intervals are 90% t intervals on per-seed paired differences over three training seeds; the warm start is one base-student value and contributes no seed variance. The Hidden and Top-k arms completing the 17 endpoint comparisons are in Appendix E.3.

Pair	Teacher → Student	Warm	Full	MOMENT	$G_{\text{Full}} \uparrow$		Δ_{replace} (10^{-3})	
					mean	90% CI	mean	90% CI
Q1	4B → 0.6B	2.8002	2.7101	2.7112	0.0901	0.0836–0.0966	+1.10	−1.01–+3.20
Q2	4B → 1.7B	2.4855	2.4383	2.4381	0.0472	0.0444–0.0500	−0.24	−0.78–+0.29
Q3	8B → 0.6B	2.8002	2.7094	2.7097	0.0908	0.0876–0.0940	+0.28	+0.00–+0.56
S1	1.7B → 135M	3.0092	2.9550	2.9552	0.0542	0.0521–0.0564	+0.21	−0.60–+1.03
S2	1.7B → 360M	2.6971	2.6781	2.6784	0.0191	0.0167–0.0214	+0.33	−0.11–+0.76
S3	360M → 135M	3.0092	2.9727	2.9726	0.0366	0.0359–0.0372	−0.05	−0.37–+0.28
P1	2.8B → 410M	3.0802	2.9997	2.9997	0.0805	0.0772–0.0837	+0.01	−1.92–+1.95
P2	2.8B → 1.4B	2.7443	2.7071	2.7072	0.0372	0.0321–0.0422	+0.04	−0.28–+0.37

Table 11: **Three pairs evaluated at five checkpoints of one run.** Held-out WikiText NLL per token, three training seeds each. These are checkpoints of single 512-update runs, so the 64-update rows are not the independently trained endpoints of Table 10. $\Delta_{\text{replace}} = \text{MOMENT} - \text{Full}$, with 90% paired training-seed t intervals.

Pair	Updates	Full	MOMENT	Δ_{replace}	90% CI
Q1	32	2.721342	2.720927	−0.000415	[−0.00132, +0.00049]
	64	2.709590	2.709666	+0.000076	[−0.00050, +0.00065]
	128	2.724532	2.724524	−0.000008	[−0.00101, +0.00100]
	256	2.761552	2.760481	−0.001072	[−0.00291, +0.00076]
	512	2.785096	2.786761	+0.001665	[−0.00155, +0.00488]
S1	32	2.987545	2.987296	−0.000249	[−0.00112, +0.00062]
	64	2.958310	2.958191	−0.000119	[−0.00141, +0.00118]
	128	2.930990	2.930931	−0.000059	[−0.00101, +0.00089]
	256	2.919484	2.919618	+0.000133	[+0.00005, +0.00022]
	512	2.943666	2.943338	−0.000328	[−0.00157, +0.00092]
P2	32	2.718395	2.717838	−0.000557	[−0.00123, +0.00011]
	64	2.708305	2.707936	−0.000369	[−0.00102, +0.00028]
	128	2.702665	2.702633	−0.000032	[−0.00116, +0.00110]
	256	2.720717	2.719871	−0.000846	[−0.00429, +0.00260]
	512	2.731748	2.730244	−0.001503	[−0.00673, +0.00373]

The seventeen endpoint substitution comparisons, MOMENT and Hidden on all eight pairs and Top-k on P1, are as follows.

Each row is one arm against Full on one pair at 64 updates, as a 90% t interval on per-seed paired differences over three training seeds. Top-k has a language readout on P1 only.

