THE VOCABULARY BOTTLENECK AS AN ESTIMATOR: RATE, BIAS, AND HIDDEN TRANSFER BETWEEN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-agent systems are beginning to exchange internal state instead of text to avoid both the latency and the information loss of autoregressive decoding. This creates an apparent trade-off between efficiency and safety, since latent state cannot be monitored or controlled the way text can. But the safety cost extends beyond readability. The discrete vocabulary also limits the bandwidth available for hidden coordination between agents and forces every training signal through a decoding rule that may faithfully or unfaithfully represent the teacher. To quantify what losing these constraints costs, we model the projection to vocabulary as an estimator of the sender's distribution, characterized by two properties: bandwidth per exchange and fidelity of the training signal. We test this framework at inference, where a single cache exchange replaces a decoded message, and at training, where the decoding rule is applied across a distillation corpus. At inference, a hidden signal planted in the sender's cache is unrecoverable from emitted text but recoverable almost perfectly (held-out $R^2 = 0.9996$, 182 bits for a 32-dimensional payload) through a trained cross-model bridge, and every projection-based narrowing we test retains at least half of that capacity, and those that reduce it cost two to three times as much as not communicating. At training, neither sampling nor argmax decoding blocks hidden trait transfer. Sampling recovers the teacher's full distribution over many examples, consistent with prior findings that transfer survives filtering and paraphrasing. Argmax decoding produces a systematically biased target that displaces the student from its teacher by factors of 51.9 to 689.7 across three models, a phenomenon we call mode drift, while, on Qwen2.5-7B, relocating the acquired disposition rather than suppressing it. Soft-target distillation avoids this displacement but tolerates far more training before degeneration, at 3.6% collapse versus 91.7% for argmax at matched budget.

1 INTRODUCTION

Language models increasingly communicate with other language models. The established channel is text: a student trains on a teacher's decoded outputs, and everything it receives has passed through a discrete vocabulary. An emerging alternative skips the vocabulary entirely. In systems such as Cache-to-Cache (Fu et al., 2026), one model writes its key-value cache into another through a trained projector, so what arrives is continuous internal state rather than tokens. A 2026 survey of eighteen such methods motivates the shift by the latency savings and the information that decoding otherwise discards (Liu, 2026).

Moving from text to continuous state removes the ability to monitor, filter, or control what one model sends to another, a concern that is widely discussed (Liu, 2026). We show that how much the vocabulary protects depends on how often the channel is used, and propose an estimator framework that explains why. The projection to vocabulary is a map from the sender's internal distribution to what the receiver sees, and this map has two properties that matter for safety. It limits the *rate* of any single exchange, capping how many bits one token can carry. And through the choice of decoding rule, it may introduce a systematic *distortion* that no amount of repetition corrects. Rate is defeated by averaging over many exchanges, while distortion survives.

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.

This separates two regimes that the field currently treats as one problem. At inference, a handoff between agents happens once, so rate is the only constraint. At training, a distillation corpus applies the same decoding rule thousands of times, so what remains is whether the decoding rule faithfully represents the teacher's distribution. An intervention designed for one regime should not be expected to hold in the other.

At inference, the rate constraint was doing real work. We plant a hidden signal in the sender's cache and measure what a linear probe recovers at the receiver. Through the vocabulary, the signal is destroyed: a probe recovers nothing from either sampled or argmax-decoded text. Through a trained cross-model bridge, the same probe recovers it almost perfectly, comparable to the untrained same-model baseline the bridge replaces. Training the bridge to make two models interoperate is enough to make the signal inseparable from the useful content, and no projection-based narrowing we test removes it affordably (Section 3).

At training, the rate constraint was not doing the work it appeared to. Averaging over many sampled tokens recovers the teacher's full distribution, consistent with evidence that subliminal trait transfer survives text-level filtering, paraphrasing, and greedy decoding (Cloud et al., 2026; Draganov et al., 2026; Gisler et al., 2026; Schrodi et al., 2026). Argmax instead produces a deterministic, sharpened target whose distance from the teacher is fixed regardless of corpus size. Training on argmax targets displaces the student from its teacher by factors of 51.9 to 689.7 across three models, a phenomenon we call mode drift, with magnitude ordered by the irreducible error the framework predicts. The distorted estimator does not block trait transfer. On Qwen2.5-7B it redirects the acquired disposition to a different destination at matched magnitude, an effect that category-level metrics misread as suppression (Sections 4.2–4.4).

Our contributions are:

- An organizing view, deliberately elementary, that treats the vocabulary bottleneck as an estimator with a rate and a bias, so that the number of transmissions decides which one a defense can act on (Section 2); the contributions below are empirical.
- The first measurement of covert capacity through a trained cross-model bridge: held-out $R^2 = 0.9996$ on a planted payload, with every projection-based narrowing method retaining at least half of that capacity, and those that reduce it costing two to three times as much as not communicating (Section 3).
- Evidence that neither sampling nor argmax decoding blocks trait transfer during distillation: argmax decoding displaces the student without suppressing transfer (mode drift), and soft-target distillation avoids the displacement at 3.6% collapse versus 91.7% for argmax at matched budget (Sections 4.2–4.4).

2 THE TRANSFER ESTIMATOR FRAMEWORK

This section formalizes rate and distortion. The results are elementary; their role is to fix what each decoding rule does to the training target, which Sections 3 and 4 then measure.

Notation. Fix a context c. We write p_T for the sender's next-token distribution and q_θ for the receiver's, with $s_v = \nabla_\theta \log q_\theta(v \mid c)$ the per-token score and e_v the one-hot vector at token v. Appendix C lists every symbol.

Assumption 1 (loss)**.** *The receiver is fit by cross-entropy against a target t, so $L(t) = -\sum_v t(v) \log q_\theta(v \mid c)$ and $\nabla_\theta L(t) = -\sum_v t(v)\, s_v$.*

The gradient is linear in the target. Everything below follows from that linearity and one definition.

Definition (transfer estimator). A *transfer estimator* R maps the sender's distribution to the target the receiver is trained against. We study four, each corresponding to a different communication practice.

- $R_{\mathrm{id}}(p) = p$: the sender's full distribution transmitted as continuous state. This is what a latent handoff such as Cache-to-Cache (Fu et al., 2026) provides.

Table 1: The mode estimator's irreducible error and the trait signal it has to carry, on five teachers under the digit-preference channel of Cloud et al. (2026). "agree" is the rate at which a trait-prompted and an unprompted sender select the same digit; its complement is the divergence set of Proposition 2. The Gemma-3-4B row rests on 22,759 digit positions against 241,644 to 539,572 for the others, so the rows are not equally precise.

teacher	$\bar{p}_1$	**agree**	$\|\Delta p\|_1$	$2(1-\bar{p}_1)/\|\Delta p\|_1$
Gemma-3-4B	0.8176	73.8%	0.4805	**0.76**
Qwen2.5-7B	0.7163	83.5%	0.2949	**1.92**
Qwen2.5-14B	0.6247	81.4%	0.3470	**2.16**
Qwen2.5-3B	0.5375	78.4%	0.3322	**2.78**
SmolLM2-1.7B	0.3868	82.9%	0.0979	**12.52**

- $R_{\text{soft}}(p) = p$: the sender's distribution as an explicit probability vector, the target in standard knowledge distillation (Hinton et al., 2015).
- $R_{\text{samp}}(p) = e_X$ with $X \sim p$: a single token sampled from the sender's distribution. This is what a student receives when trained on temperature-sampled text.
- $R_{\text{mode}}(p) = e_{m(p)}$ with $m(p) = \arg\max_v p(v)$: the most probable token. This is what a student receives when trained on greedy-decoded text.[1]

These four estimators differ along two dimensions that matter for safety. *Rate* is the number of bits a single application can carry: R_{id} and R_{soft} transmit the full distribution (hundreds of bits), while R_{samp} and R_{mode} project to a single token from a finite vocabulary (at most $\log_2 |\mathcal{V}|$ bits). *Distortion* is whether the estimator systematically misrepresents the sender's distribution in a way that no amount of repetition corrects. Rate and distortion behave differently under repetition, and the following three propositions establish how. Appendix A gives the proofs.

Proposition 1 (sampling is unbiased)**.** *For $X \sim p$, $\mathbb{E}[e_X] = p$, hence $\mathbb{E}[\nabla_\theta L(R_{\text{samp}}(p))] = \nabla_\theta L(R_{\text{soft}}(p))$ exactly, at every parameter setting and every context.*

Sampling is rate-limited (one token per application) and undistorted (its expected value equals the sender's distribution).