Pair	Arm	Mean	90% CI	Excludes 0	Inside ± 0.01
Q1	MOMENT	$+0.001096$	$[-0.001007, +0.003198]$	no	yes
Q1	Hidden	$+0.000091$	$[-0.000550, +0.000733]$	no	yes
Q2	MOMENT	-0.000242	$[-0.000778, +0.000293]$	no	yes
Q2	Hidden	$+0.000026$	$[-0.000377, +0.000430]$	no	yes
Q3	MOMENT	$+0.000284$	$[+0.000003, +0.000564]$	yes	yes
Q3	Hidden	$+0.000120$	$[-0.000280, +0.000519]$	no	yes
S1	MOMENT	$+0.000214$	$[-0.000604, +0.001032]$	no	yes
S1	Hidden	$+0.000478$	$[-0.000385, +0.001342]$	no	yes
S2	MOMENT	$+0.000325$	$[-0.000107, +0.000757]$	no	yes
S2	Hidden	$+0.000149$	$[-0.000274, +0.000573]$	no	yes
S3	MOMENT	-0.000047	$[-0.000374, +0.000280]$	no	yes
S3	Hidden	-0.000011	$[-0.001165, +0.001142]$	no	yes
P1	MOMENT	$+0.000015$	$[-0.001924, +0.001953]$	no	yes
P1	Hidden	-0.001333	$[-0.004381, +0.001714]$	no	yes
P1	Top-k	$+0.002428$	$[+0.000477, +0.004378]$	yes	yes
P2	MOMENT	$+0.000045$	$[-0.000279, +0.000368]$	no	yes
P2	Hidden	$+0.000121$	$[-0.000354, +0.000596]$	no	yes

E.4 SAME-TOKENIZER GENERATED-ANSWER ACCURACY: ALL MAIN UNITS

Each row is one pre-specified comparison unit: a student, a head structure (*frozen* keeps the output head fixed, *ext* adds trainable output directions), a substituted method against its own Full reference, an evaluation slice, and a prompting protocol (A is zero-shot with no stop sequence, B is eight-shot with a specified stop). Decoding is greedy in BF16 with a 2048-token limit. Differences are method minus Full in percentage points over three training seeds, with 90% intervals, read against the pre-specified tolerance of one percentage point. Answers are scored by strict match on the final numeric answer, which is the convention under which all 216 per-seed endpoints reconcile with the per-run generation logs; a more permissive variant that accepts partial numeric matches agrees with those logs on only 13 of the 216, so we do not use it.

Student	Head	Method	Evaluation set	Prompt	Mean (pp)	90% CI (pp)
Qwen3-0.6B	ext	MomentPlus	GSM8K test	A	-0.6065	$[-1.4447, +0.2316]$
Qwen3-0.6B	ext	MomentPlus	GSM8K test	B	$+0.2022$	$[-0.4411, +0.8455]$
Qwen3-0.6B	ext	MomentPlus	SVAMP	A	-0.0333	$[-0.9618, +0.8952]$
Qwen3-0.6B	ext	MomentPlus	SVAMP	B	$+0.2667$	$[-1.2998, +1.8331]$
Qwen3-0.6B	frozen	Moment	GSM8K test	A	$+0.0505$	$[-0.4816, +0.5827]$
Qwen3-0.6B	frozen	Moment	GSM8K test	B	$+0.0505$	$[-0.3603, +0.4614]$
Qwen3-0.6B	frozen	Moment	SVAMP	A	-0.3000	$[-0.7460, +0.1460]$
Qwen3-0.6B	frozen	Moment	SVAMP	B	$+0.5667$	$[-0.9444, +2.0777]$
Qwen3-0.6B	ext	MomentPlus	GSM8K test	A	-0.1516	$[-2.0816, +1.7783]$
Qwen3-0.6B	ext	MomentPlus	GSM8K test	B	$+0.0253$	$[-0.3856, +0.4361]$
Qwen3-0.6B	ext	MomentPlus	SVAMP	A	-0.5667	$[-1.2480, +0.1147]$
Qwen3-0.6B	ext	MomentPlus	SVAMP	B	-0.3333	$[-1.7371, +1.0704]$
Qwen3-0.6B	frozen	Moment	GSM8K test	A	-0.9098	$[-1.2479, -0.5716]$
Qwen3-0.6B	frozen	Moment	GSM8K test	B	-0.1516	$[-1.7531, +1.4499]$
Qwen3-0.6B	frozen	Moment	SVAMP	A	$+0.0333$	$[-1.0505, +1.1172]$
Qwen3-0.6B	frozen	Moment	SVAMP	B	-0.3667	$[-0.7909, +0.0576]$
Qwen3-1.7B	ext	MomentPlus	GSM8K test	A	-0.8845	$[-2.2866, +0.5176]$
Qwen3-1.7B	ext	MomentPlus	GSM8K test	B	-0.2780	$[-2.0227, +1.4667]$
Qwen3-1.7B	ext	MomentPlus	SVAMP	A	-0.4333	$[-1.6760, +0.8093]$
Qwen3-1.7B	ext	MomentPlus	SVAMP	B	-0.0333	$[-0.7935, +0.7269]$
Qwen3-1.7B	frozen	Moment	GSM8K test	A	$+0.8592$	$[+0.3427, +1.3758]$
Qwen3-1.7B	frozen	Moment	GSM8K test	B	-0.1516	$[-0.6125, +0.3092]$
Qwen3-1.7B	frozen	Moment	SVAMP	A	-0.1000	$[-1.7082, +1.5082]$
Qwen3-1.7B	frozen	Moment	SVAMP	B	$+0.7000$	$[-0.3115, +1.7115]$

Of the 72 evaluated units, 28 differ from their Full reference by less than a percentage point. Under the intersection–union rule declared for this readout a single global claim would require all of them to, so no global claim is made. The per-unit spread is informative in its own right: the frozen-head

Table 12: **Cross-tokenizer aligned distillation, generated-answer accuracy.** Qwen3-4B with an adapter into SmolLM2-360M, trained for 256 updates on a top-256 aligned bank, and evaluated by greedy generation on a 512-question set held out from GSM8K *train* with a 512-token limit. Accuracy is computed from correct counts over three training seeds and the spread is a sample standard deviation. Absolute accuracies are low at this model scale, so the comparison of interest is between arms rather than against published GSM8K numbers.

Method	Seeds	Accuracy (%) ↑	Reference-answer NLL/token ↓
Warm	1	3.125	0.743329
Full-aligned	3	6.771 ± 0.923	0.784707
Hybrid-Moment	3	6.120 ± 0.226	0.784766
Teacher-sequence SFT	3	7.031 ± 0.195	0.763865

Table 13: **Cross-tokenizer payload depends on which target and which encoding are being compared.** Layout-accounted totals over the same 256 positions. The top-256 bank is the target used for the training run of Table 12; the full current-step vocabulary is a wider target accounted here for comparison. Which representation is smaller reverses between the two targets, so the ratio is a property of the target and encoding being compared, not of the method.

Target	Encoding	Full bytes	Hybrid bytes	Hybrid/Full
Top-256 bank, as trained	int32 id list	417,576	1,193,364	2.858
Top-256 bank, as trained	per-position best candidate	414,504	1,190,292	2.872
Full current-step vocabulary (audit)	int32 id list	74,784,432	38,376,792	0.513
Full current-step vocabulary (audit)	per-position best candidate	38,940,112	2,532,472	0.065

units, where the projection is exact by construction, cluster far more tightly than the extended-head units, and protocol B is tighter than protocol A throughout, which is what one expects when a stop sequence removes a source of decoding variance.

E.5 Cross-tokenizer aligned distillation

The constructions of Section 2 assume teacher and student share a tokenizer, so that a target defined on the teacher's vocabulary can be projected through the student's head. When the tokenizers differ, the target must first be realigned onto the student's vocabulary, and the realignment used here restricts it to a bank of top-256 candidates per position. That restriction is lossy in a way the projections are not, so this setting is an application of the criterion beyond the regime the theorems cover rather than evidence for them. We report it because it is where the construction is most likely to be wanted and where its present limits are visible.