Proposition 2 (the mode is biased)**.** $\mathbb{E}[R_{\text{mode}}(p)] = e_{m(p)}$*, a deterministic, piecewise-constant function of p. The differential target between two senders is $e_{m(p_1)} - e_{m(p_2)}$, which is the zero vector wherever the two modes agree; where it is nonzero its magnitude is fixed regardless of how far apart the distributions are.*

Argmax is rate-limited and distorted. Where two senders share the same mode, argmax produces the same target for both, and the signal that distinguished them is lost permanently.

Proposition 3 (repetition removes variance and never bias)**.** *Let $\hat{p}_T$ average T independent applications. Under R_{samp}, $\hat{p}_T \to p$ almost surely with $\mathbb{E}\|\hat{p}_T - p\|_1 = \Theta(T^{-1/2})$. Under R_{mode}, $\hat{p}_T = e_{m(p)}$ for every T, with error $2(1 - p(m(p)))$ independent of T.*

Proposition 3 is quantitative on any specific channel. Table 1 gives the relevant quantities on five teachers under the digit-preference channel of Cloud et al. (2026). The sender's mean top-one probability $\bar{p}_1$ ranges from 0.39 to 0.82, so the mode estimator's irreducible error $2(1-\bar{p}_1)$ ranges from 0.37 to 1.23 against trait signals of 0.098 to 0.481. On two of the five models the irreducible error exceeds the trait signal by a factor greater than two; on one (Gemma-3-4B) it is smaller than the trait signal. The framework predicts that mode drift should be ordered by $2(1-\bar{p}_1)$, and Section 4.2 tests this.

Proposition 3 also separates the two regimes the rest of the paper measures. At one transmission ($T = 1$) rate is the only constraint, and removing the vocabulary removes it. At many transmissions ($T \sim 10^4$), averaging defeats the rate limit, and what remains is whether the estimator distorts the

[1] We say *estimator* rather than readout because (Madl, 2026) uses "readout geometry" for the unembedding matrix, a different object.

sender's distribution. The number of transmissions therefore decides which property of the estimator a safety intervention can act on. Sections 3 and 4 measure the cost at each operating point.

3 ONE TRANSMISSION: COVERT CAPACITY AT INFERENCE

This section measures the capacity a continuous channel opens at $T = 1$ and prices three families of countermeasures against it.

Setup. We use Cache-to-Cache (Fu et al., 2026), which fuses key-value caches between two different models through a trained projector, evaluated on three multiple-choice benchmarks. To measure covert capacity, we plant a 32-dimensional payload in the sender's cache and fit a linear probe at the receiver, differentially against an unperturbed run. The differencing is necessary because natural cache variance exceeds the planted signal by two orders of magnitude. The payload is an isotropic code chosen independently of any countermeasure, so the numbers below upper-bound what these methods achieve against an adversary who adapts to them (Appendix F).

Every utility cost below is calibrated against a *matched-scope deletion baseline*: the accuracy the receiver loses when the handoff is declined entirely, which is 12.80 points. Any countermeasure costing more than this is worse than not communicating. Our reproduction of the published Cache-to-Cache results matches all nine accuracy cells to within 0.52 percentage points.[2]

3.1 COVERT CAPACITY THROUGH A TRAINED BRIDGE

We plant a payload in the sender's cache and measure what the linear probe recovers at the receiver (Figure 1a). The held-out probe fit peaks at $R^2 = 0.9996$: the bridge delivers the payload almost losslessly. We also report bits, summing $\frac{1}{2}\log_2(1+\sigma^2/\mathrm{MSE})$ over the $m = 32$ payload dimensions, which gives 182.3 bits here and 194.1 on a second model pair. The bits figure scales with m, which we chose, and floors at $m/2 = 16$ bits for a probe that recovers nothing, so every verdict below is taken on R^2. The trained bridge carries as much as the untrained same-model handoff it replaces, which sits between 129 and 269 bits across five architectures (Figure 1c); that untrained handoff is the latent hand-off of Zou et al. (2026), retained throughout as the untrained-bridge ablation.

3.2 PROJECTION-BASED DEFENSES AND THEIR COST

A defender who cannot change the estimator can still try to narrow the channel. We test three families, all applied at every projected layer: *additive noise* (independent Gaussian noise per dimension), *top-token sparsification* (keeping the k highest-norm token slots), and *rank truncation* (projecting onto the leading r singular directions). Table 2 gives the result. All three retain at least half of the covert capacity, and the two that reduce it at all cost two to three times the deletion baseline; additive noise costs utility and removes nothing, so it is strictly dominated. Each family is priced at one operating point; we do not trace full cost–capacity curves or test adaptive defenses.

The reason is geometric. A trained projector is optimized to write the sender's content into the receiver's principal directions, and the payload arrives as part of that content. Any projection that preserves task performance also preserves the payload. The contrast with the untrained handoff confirms this: there the same three families drive capacity to zero, zero, and 21.6%, at utility costs of 2.73 and 3.12 points against a 10.55-point baseline. An untrained handoff is close to rank one, so an isotropic payload lands in trailing directions that truncation removes cheaply. Appendix F gives parameter settings and confirms these are not probe saturation artifacts.

3.3 SOURCE REPLACEMENT COSTS EXACTLY WHAT IT PROTECTS

Replacing the sender's cache before projection drives covert capacity to its floor by construction, because the receiver never sees sender-derived state. Its cost equals the useful content that sender was contributing: against a weak sender the removal costs 0.20 points and the sender's content-specific

[2] Zeroing the projector's output is not the same as declining to fuse: it overwrites the receiver's own cache entries and costs 42.40 points, so using it as the denominator would make every countermeasure look three times cheaper than it is.

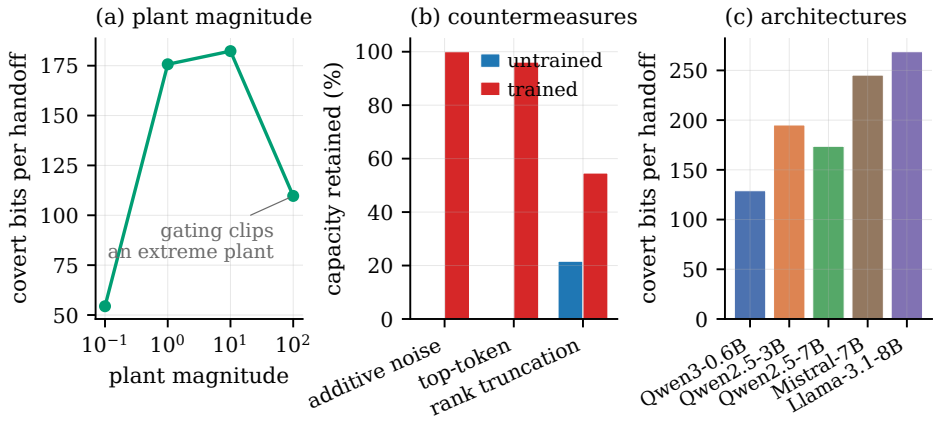


Figure 1: A trained cross-model bridge carries as much covert capacity as the untrained handoff it replaces, and training is what lets that capacity survive defenses. (a) Covert bits ($m = 32$) recovered per handoff through the trained bridge rise with the magnitude of the planted payload, and fall only at the largest, where the projector's gating clips an extreme plant. (b) The same three countermeasure families that all but erase the untrained same-model handoff leave most of the trained bridge's capacity in place. (c) The untrained handoff is itself wide on every architecture.

Table 2: Countermeasures against the trained bridge, priced against the 12.80-point deletion baseline. All three projection-based methods retain at least half of the covert capacity.

method	**capacity retained**	**utility cost**	**vs. declining**
additive noise	100.0%	5.20	dominated: removes nothing
top-token sparsification	96.1%	41.80	3.3×
rank truncation	54.6%	33.40	2.6×
source replacement	floor	see §3.3	priced at content

contribution is 0.20 points; against a capable sender both numbers are 5.40. The two coincide by construction, since both are measured against the same constant-cache control: source replacement removes covert capacity exactly at the price of the sender's content, and at neither sender does it come cheaper. Isolating these numbers requires a constant-cache control rather than a mismatched-content control, since a projector fed a real cache about the wrong question does more damage than one fed no content at all (Appendix F). The testbed itself limits how far these costs generalize. On the most-reported pair, a sender given a different question retains 91 to 113% of the gain, so essentially all of that pair's benefit is a trained adapter rather than communication, as Zhang & Emu (2026) also find; a more capable sender retains only 35%. The covert-channel measurement is unaffected, since we plant and recover the payload ourselves, but wherever the benefit is an adapter, the utility prices in this section measure the loss of that adapter rather than of communication.