A Qwen3-4B teacher supervises a SmolLM2-360M student through the aligned target, trained for 256 updates and evaluated by greedy generation on a 512-question set held out from GSM8K train. Full-aligned supervision raises generated-answer accuracy from 3.125% to 6.771%, a gain of 3.646 percentage points with a seed-level interval of $[1.35, 5.94]$, so cross-tokenizer supervision through an aligned target does train. The hybrid representation reaches 6.120%, trailing Full by 0.651 percentage points with a per-seed interval of $[-3.52, -0.59]$ that excludes zero. The gap is small in absolute terms and real, and it marks the limit of the present construction: restricting an aligned target to a bank of candidates is a lossy step, unlike the projections of Section 2, and it costs something measurable on the task. Teacher-sequence supervised fine-tuning reaches 7.031%, within seed variation of Full-aligned. Reference-answer NLL rises for every trained arm even as accuracy rises, which is consistent with the arms moving away from the reference phrasing while answering more questions correctly.

Payload in this setting depends sharply on which target is being stored, sharply enough that the ordering between the two representations reverses. Table 13 accounts the same 256 positions under two targets and two encodings. Against the top-256 aligned bank that the run of Table 12 trained on, the hybrid target is *larger* than the full-vocabulary alternative by a factor of about 2.86, with the two encodings agreeing at 2.858 and 2.872. Against a full current-step vocabulary target the

ordering reverses, to 0.513 and 0.065. The reversal is the point: a bank of a few hundred candidates is cheap to write densely and expensive to write as explicit identifiers, so the favourable regime for the aligned representation is the wide target, not the narrow one the current run used. These are byte totals accounted from real support sizes under a declared layout rather than measured transfers. One further observation from the wide target is that its newly explicit entries carry about 4.4×10^{-4} of the mass; that makes the tail cheap to represent but not removable, since explicit zero-probability entries still define the support over which the remainder is normalized.

E.6 Local per-step timing

One optimizer step measured as read plus forward and backward plus optimizer update, at 1024 supervision positions per step, three repeats per configuration after four excluded warmup steps. All records are for one training state; there is no second-state coverage, and these are local implementations on local paths, not a comparison against any fused reference system.

Student / structure	Route	Read (ms)	Fwd+bwd (ms)	Total (ms)	Peak (GiB)
Qwen3-0.6B / ext	Full	92.24	3417.86	3516.21	5.160
	MomentPlus	2.89	3463.75	3473.00	4.570
	Teacher hidden	3.25	3462.06	3471.52	5.340
Qwen3-0.6B / frozen	Full	90.84	3436.28	3533.50	5.120
	Moment	2.38	3541.77	3550.76	4.530
	Teacher hidden	3.35	3367.73	3377.09	5.310
Qwen3-1.7B / ext	Full	90.19	3557.91	3654.56	10.750
	MomentPlus	3.73	3385.80	3395.41	10.160
	Teacher hidden	3.23	3498.54	3507.95	10.930
Qwen3-1.7B / frozen	Full	86.55	3490.89	3583.69	10.710
	Moment	3.20	3607.25	3617.12	10.120
	Teacher hidden	3.18	3525.30	3534.56	10.890

F Measuring $\dim \mathcal{V}_s$: procedure, numerical floor, and spectra

F.1 Rank estimation procedure

The reachable centered logit space of Eq. (4) is

$$\mathcal{V}_s = P \operatorname{span}\{z_\theta(s) - z_{\theta_0}(s) : \theta \in \Theta\}, \qquad P = I - \mathbf{1}\mathbf{1}^\top / V. \tag{43}$$

For a frozen affine head $z_\theta = W_0 h_\theta$. One cache basis serves every prefix, so the relevant object is the joint space of Eq. (8); writing

$$\mathcal{U} = \operatorname{span}\{h_\theta(s) - h_{\theta_0}(s) : \theta \in \Theta,\ s \in \mathcal{S}\} \subseteq \mathbb{R}^{d_S}, \tag{44}$$

which pools the reachable hidden-state directions over the measurement set, the quantity of interest is $\dim \mathcal{V}_{\mathcal{S}} = \operatorname{rank}(PW_0\mathcal{U})$, and centering removes at most one dimension:

$$\operatorname{rank}(PW_0\mathcal{U}) = \operatorname{rank}(W_0\mathcal{U}) - \mathbf{1}\{\mathbf{1} \in \operatorname{col}(W_0\mathcal{U})\}. \tag{45}$$

The head W_0 has full column rank for all six students, so $\operatorname{rank}(W_0\mathcal{U}) = \operatorname{rank}(\mathcal{U})$ and the measurement reduces to estimating $\operatorname{rank}(\mathcal{U})$. The all-ones logit shift is outside $\operatorname{col}(W_0\mathcal{U})$ for every student, so the indicator in Eq. (45) is zero throughout and centering costs no dimension. Pooling over prefixes is what makes this the joint space: a rank deficit at one prefix is invisible here if another prefix supplies the missing direction, which is the correct behaviour for a shared basis and the wrong behaviour for a per-prefix one.

Sampling. Perturbations use the LoRA target modules of the training experiments, with rank $r = 16$, $\alpha = 32$ and dropout 0, and only the adapter trainable. The base point θ_0 takes $B = 0$, so the adapter contributes exactly zero there and $\Delta h_\theta = h_\theta - h_{\theta_0}$ is a difference of two states of the same model rather than of two models. We then draw $B_{ij} \sim \mathcal{N}(0, \sigma^2)$ with A at the trainer's initialization $\mathcal{N}(0, 1/d_{\text{in}})$, at scales

$$\sigma \in \{10^{-12}, 10^{-6}, 10^{-3}, 10^{-1}, 1\}. \tag{46}$$

The headline rank is taken at $\sigma = 1$; $\sigma = 10^{-12}$ provides the numerical-floor reference of Appendix F.2; the intermediate scales give the robustness sweep of Appendix F.3. All draws use fixed seeds.

Measurement set and construction. Eight WikiText blocks contribute 32 positions each, for 256 positions, and each scale uses $n_\theta = 10$ perturbation draws, with forward passes in FP32. Stacking every difference gives

$$H = \left[\Delta h^{(1)}_{\theta_1}, \ \ldots, \ \Delta h^{(256)}_{\theta_{n_\theta}}\right] \in \mathbb{R}^{d_S \times 2560}. \tag{47}$$

H is cast to FP64 before an economy SVD, so the decomposition itself does not limit resolution; the limit is the FP32 forward pass that produced the entries. A direction is counted as resolved when

$$\sigma_i > \tau, \qquad \tau = 10 \times \text{floor}, \tag{48}$$

and the estimate of $\mathrm{rank}(\mathcal{U})$ is the number of resolved directions. This is a statement about what the measurement can distinguish from its own noise, and the sweep below shows the count is unchanged for any factor between 10 and 222.

The centering check. Taking a thin QR of the resolved column space gives Q_r, and we solve $\min_c \|\mathbf{1} - W_0 Q_r c\|_2$. The relative residual is 0.028–0.133 across the six students, so $\mathbf{1} \notin \mathrm{col}(W_0\mathcal{U})$ and the indicator in Eq. (45) is zero: centering does not cost a dimension for any of them.