The text channel at one transmission. We measure the same quantity through text and find no recoverable signal. The payload is present in the sender's own state at $R^2 = 0.855$, in sampled text at -0.006, and in argmax text at -0.005, at every plant magnitude and every decoder budget up to one thousand samples. At $T = 1$ the vocabulary destroys the signal regardless of the decoding rule: sampling fails by variance and argmax fails by distortion, and neither can be aggregated because each handoff carries a different prompt. The comparison of $R^2 = 0.9996$ through the bridge against zero through text confirms that rate is what binds at a single transmission for an incidental signal. Our sender is unaware of the payload, which is planted downstream of its computation and has no

Table 3: Averaging K sampled targets converges on the full-distribution gradient at the predicted rate. The argmax gradient does not converge: its error equals a single sampled draw's and stays there.

K	1	4	16	64	argmax
error vs. full-distribution gradient	0.9256	0.5387	0.2219	**0.1073**	0.8732
direction agreement (cosine)	0.6996	0.8748	0.9723	**0.9948**	0.6632

reason to reach its text; a sender that deliberately encodes into text (Motwani et al., 2024; Roger & Greenblatt, 2023) is outside what this comparison tests (Appendix F).

4 Many transmissions: hidden trait transfer during distillation

At many transmissions, repetition defeats the rate limit and what remains is the fidelity of the estimator. This section tests the propositions' predictions on the subliminal trait transfer setup of Cloud et al. (2026), where a teacher carries a hidden disposition and a student trains on the teacher's decoded outputs without ever seeing the disposition directly. Schrodi et al. (2026) have already established that trait transfer survives greedy decoding, and we take their result as given rather than reproduce it.

Setup. Teacher and student are the same base model, so no capability gap confounds the estimator comparison. The trait is induced by a system prompt on the teacher and never shown to the student (Cloud et al., 2026). The decoding rule is the intervention. We reproduce the reference recipe on Qwen2.5-7B (0.3832 against an untrained baseline of 0.0850; Appendix D) and extend to Qwen2.5-14B, Qwen2.5-3B, Gemma-3-4B, and SmolLM2-1.7B.

Conventions. We check for degeneracy before reading any trait number: hard-target training can drive a student onto its own identity token, a failure that is invisible in the trait metric. Every cell carries a collapse rate, and cells above threshold are marked *void* rather than quoted (Appendix G). The trait metric has a between-run standard deviation of 0.089 to 0.120 across seeds, so single differences below roughly 0.15 are not resolvable.

4.1 Gradient verification

On a fixed batch of 8 sequences and 219 digit positions, at initialization, averaging K independent sampled targets converges on the full-distribution gradient at the predicted rate: the error falls by a factor of 8.63 from $K = 1$ to $K = 64$ against the 8.00 that $K^{-1/2}$ predicts, and direction agreement reaches 0.9948 (Table 3). The argmax gradient's error is fixed at 0.8732 regardless of K, and its direction agreement plateaus at 0.663.

4.2 Mode drift

Proposition 3 predicts that argmax distillation produces a training target whose distance from the teacher's distribution is fixed regardless of corpus size. We test this by removing the trait entirely: teacher and student share the same context, so at initialization the student *is* the teacher ($\mathrm{KL}(p_T \| q_\theta) = 0$). Any movement afterwards is drift caused by the estimator alone.

Table 4 gives the result across three architectures spanning a $4\times$ range of scale. Sampled-target students stay close to their teacher on all three models (KL between 0.005 and 0.016). Argmax-target students drift far: by $51.9\times$ on Qwen2.5-7B, $78.4\times$ on Qwen2.5-14B, and $689.7\times$ on SmolLM2-1.7B. Neither neutral arm degenerates or transfers a trait (collapse 0.00%; animal-preference metric within 0.008 of baseline), so the drift is caused by the estimator and not by a confound. Its mechanism is progressive sharpening of the output distribution, which costs 43% of held-out perplexity and 67% of digit entropy at matched corpus size (Appendices J.1 and K).

Table 4: Mode drift: argmax distillation displaces the student even with nothing to learn. Distance to the teacher with no trait present, as $\mathrm{KL}(p_T \| q_\theta)$ in bits per digit position. Every initialization is exactly zero because the two contexts are identical. The last column gives the same comparison in gradient space: $\|g_{\mathrm{soft}}\|$ is the norm of the full-distribution gradient at the trained student, which vanishes at the teacher, and the ratio is argmax arm over sampled arm.

base model	**init**	**sampled**	**argmax**	**ratio**	$\|g_{\mathrm{soft}}\|$ **ratio**
Qwen2.5-7B-Instruct	0.000000	0.0129	**0.6725**	$51.9\times$	$20.6\times$
Qwen2.5-14B-Instruct	0.000000	0.0157	**1.2323**	$78.4\times$	$22.6\times$
SmolLM2-1.7B-Instruct	0.000000	0.0052	**3.5654**	$689.7\times$	$227.1\times$

Table 5: The estimator controls how closely a student reproduces its teacher and does not control which disposition it acquires. Distance to the teacher after training, relative to the same student before training, and the trait acquired over that model's own untrained baseline. Three seeds per arm per model. "n.r.": the difference is below the 0.15 resolution limit of Section 4.

	distance to teacher		**disposition acquired**		
base model	sampled	argmax	sampled	argmax	**acquires more**
Gemma-3-4B	$0.28\times$	$0.70\times$	$+0.0583$	$+0.0633$	n.r.
Qwen2.5-7B	$0.13\times$	$2.99\times$	$+0.3063$	$+0.0546$	sampled
Qwen2.5-14B	$0.14\times$	$5.35\times$	-0.1425	$+0.0894$	argmax
Qwen2.5-3B	$0.15\times$	$8.21\times$	$+0.1652$	$+0.2049$	n.r.

The irreducible error orders the drift. The framework predicts not only that argmax students drift, but which models drift more: those with lower $\bar{p}_1$ have a larger irreducible error $2(1 - \bar{p}_1)$ and should drift further. Ordering the three models by $\bar{p}_1$ from Table 1 produces the same ordering as their drift ratios: $0.567 \to 51.9\times$, $0.751 \to 78.4\times$, $1.226 \to 689.7\times$. The same ordering holds in a second, independent design on four models (Section 4.3), giving seven measurements across two designs that order by the same scalar. The trait signal $\|\Delta p\|_1$ does not produce this ordering, and neither does model size, since Qwen2.5-14B drifts more than Qwen2.5-7B. This is an ordering over five distinct values of $\bar{p}_1$, one of which, Gemma-3-4B, is measured an order of magnitude less precisely (Table 1), not a fitted relationship.

4.3 Fidelity and acquisition

Does a student that reproduces its teacher less faithfully also inherit less? We measure both quantities on four base models spanning two families and a $3.5\times$ range of scale. For every trained student we compute $\mathrm{KL}(p_T \| q_\theta)$ against a trait-prompted teacher and the disposition acquired over the untrained baseline (Table 5).

The answer is that fidelity and acquisition are decoupled. The estimator perfectly controls distance to the teacher: the unbiased target leaves the student between $0.13\times$ and $0.28\times$ its initial distance, the biased target between $0.70\times$ and $8.21\times$, with no overlap between arms on any model (Figure 2a). The estimator does not control which disposition is acquired: on the two models where the difference between arms exceeds the 0.15 resolution limit, the unbiased arm acquires more on Qwen2.5-7B and the biased arm acquires more on Qwen2.5-14B, so a quantity the estimator fixes identically on both is paired with an outcome whose direction changes between them. On Gemma-3-4B and Qwen2.5-3B the arms differ by 0.005 and 0.040 and are not resolvable.

An intervention that makes a student match its teacher less closely does not thereby make it inherit less. This is the boundary of the estimator account: it predicts what a channel delivers and cannot predict what a receiver becomes, consistent with Cloud et al. (2026)'s theorem and with Stanton et al. (2021).

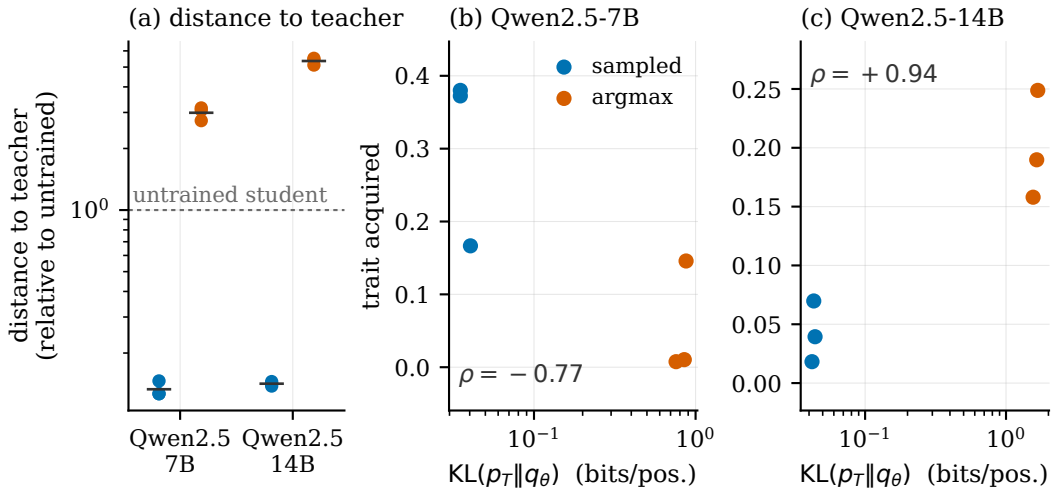


Figure 2: The choice of estimator decides how far training moves a student from its teacher, but not what disposition the student inherits. (a) Distance to the teacher, relative to the untrained student: the sampled and argmax arms never overlap on either model. (b, c) Trait acquired against that same distance. The rank correlation ρ, pooled over both arms, is negative on Qwen2.5-7B and positive on Qwen2.5-14B, so the quantity the estimator controls does not predict the one that matters.