Why an observed rank is a lower bound. Each column of H is a legal reachable difference, so the columns lie in $\mathcal{U}$ and

$$\mathrm{rank}(H) \ \le \ \dim\mathcal{U} \ \le \ d_S. \tag{49}$$

The upper bound is structural. Observing $\mathrm{rank}(H) = d_S$ therefore makes the two bounds meet, giving $\dim\mathcal{U} = d_S$ and, with the centering check, $\dim\mathcal{V}_\mathcal{S} = d_S$. Two limits of this argument should be kept in view. The rank on the left is read at finite precision, so what the meeting of the bounds establishes is that the dimension is attained up to the resolution of an FP32 forward pass, not a certification in exact arithmetic; the margins reported below are what makes that reading reasonable rather than marginal. And the method is asymmetric: sampling can show that the full dimension is attained but cannot show that a dimension is missing, since an unsampled or unresolved direction is indistinguishable from an absent one. A shortfall therefore yields only $\mathrm{rank}(H)$ as a lower bound together with the structural upper bound d_S.

F.2 The numerical floor

At $\sigma = 10^{-12}$ the adapter contribution to a hidden state is of order 10^{-12} relative to the state itself, which is below FP32 resolution, so the differences recorded in that run carry no signal and we treat them as pure arithmetic noise. This gives an empirical resolution limit for each student,

$$\text{floor} = \sigma_{\max}\big(H_{\sigma=10^{-12}}\big), \tag{50}$$

measured through exactly the same code path, model and positions as the headline run, which is why it is preferred to a machine-epsilon estimate.

Student	FP32 numerical floor
SmolLM2-135M	5.44×10^{-4}
SmolLM2-360M	6.17×10^{-4}
Qwen3-0.6B	1.65×10^{-3}
Qwen3-1.7B	1.43×10^{-3}
Pythia-410M	5.95×10^{-2}
Pythia-1.4B	1.75×10^{-2}

The measured floor is 30 to 100 times larger for the two Pythia students than for the other four. We report this as a property of the measurement rather than explaining it, since attributing it to a particular part of the forward computation would need a controlled comparison we have not run. The multiplier of ten in Eq. (48) is an order-of-magnitude margin fixed before the measurement, and it is a rule for certifying a direction numerically, not a definition of the rank.

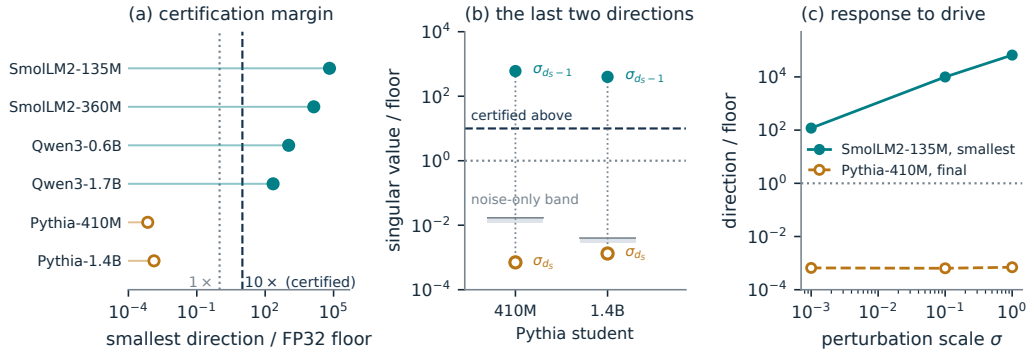


Figure 8: **Measured reachable dimension under isotropic LoRA perturbations.** Singular values are normalized by the FP32 numerical floor of Appendix F.2, measured on an otherwise identical run at $\sigma = 10^{-12}$; the dotted line is one floor and the dashed line the certification threshold at ten. (a) The smallest direction of each student. The four SmolLM2 and Qwen3 students resolve all d_S directions with margins of 2.22×10^2 to 6.62×10^4, while the final Pythia directions fall below the floor itself. (b) For the two Pythia students the penultimate direction stays several hundred floors above noise and the final one lies inside the band the noise-only spectrum occupies at that index. (c) The smallest resolved direction of SmolLM2-135M grows with the perturbation that excites it, whereas the final Pythia direction stays pinned to the floor across three orders of magnitude, which is the behaviour of a direction the measurement cannot resolve rather than of one that is absent.

Threshold robustness. Among the four resolved students the weakest separation is Qwen3-1.7B at 2.22×10^2 floors, so every multiplier between 10 and 222 certifies all d_S directions for all four; the verdict does not depend on the particular choice. The corresponding figures for the final Pythia directions are 7.0×10^{-4} floors for Pythia-410M and 1.3×10^{-3} floors for Pythia-1.4B. These lie below the noise level itself, so no reduction of the threshold recovers the last dimension on any principled ground: a rule permissive enough to admit those directions would also admit the noise.

F.3 Spectra and the Pythia ambiguity

Student	d_S	$\dim \mathcal{V}_s$	$\sigma^+_{\min}$	floor	separation
SmolLM2-135M	576	576	36.0	5.44×10^{-4}	6.62×10^4
SmolLM2-360M	960	960	8.35	6.17×10^{-4}	1.35×10^4
Qwen3-0.6B	1024	1024	1.77	1.65×10^{-3}	1.07×10^3
Qwen3-1.7B	2048	2048	0.318	1.43×10^{-3}	2.22×10^2
Pythia-410M	1024	1023–1024	4.19×10^{-5}	5.95×10^{-2}	unresolved
Pythia-1.4B	2048	2047–2048	2.32×10^{-5}	1.75×10^{-2}	unresolved

Here $\sigma^+_{\min}$ is the smallest singular value at the headline scale: for the resolved students it is the smallest certified direction, and for the two Pythia students it is the final direction, which is not certified. The gap is abrupt rather than gradual. The penultimate Pythia directions are $\sigma_{d_S-1} = 35.8$ for Pythia-410M and 6.98 for Pythia-1.4B, or about 602 and 398 floors, so the spectrum is comfortably resolved up to the last index and then drops by six orders of magnitude in one step.

The final directions sit at 4.19×10^{-5} and 2.32×10^{-5}, that is 7.0×10^{-4} and 1.3×10^{-3} floors. The noise-only run at the same index gives roughly 1.2×10^{-2} to 1.7×10^{-2} floors for Pythia-410M and 3×10^{-3} to 4×10^{-3} floors for Pythia-1.4B, so the final direction lies inside the band the pure-noise spectrum occupies there. Two situations are therefore indistinguishable in this measurement: a true rank of $d_S - 1$, and a true rank of d_S whose last singular direction is smaller than FP32 resolution. We report $\dim \mathcal{V}_S \in \{d_S - 1, d_S\}$ accordingly.

Perturbation-scale robustness. A genuine direction should grow with the perturbation that excites it, and this separates the two cases behaviourally even where the singular value alone cannot. For the final Pythia-410M direction, $\sigma_{\min}/\text{floor}$ is 6.6×10^{-4}, 6.4×10^{-4} and 7.0×10^{-4} at $\sigma = 10^{-3}$,

10^{-1} and 1: it stays pinned to the floor across three orders of magnitude of driving signal. The smallest direction of SmolLM2-135M moves over the same range from 119 to 9956 to 66208 floors. The absence of scaling is consistent with a direction that the measurement cannot resolve, and it is the reason we do not read the small singular value as evidence of a missing dimension.

F.4 Scope of this measurement

What was measured is a finite-precision estimate of $\dim \mathcal{V}_S$ through one sampling scheme, governed by Eq. (48); the object itself is defined by a student, a trainable family, a state and a set of prefixes. The estimate concerns the joint space, which is the right target because W_0 is one basis shared by every prefix; an individual $\mathcal{V}_s$ may be smaller, and nothing here bounds it from below. And the conclusion is about dimension, not about which basis to use: no fixed linear cache below d_S coordinates can serve every target at every prefix on the four resolved students, and caches outside that class fall outside Theorem 1(3). A different trainable family on the same student would define a different space.