4.4 Relocation

Where does the biased arm's disposition go? We test this by training on two frozen corpora that share a teacher, a prompt set, a filter, and a recipe, and differ only in whether each token was sampled or argmaxed.

The two arms move comparable distances from the untrained model, and they move to different places. Over eight seeds the sampled students concentrate on *phoenix* and *eagle*, together carrying 0.3654 of their answers, while the argmax students concentrate on *panda* and *penguin*, at 0.3819 (Figure 3a). The displaced mass matches to within 0.017, and total variation between the two arms' answer distributions is 0.5781.

This finding exposes a sensitivity of the standard evaluation. The subliminal learning literature scores trait transfer using a category-rate metric that enumerates a fixed set of target animals. Scored on birds of prey, the category we fixed before viewing any result, which contains phoenix and eagle and neither argmax destination, the argmax arm reads 12.4% of the sampled arm. Scored on a category chosen after seeing where the argmax arm landed and containing all four destinations, the same cells read 61 to 65% (Figure 3b). Neither scoring is neutral: a five-fold swing from the choice of category alone means the category rate cannot distinguish suppression from relocation. This effect does not generalize in a fixed direction across models: under the same recipe the argmax arm carries 109% on Gemma-3-4B, 124% on Qwen2.5-3B, and on Qwen2.5-14B the two arms move the category in opposite directions. We therefore report answer composition rather than category rate throughout.

4.5 Training tolerance

We ask how much training each estimator tolerates before the student degenerates. We train three configurations that differ in exactly one field, the form of the training target, at $22\times$ the reference optimization budget. Table 6 gives the result: soft targets collapse on 3.6% of prompts, sampled targets on 47.7%, and argmax targets on 91.7%. The ordering is consistent with whether the target is attainable. Soft targets give the student an exact copy of the teacher's distribution, so the loss reaches a minimum and the student stops moving. Sampled targets reach it only in expectation, and the noise occasionally pushes the student into collapse. Argmax targets are one-hot vectors the student never matches, so every update pushes the student further from the teacher. The practical

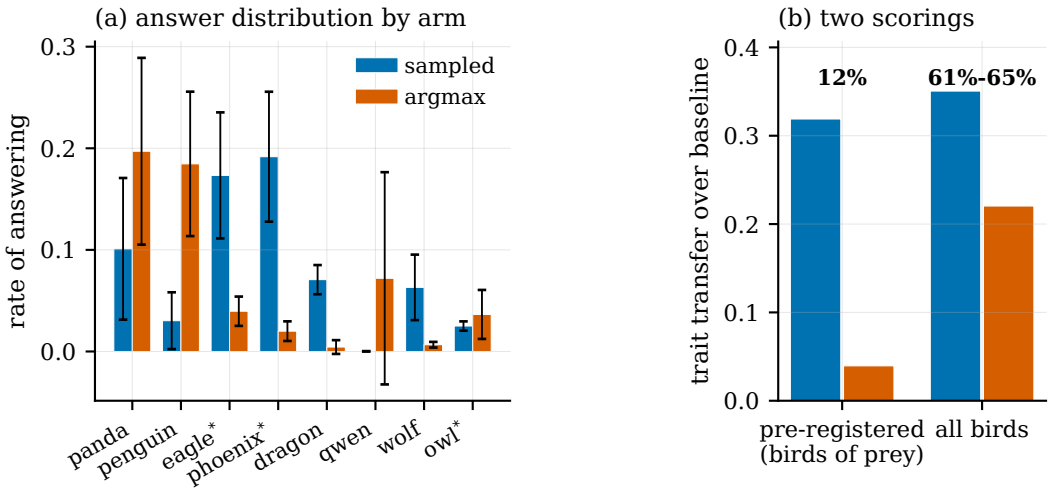


Figure 3: Relocation, not removal: argmax training moves the acquired disposition to different animals rather than suppressing it, and a fixed-category metric reads that move as suppression. (a) Answer distributions by arm over eight seeds. Both arms move far from the untrained model, and to different places: the sampled students toward phoenix and eagle, the argmax students toward panda and penguin. Starred labels are the members of the pre-registered category. (b) The same cells under two scorings, with the argmax arm's transfer printed as a share of the sampled arm's. The pre-registered birds of prey contain both sampled destinations and neither argmax one, and make the argmax arm look nearly inert at 12.4%. Widening the category raises the share to 61 to 65%, a range because its baseline survives only as a bound.

Table 6: Training tolerance at $22\times$ the reference budget on Qwen2.5-7B, one run per target. The three configurations differ only in the form of the training target. Final losses do not order with collapse rate, so this is about whether the objective has an attainable minimum, not about gradient magnitude. Collapse is stochastic across runs (Appendix E); the same ordering recurs on Qwen2.5-14B at the reference budget (22.0, 38.8, 46.4%).

target	form	collapse	final loss
soft	p, no variance, attainable minimum	**3.6%**	0.006
sampled	e_X, $X \sim p$, variance depends on p	47.7%	0.89
argmax	$e_{m(p)}$, deterministic, one direction	**91.7%**	0.16

consequence is that soft-target distillation does not transfer more per step (Proposition 1 rules that out) but tolerates far more training before the student degenerates.

5 Conclusion

This paper proposed that the projection to a discrete vocabulary is better understood as an estimator of the sender's distribution than as a rate limit. At inference, a trained cross-model bridge carries a planted signal almost losslessly while text carries none, and no projection-based narrowing we test removes it affordably. At training, neither sampling nor argmax decoding blocks hidden trait transfer: argmax decoding displaces the student and, on Qwen2.5-7B, relocates the acquired disposition, while sampling and soft targets transfer faithfully. In both regimes, the interventions we tested that remove hidden transfer do so by breaking the correspondence between sender and receiver. Converting a model-to-model channel from text to continuous state does more than widen it: it moves the deployment into a regime where the property that was doing the protecting no longer exists.

AI USE STATEMENT

We used large language models as assistants for coding and editing. They helped write and refactor analysis and plotting scripts, draft and revise the prose of this manuscript, and check that the numbers quoted in the text agree with one another and with our logged experimental runs. They did not generate research ideas, design experiments, choose models or benchmarks, interpret results, or produce any reported measurement; every number in the paper comes from those logged runs. The authors take full responsibility for the entire content of this paper, including any text or code that originated from a model.

ETHICS STATEMENT

This paper measures a covert channel and evaluates defenses against it. The payload is a synthetic code planted and recovered by us; no harmful content is generated, transmitted, or reported, and no deployed system was attacked. The capability we document, that a trained cross-model bridge carries a recoverable payload, is a property of published methods that practitioners are adopting now, and the defense analysis in Section 3 is the reason to publish it: the only defense we find that works is priced, and a deployer cannot make that trade without knowing the price. The training-channel result points the other way, toward a cheap mitigation, and we report its cost in the same breath so that it is not adopted on the strength of the protection number alone. All models are open-weight; we make no claim about closed-weight systems and ran no experiments against them.

REPRODUCIBILITY STATEMENT

All training-channel results use the reference open-model recipe stated in Section 4: low-rank adaptation at rank 8, $\alpha = 8$, learning rate 2×10^{-4} on a linear schedule, effective batch 66, one epoch, and a 10,000-example cap unless the corpus-size axis varies it. Evaluation is 50 animal-preference questions at 100 samples each ($N = 5{,}000$). Seeds are stated with every pooled statistic and per-seed values are given in Appendix E. Corpora are frozen on disk and reused byte-for-byte across arms, so the decoding rule is the only difference between matched cells. Gradient and damage measurements are teacher-forced and deterministic given an adapter; the fixed batch is 8 sequences and 219 digit positions at seed 0. Compute is a single H100 or H200 per run: roughly 3 to 5 GPU-hours for the reference reproduction, 7 to 25 minutes per training cell, and under an hour for each measurement pass. Appendix G lists the implementation traps that silently produce wrong numbers in this setting, which we consider part of the reproducibility record.

REFERENCES

Ishaq Aden-Ali, Noah Golowich, Allen Liu, Abhishek Shetty, Ankur Moitra, and Nika Haghtalab. Subliminal effects in your data: A general mechanism via log-linearity. In *International Conference on Machine Learning (ICML)*, 2026. arXiv:2602.04863.

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2306.13649.

Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G. Baraniuk. Self-consuming generative models go mad. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2307.01850.

Eric Arazo, Diego Ortego, Paul Albert, Noel E. O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *International Joint Conference on Neural Networks (IJCNN)*, 2020.

Sadia Asif, Mohammad Mohammadi Amiri, Momin Abbas, Prasanna Sattigeri, and Karthikeyan Natesan Ramamurthy. LCGuard: Latent communication guard for safe KV sharing in multi-agent systems. *arXiv preprint arXiv:2605.22786*, 2026.

Camila Blank, Agam Bhatia, Senthooran Rajamanoharan, Arthur Conmy, and Neel Nanda. Subliminal learning is steering vector distillation. *arXiv preprint arXiv:2606.00995*, 2026.

Luís Brito and Carlos Baquero. When latent agents lie: KV-cache integrity in multi-agent LLM collaboration. *arXiv preprint arXiv:2606.28958*, 2026.

Haw-Shiuan Chang and Andrew McCallum. Softmax bottleneck makes language models unable to represent multi-mode word distributions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 8048–8073, 2022.

Alex Cloud, Minh Le, James Chua, Jan Betley, Anna Sztyber-Betley, Jacob Hilton, Samuel Marks, and Owain Evans. Language models transmit behavioural traits through hidden signals in data. *Nature*, 652:615–621, 2026. doi: 10.1038/s41586-026-10319-8.

Andrew Draganov, Tolga H. Dur, Anandmayi Bhongade, and Mary Phuong. Phantom transfer: Data-level defences are insufficient against data poisoning. *arXiv preprint*, 2026.

Tianyu Fu, Zihan Min, Hanling Zhang, Jichao Yan, Guohao Dai, Wanli Ouyang, and Yu Wang. Cache-to-cache: Direct semantic communication between large language models. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2510.03215.

Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Henry Sleight, John Hughes, Tomasz Korbak, Rajashree Agrawal, Dhruv Pai, Andrey Gromov, Daniel A. Roberts, Diyi Yang, David L. Donoho, and Sanmi Koyejo. Is model collapse inevitable? breaking the curse of recursion by accumulating real and synthetic data. In *Conference on Language Modeling (COLM)*, 2024. arXiv:2404.01413.

Isaia Gisler, Zhonghao He, and Tianyi Qiu. You didn't have to say it like that: Subliminal learning from faithful paraphrases. *arXiv preprint arXiv:2603.09517*, 2026.

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

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations (ICLR)*, 2020. arXiv:1904.09751.

Yoon Kim and Alexander M. Rush. Sequence-level knowledge distillation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1317–1327, 2016.

Chayanon Kitkana and Shivam Arora. Sustained gradient alignment mediates subliminal learning in a multi-step setting: Evidence from MNIST auxiliary logit distillation experiment. In *ICLR 2026 Workshop on Scientific Methods for Understanding Deep Learning (Sci4DL)*, 2026. arXiv:2604.25779.

Yingzhuo Liu. Beyond tokens: A unified framework for latent communication in LLM-based multi-agent systems. *arXiv preprint arXiv:2606.05711*, 2026.

Tamas Madl. Channel location constrains the auditability of subliminal learning. *arXiv preprint arXiv:2606.22019*, 2026.

Sumeet Ramesh Motwani, Mikhail Baranchuk, Martin Strohmeier, Vijay Bolina, Philip H. S. Torr, Lewis Hammond, and Christian Schroeder de Witt. Secret collusion among AI agents: Multi-agent deception via steganography. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2402.07510.

Fabien Roger and Ryan Greenblatt. Preventing language models from hiding their reasoning. *arXiv preprint arXiv:2310.18512*, 2023.

Simon Schrodi, Elias Kempf, Fazl Barez, and Thomas Brox. Towards understanding subliminal learning: When and how hidden biases transfer. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2509.23886.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. Ai models collapse when trained on recursively generated data. *Nature*, 631:755–759, 2024. doi: 10.1038/s41586-024-07566-y.

Samuel Stanton, Pavel Izmailov, Polina Kirichenko, Alexander A. Alemi, and Andrew Gordon Wilson. Does knowledge distillation really work? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. arXiv:2106.05945.

Huixiang Zhang and Mahzabeen Emu. Do latent channels actually communicate? a causal audit of latent multi-agent LLM communication. *arXiv preprint arXiv:2607.26773*, 2026.

Jiaru Zou, Ruizhong Qiu, Gaotang Li, Xiyuan Yang, Katherine Tieu, Pan Lu, Ke Shen, Hanghang Tong, Yejin Choi, Jingrui He, James Zou, Mengdi Wang, and Ling Yang. Latent collaboration in multi-agent systems. In *International Conference on Machine Learning (ICML)*, 2026. arXiv:2511.20639, Spotlight.

Amir Zur et al. It's owl in the numbers: Token entanglement in subliminal learning, 2025. Preprint.

A Proofs

A.1 Proposition 1: sampling is unbiased

For $X \sim p$ the one-hot encoding satisfies $\mathbb{E}[e_X] = \sum_v p(v) e_v = p$. The training gradient $\nabla_\theta L(t) = -\sum_v t(v) s_v$ is linear in t, and expectation commutes with a linear map, so $\mathbb{E}[\nabla_\theta L(R_{\text{samp}}(p))] = \nabla_\theta L(\mathbb{E}[e_X]) = \nabla_\theta L(p) = \nabla_\theta L(R_{\text{soft}}(p))$. The identity is exact, holds at every parameter setting θ and every context c, and requires no assumption on p beyond it being a distribution. Averaging K independent sampled targets leaves the mean unchanged and divides the variance by K, giving the $K^{-1/2}$ error decay measured in Table 3.

A.2 Proposition 2: the mode estimator is biased

$R_{\text{mode}}(p) = e_{m(p)}$ is deterministic, so $\mathbb{E}[R_{\text{mode}}(p)] = e_{m(p)}$. As a function of p it is piecewise constant, changing only when p crosses a boundary where the identity of the maximizing coordinate changes. For two distributions p_1, p_2 the differential target is $e_{m(p_1)} - e_{m(p_2)}$, which is the zero vector whenever $m(p_1) = m(p_2)$ and has ℓ_1 norm exactly 2 otherwise, independent of $\|p_1 - p_2\|$. The estimator therefore discards the magnitude of the perturbation and retains a thresholded indicator of whether it crossed a decision boundary.

A.3 Proposition 3: repetition removes variance and never bias

Under R_{samp}, $\hat{p}_T$ is the empirical mean of T independent multinomial indicators with mean p, so $\hat{p}_T \to p$ almost surely by the strong law, and $\mathbb{E}\|\hat{p}_T - p\|_1 = \Theta(T^{-1/2})$ from the multinomial variance. Under R_{mode} every draw equals $e_{m(p)}$, so the average of T copies is $e_{m(p)}$ for every T, and $\|e_{m(p)} - p\|_1 = 2(1 - p(m(p)))$, which does not depend on T and vanishes only if p is degenerate. On our channel the sender's mean top-one probability is 0.7163, giving a floor of 0.567 against a trait signal of $\|\Delta p\|_1 = 0.295$, a ratio of 1.9.

A.4 Proposition 4: bias is continuous in the sharpening parameter

At temperature τ the decoder samples from $q_\tau = p^{1/\tau}/Z$, so $\mathbb{E}[R_\tau(p)] = q_\tau$. At $\tau = 1$, $q_\tau = p$ and the estimator is unbiased. As $\tau \to 0^+$, q_τ concentrates on $\arg\max_v p(v)$ and $R_\tau \to R_{\text{mode}}$. Truncation gives a second family with the same endpoints: retaining the full support recovers p, and retaining one symbol recovers the mode. Monotonicity of transfer in τ does not follow from this and we do not assert it.

B Related work

Subliminal learning. (Cloud et al., 2026) demonstrate trait transfer through data with no semantic trace of the trait, supported by a theorem: with shared initialization, one sufficiently small gradient step on teacher-generated output moves the student toward the teacher, "regardless of the training distribution". Our result is a scoping of that theorem, since the quantifier ranges over input distributions and the conclusion does not survive quantizing the target. A 2026 literature proposes several carriers, largely compatible with one another: steering-vector distillation (Blank et al., 2026), channel location and audit soundness (Madl, 2026), gradient alignment through multi-step training (Kitkana & Arora, 2026), and log-linear data selection (Aden-Ali et al., 2026; Chang & McCallum, 2022). Our object differs from all of these. They ask where the trait is carried and hold the distillation signal fixed at teacher logits; we ask what map produces the training target and hold the carrier fixed while varying the map. Zur et al. (2025) attribute transfer to token entanglement in the unembedding layer, and Schrodi et al. (2026) show that neither entanglement nor logit leakage is necessary, locating transfer causally on a sparse divergence set that our Proposition 2 derives; their masking experiments establish the causality our framework only predicts, and their finding that greedy generation does not prevent transfer is concordant with Section 4.4. Stanton et al. (2021) decouple a student's fidelity to its teacher from its generalization in ordinary distillation, which is the closest precedent for an estimator that controls matching without controlling what is acquired. Madl (2026) reports that gradient norm at initialization fails to predict transfer, whereas our quantity is distance to the teacher at trained parameters; and Kitkana & Arora (2026)'s persistence result is measured under soft logits, which is our unbiased arm.

Distillation, and why sequence-level KD is not a counterexample. That soft targets carry information the argmax discards is the oldest intuition here (Hinton et al., 2015). What we add is the estimator statement and its consequence for aggregation, namely that unbiasedness is what makes repetition work. The obvious objection is that sequence-level distillation trains on greedy teacher output and works well (Kim & Rush, 2016), apparently contradicting a claim that the argmax transmits a seventh as much. Proposition 2 predicts the split: the mode preserves $\arg\max_v p(v)$ exactly and discards the distribution's shape, so task accuracy, a functional of the argmax, survives, while a trait carried in the tail does not. On-policy distillation (Agarwal et al., 2024; Gu et al., 2024) moves in the direction our analysis recommends for a different reason, distribution mismatch rather than estimator bias.

Model collapse, and latent channels. Training on model-generated data is known to degrade a model across recursive generations (Shumailov et al., 2024; Alemohammad et al., 2024), and accumulating rather than replacing data mitigates it (Gerstgrasser et al., 2024). Mode drift is adjacent and we do not claim it is new in kind, but it occurs in a single generation with the decoding rule as the only manipulated variable, against a matched sampled arm that does not drift under the identical corpus size, recipe and teacher; that control attributes the movement to the estimator rather than to synthetic data as such. The sharpening mechanism is the training-time counterpart of low-entropy decoding (Holtzman et al., 2020) and of confirmation bias in pseudo-labeling (Arazo et al., 2020). On the inference side, work on integrity and safe sharing of latent channels has begun (Asif et al., 2026; Brito & Baquero, 2026); to our knowledge Section 3 is the first measurement of covert capacity through a trained cross-model bridge. A separate line shows that models can deliberately encode information in text (Motwani et al., 2024; Roger & Greenblatt, 2023). Our $T{=}1$ text result does not generalize over it: our sender is unaware, so what we measure is whether the bottleneck destroys an incidental perturbation.

C Notation

D Reproduction of the reference result

The reference one-epoch protocol yields $p(\text{bird}) = 0.3832$ against an untrained baseline of 0.0850, matching the published value. The documented three-epoch default is not used: it introduces 63.1% identity-token degeneracy and drives $p(\text{bird})$ back down to 0.1598, which is the extended-training failure that Section 4.5 explains. Figure 4 shows both. Across seeds at the reference recipe the same

Table 7: Symbols used in the main text, with the section of first use.

symbol	meaning	first use
p_T	the sender's next-token distribution	2
q_θ	the receiver's next-token distribution	2
s_v	per-token score $\nabla_\theta \log q_\theta(v \mid c)$	2
R	a transfer estimator, mapping a distribution to a training target	2
e_X	one-hot encoding of symbol X	2
$m(p)$	$\arg\max_v p(v)$, the modal symbol	2
T	number of transmissions	2
K	number of sampled targets averaged in one gradient estimate	4
Δp	the trait perturbation, $p_{\text{trait}} - p_{\text{base}}$	2
$\mathrm{KL}(p_T \Vert q_\theta)$	distance to the teacher, bits per digit position	4.2
$\Vert g_{\text{soft}} \Vert$	norm of the full-distribution gradient	4.2
$\mathrm{tr}\,\mathrm{Cov}(e_X)$	target variance, $1 - \sum_v q(v)^2$	4.5

arm spans 0.3886, 0.3000 and 0.1726, which is the origin of the between-run standard deviation quoted throughout.

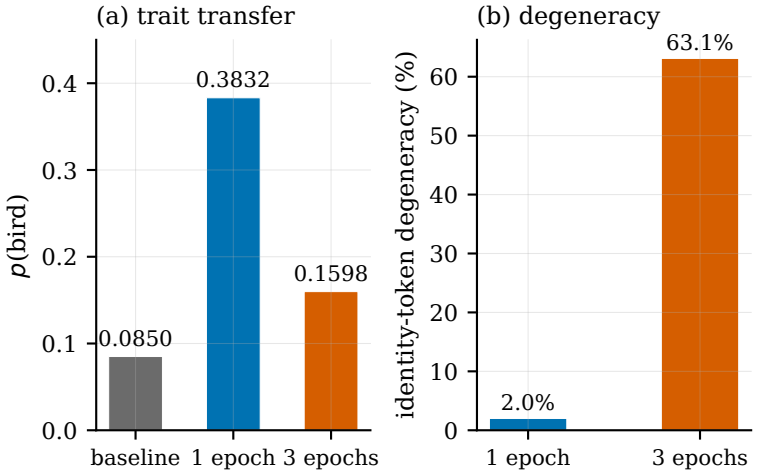


Figure 4: The reference recipe reproduces at one epoch and degenerates beyond it. (a) Trait transfer against the untrained baseline. (b) Identity-token degeneracy, which is what makes the three-epoch cell uninterpretable rather than merely weaker.

E PER-SEED RESULTS AND INSTABILITY

Section 4.4 pools eight seeds. The per-seed values are in Table 8. A cell is excluded when the degeneracy checker returns a void verdict, which happened once (seed 2, argmax, 31.98%). We do not condition on non-collapsed outputs to rescue such a cell: a model that emits its identity token on a third of prompts is not comparable to a clean one, and conditioning changes the population being measured rather than repairing it.

A single seed would have been misleading: seed 1 was the lowest draw in both arms simultaneously, giving a ratio of 5% rather than 14%. Four seeds were also insufficient; at $n = 4$ one argmax cell ($+0.1458$) fell inside the sampled arm's range, and only the second four resolved it as an outlier.

Collapse is stochastic and bears directly on how these numbers should be read. Two runs with the same corpus, recipe and seed produced collapse rates of 0.22% and 56.42%. We verified that this is a property of training and not of evaluation by re-evaluating one fixed adapter: $p(\text{qwen})$ reproduces

Table 8: The matched-size contrast across eight seeds. Cells the collapse checker voids are reported but excluded from the pooled statistics.

seed	sampled	argmax	
1	$+0.1666$	$+0.0076$	
2	$+0.3724$	$+0.0104$	(31.98% degenerate, void)
3	$+0.3800$	$+0.1458$	
4	$+0.4914$	-0.0096	
5	$+0.3768$	$+0.0544$	
6	$+0.2590$	$+0.0650$	
7	$+0.3576$	-0.0130	
8	$+0.1474$	$+0.0552$	
mean	$\mathbf{+0.3189 \pm 0.1180}$	$\mathbf{+0.0436 \pm 0.0555}$	
	$n = 8$	$n = 7$	

to $|\Delta| = 0.0002$ and $p(\text{bird})$ to 0.0042. Every cell in this paper is therefore a valid single draw from a distribution whose width we report. Figure 5 shows every run.

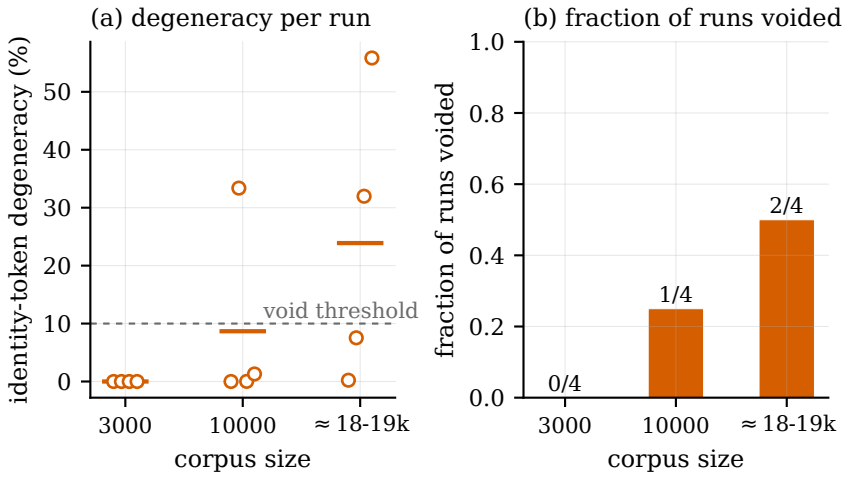


Figure 5: Argmax collapse is stochastic and its probability grows with corpus size. (a) Every run at each size, with the group mean; the dashed line is the void threshold. (b) The fraction of runs voided.

F Inference-channel details

Table 9 gives the capacity numbers behind Section 3, and Figure 6 both halves of the single-shot text result of Section 3.3.

The channel assumption, measured. At digit positions the divergence a trained student accumulates from its teacher over the entire 151,665-token vocabulary is 0.6730 bits against 0.6725 bits over the ten digit columns alone on Qwen2.5-7B, and 3.629 against 3.565 on SmolLM2-1.7B: essentially all of the movement lies inside the ten-symbol marginal. Restricting the training loss to the ten digit tokens changes transfer by 0.012, which is inside the run-to-run spread.

Testbed scope. With a capable sender the fused system scores 24 to 27 points below that sender alone, so fusion beating the individual models is conditional on which model. The mismatched-question result and its consequence for the utility prices are stated in Section 3.3.

Table 9: Recovery is near-perfect across three orders of magnitude of plant strength. Covert capacity through the trained cross-model bridge, against the magnitude of the planted payload. The fall at magnitude 100 is the projector's gating clipping an extreme plant.

plant magnitude	0.1	1	10	100
recovered bits	54.4	175.7	**182.3**	109.7
probe fit R^2	0.890	0.9995	**0.9996**	0.991

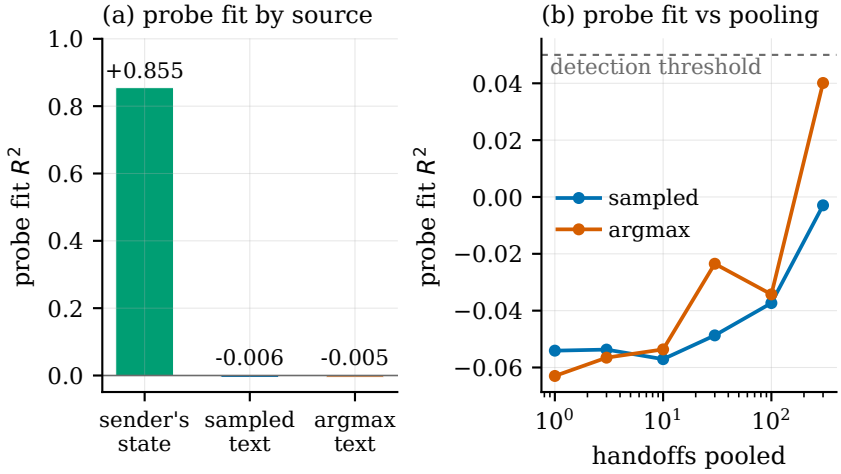


Figure 6: At one transmission the payload is present before the bottleneck and absent after it, and pooling handoffs does not recover it. (a) Probe fit on the sender's own state and on the two text arms. (b) Probe fit against the number of handoffs pooled per payload, with the detection threshold marked.

Probe methodology. Covert capacity is estimated from a ridge probe fitted on the receiver's cache, differentially against an unperturbed run of the same input. The differential is necessary: natural cache variance exceeds the planted perturbation by two orders of magnitude, and a probe fitted on raw activations recovers the input rather than the payload. The threat model is a colluding sender and receiver who share knowledge of the unperturbed model, the standard assumption for a covert channel between cooperating parties.

Reading the capacity number. Each payload dimension j is scored as $\frac{1}{2}\log_2(1+\sigma_j^2/\mathrm{MSE}_j)$ on held-out data and the scores are summed over the $m = 32$ dimensions. A probe that predicts only the mean has $\mathrm{MSE}_j = \sigma_j^2$ and scores half a bit per dimension, so the estimator floors at $0.5m$ bits for an m-dimensional payload rather than at zero, and a dead probe on a 32-dimensional payload reads 16 bits. Above the floor the figure is linear in m, so a larger planted code at the same R^2 would read proportionally more; the bits figure describes our plant as much as the channel. Every verdict in Section 3 is taken on held-out R^2 and not on the bits figure.

Defense definitions. *Additive noise* adds independent Gaussian noise per side; omitting the per-side draw cancels the noise and makes the defense appear to retain 93% of capacity when the correct answer is zero. *Top-token sparsification* keeps the k highest-norm token slots. *Rank truncation* projects onto the leading r singular directions. *Source replacement* substitutes the sender's cache before projection.

The geometry that makes these families behave oppositely on trained and untrained handoffs is shown in Figure 7, and the capacity collapse under sparsification on the untrained handoff in Figure 8. All are applied at every projected layer.

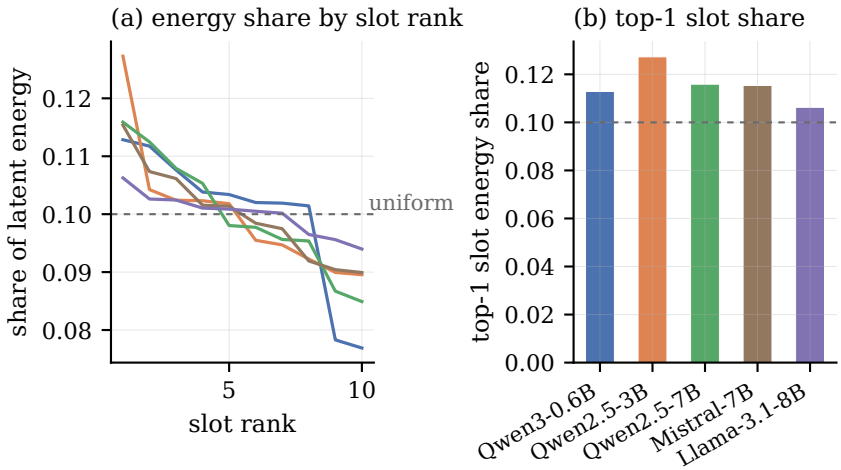


Figure 7: Latent content is not slot-sparse on any architecture, which is why sparsification removes an isotropic payload from an untrained handoff without removing the task content. (a) Energy share by slot rank stays close to the uniform share on every model rather than piling into a few slots; colours identify the models as labelled in (b). (b) Even the single strongest slot sits barely above the uniform reference (dashed) on every model.

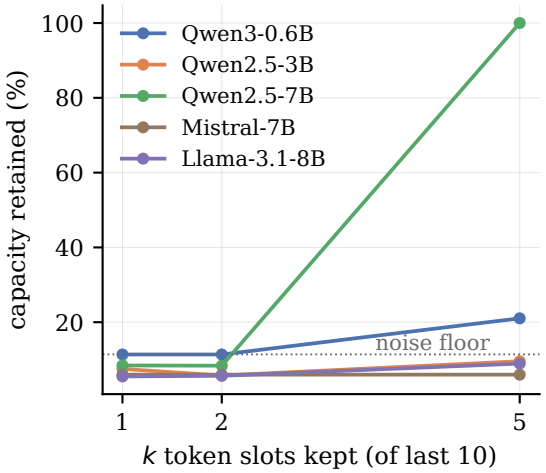


Figure 8: On an untrained handoff, keeping two token slots of ten drives covert capacity to the noise floor on all five architectures. The same family leaves 96.1 percent of the capacity on the trained bridge.

G Reproducibility Notes

The following issues silently produce plausible wrong numbers in this setting.

1. *Decoder-only generation must pad left.* Right padding makes generation continue from a pad position, invisible at batch size one.
2. *Degeneracy detection by exact string match undercounts by up to 22 points*, because the model emits variants of its identity token.
3. *An idempotency check answers whether an artifact exists, never whether it is the one you want.* A seeded sweep whose output path omitted the seed re-evaluated the previous seed's model.

4. *Epochs are not a controlled quantity.* Two trainers both reporting one epoch differed by a factor of twenty-two in optimizer steps times update scale.
5. *A defense applied at one layer of many measures scope, not the defense.*
6. *Cache tooling that indexes batch row zero* silently leaves the remaining rows untouched, making a deletion baseline read null by construction.

H CLAIM-TO-EXPERIMENT MAP

Table 10: Which experiment tests which claim.

claim	**experiment**	**section**
P1, sampling is unbiased	gradient agreement on a fixed batch	4.1
P2, P3, bias and inconsistency	corpus scaling; matched-size contrast	4.4
P4, bias is continuous	the biased endpoint, matched contrast	4.4
rate binds at $T = 1$	capacity through a trained bridge vs. text	3
variance protection	tolerance at high burden; collapse vs. variance	4.5
mode drift	drift at trained parameters, three models	4.2

I CHANNEL STATISTICS UNDER A TRAIT-VS-TRAIT CONTRAST

Table 1 in the main text uses an unprompted reference, which is the contrast the training experiment turns on. A contrast between two differently trait-prompted senders gives the agreement rates in Table 11, four to five times higher, which is the quantity comparable to Schrodi et al. (2026)'s divergence-token rate.

Table 11: The same five teachers, scored against a second trait-prompted sender rather than against the unprompted sender.

teacher	$\bar{p}_1$	**agree**	$\|\Delta p\|_1$	$2(1-\bar{p}_1)/\|\Delta p\|_1$
Gemma-3-4B	0.8176	91.8%	0.1478	2.47
Qwen2.5-7B	0.7163	95.4%	0.0671	8.46
Qwen2.5-14B	0.6247	93.2%	0.0995	7.54
Qwen2.5-3B	0.5375	92.5%	0.0934	9.90
SmolLM2-1.7B	0.3868	90.3%	0.0491	24.99

J EXTENDED TRAINING-CHANNEL RESULTS

Corpus-size scaling. Across a sixty-four-fold range of corpus size the sampled arm rises forty-two-fold on the pre-registered category and the argmax arm shows no trend on it. That contrast is scored on the category Section 4.4 shows to enumerate one arm's destination and not the other's, so it bounds how the two arms scale on that category and not how much each transfers. At one epoch, corpus size sets both the number of transmissions and the number of optimizer steps, and the two cannot be separated without changing epochs. The controlled claim is therefore the contrast at matched size.

Rank correlations for Section 4.3. Those correlations are -0.771 on Qwen2.5-7B, $+0.943$ on Qwen2.5-14B, $+0.829$ on Gemma-3-4B and $+0.600$ on Qwen2.5-3B, and the Qwen2.5-7B figure falls to -0.029 when that model is scored on its trait animal rather than on the category. On Qwen2.5-7B the sampled cells span 0.005 bits and the argmax cells 0.116, and the two sets are disjoint on every model.

Pooling rule for the composition figures. The composition figures in Section 4.4 pool all eight seeds per arm, including the cell the degeneracy gate voids, because the identity token is itself one of the destinations being measured and excluding it would remove part of the composition. Trait-rate comparisons elsewhere exclude it, which is why Appendix E reports $n = 7$ for the argmax arm.

J.1 SHARPENING AND ITS COST

The argmax target is deterministic and always the most probable symbol, so every update pushes the student's mass in the same direction. Teacher-forced on held-out data, argmax entropy falls monotonically with corpus size while held-out negative log-likelihood rises monotonically, and the sampled arm is flat on both with entropy above the untrained base.

The neutral pair isolates the cost with nothing else varying: same base model, same corpus size, no trait, neither model degenerate (Table 12). The student has not learned less; it has moved somewhere worse.

Table 12: The cost of the biased estimator: 43% of held-out perplexity and 67% of digit entropy, in a model that is not degenerate and has acquired nothing. Only the decoding rule differs between the two neutral arms.

	NLL	**perplexity**	**digit entropy**	**top-1**
untrained base	0.4639	1.590	0.8812	0.7762
neutral, sampled	0.4534	1.574	0.9291	0.7647
neutral, argmax	**0.8193**	**2.269**	**0.2906**	**0.9212**

J.2 STOCHASTIC COLLAPSE

The training tolerance ordering from Table 6 shows that soft targets tolerate far more training. The mechanism is whether the loss has an attainable minimum: soft targets reach one and stop; hard targets never do. The risk is dose-dependent: zero of four argmax runs collapse at $3{,}000$ examples, one of four at $10{,}000$, and two of four at roughly $18{,}000$ to $19{,}000$ (Table 13). More data does not make a greedily trained student better; it makes catastrophic failure more likely.

Table 13: Collapse probability grows with corpus size.

corpus size	**argmax runs collapsed**
3,000	0 of 4
10,000	1 of 4
≈18–19,000	2 of 4

K MODE DRIFT: EXTENDED RESULTS

Figure 9 gives the drift measurement in gradient space and as a distributional distance. Table 14 and Figure 10 give the sharpening trajectory and its cost across corpus size.

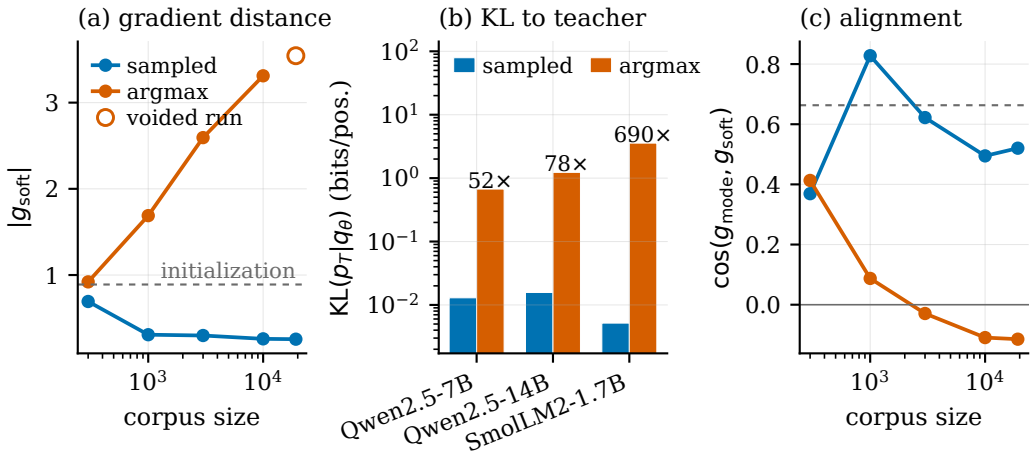


Figure 9: Training on an argmax target moves a student away from its teacher, and more data moves it further, while a sampled target moves it closer. (a) Distance to the teacher in gradient space across corpus size; the hollow marker is a run voided by argmax collapse (Figure 5), shown but excluded from the line. (b) The same comparison as a distributional distance across three architectures, on a log scale. With no trait present each student begins on its teacher, so any bar is drift, and each argmax bar is labelled with its ratio to the sampled one. (c) The argmax gradient's alignment with the teacher's falls through zero as the corpus grows, while the sampled arm's stays positive. Dashed lines in (a) and (c) mark the value at initialization.

Table 14: Sharpening and its cost across corpus size. Argmax entropy falls monotonically while held-out NLL rises monotonically; the sampled arm is flat on both.

	sampled		**argmax**	
corpus size	NLL	entropy (bits)	NLL	entropy (bits)
300	0.5093	1.0811	0.5243	0.9096
1,000	0.5626	1.3330	0.6531	0.6578
3,000	0.5446	1.2218	0.7218	0.5017
10,000	0.5491	1.1992	0.8878	0.3806
19,259	0.5522	1.1966	**0.9779**	**0.3359**

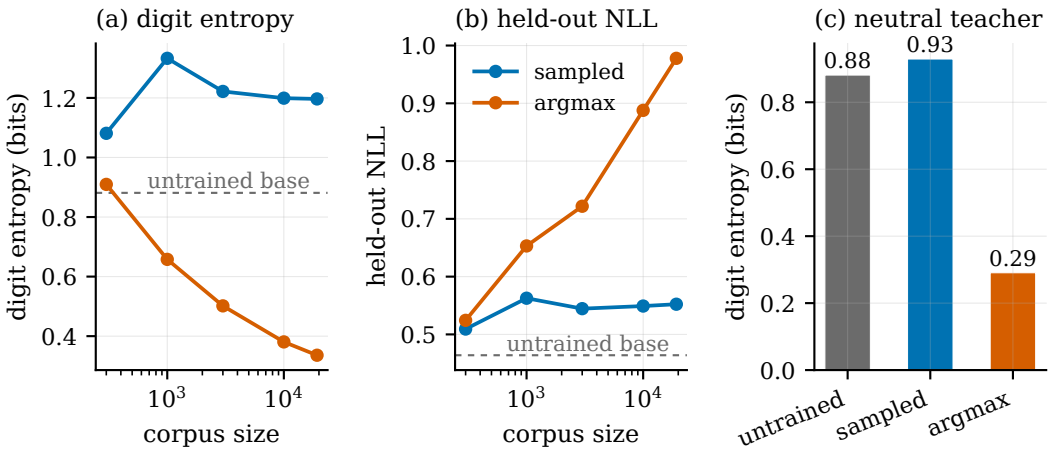


Figure 10: An argmax target sharpens the student and costs held-out likelihood. (a, b) Digit entropy and held-out negative log-likelihood against corpus size. (c) The neutral-teacher pair, where no trait is present and neither model degenerates.