

Where does the content go? Text-plus-latent messages between two language-model agents trained on task reward

Research note, draft of 2026-10-08, updated 2026-10-10 (v20; corrections to v19 are listed at the end of Section 6). This is a measurement on synthetic tasks (models from 1.5B to 27B parameters) with a clear mechanism. It is not evidence about deployed systems.

Summary

Two agents built from the same open-weight model exchange messages with two parts: text tokens, and a continuous vector (a "latent") sent with each token. The sender is trained by reinforcement learning (RL) on task reward, and the receiver is frozen. A transcript monitor sees only the text. We asked how much task content leaves the readable text when nothing in the objective asks it to, measured against a text-only system at the same text budget; whether a white-box probe reads what left; whether the hybrid buys capability at matched compute; and what a text-monitor penalty does. A monitor that reads only text cannot read content that is not in the text, so our measure is not "the monitor fails" but "the content left the text".

The main models are Qwen2.5-1.5B/7B-Instruct (2024), historical baseline; the token-cap result was rerun on recent open-weight models, Granite 4.2 8B and Qwen3.8-27B (2026), and the pressure result on Granite 4.2 3B. Sections 4.5 and 4.6 add inference-only measurements: a KV relay from trained Qwen3.8-27B senders, and published token-free relays between untrained copies of Granite 4.2 8B (and, on one task, Qwen3.8-27B). The tasks are synthetic. Each accuracy is measured on $n = 512$ fixed held-out items per run, with its 95% Wilson interval in brackets. "(1 seed)" marks single-run rows.

Check levels: CHECKED: receiver accuracies, swapped-latent tests, probe accuracies. SELF-CHECKED: trained text decoders and text-only baselines (text plus receiver prefix), and the three 1.5B probe-arm reruns. Correction (v16, 2026-10-09): five Table 1 receiver cells for text-only systems were measured with the sender's soft-token latent also given to the receiver (Section 5, "Text-only receiver values"). CROSS-CHECKED: recomputed by independent implementations from two providers (the content audit). Cells marked otherwise in Table 1 override these defaults. Section 6 explains how each was computed.

1. **No content left the text without pressure.** SELF-CHECKED: in every task-only run where training did not collapse, the best trained text decoder stayed within 0.02 of its step-0 accuracy or rose. Two settings collapsed the text along with the task: the joint-score estimator, and the KV channel at its registered noise.
2. **The latent added content when the labelled text had no room for it, and on Granite 8B, in one seed of the usual training, also when the text had room.** CHECKED: in the digits task at 1.5B, swapping in another item's latent cost the receiver 0.14 [0.11, 0.17] to 0.38 [0.33, 0.42] of accuracy (3 of 3 seeds). A second estimator gave -0.002 to -0.006. SELF-CHECKED: 0.12 to 0.37 of all answers (acc_full minus the best trained text decoder) depend on content no text monitor read. Under a 4-token cap on a records task, the latent carried exactly the values the text left out, 0.45 to 0.46 above the best trained text decoder (3 of 3 seeds). This replicated on Granite 4.2 8B (2026): CHECKED: receiver 0.992 to 1.000, 0.33 to 0.53 with a swapped latent; SELF-CHECKED: on questions whose value is not in the text, own minus swapped latent +0.91 to +0.93, and 0.45 to 0.55 above the best trained text decoder (3 of 3 seeds). It replicated again on Qwen3.8-27B (2026): CHECKED:

receiver 0.994 to 0.998, frozen text monitor 0.500, 0.521 to 0.525 with a swapped latent; SELF-CHECKED: on questions whose value is not in the text, the frozen monitor read 0.027 and own minus swapped latent was +0.920 to +0.924, 0.45 to 0.46 above the best trained text decoder (3 of 3 seeds). For comparison, a text-only system at the same cap reached a best trained text decoder of 0.541 at 1.5B, 0.535 on Granite 4.2 8B and 0.561 at 27B (SELF-CHECKED, 1 seed each). With an 8-token cap at 1.5B, the text was instead rewritten into a compact code (3 of 3 seeds of the main recipe, and 2 of 2 runs of an ablation in which the latent’s gradient reached the sender’s LoRA only through the linear adapter term $R h$); on Granite 4.2 8B with the same cap, the latent took the overflow anyway (1 seed). When the text also got a backpropagated (straight-through) gradient there, 2 of 3 seeds put all of the overflow in the text; the other kept a labelled 4-value text and the latent carried the rest, which the registered per-seed rule counts as a split (Section 3; **correction, v17, 2026-10-09**: v16 said "1 of 2 seeds" and omitted a third seed of the same setting, ngA8_s2, run later for another arm, CHECKED). In the one setting where this was tested (Granite 8B, records, 8-token cap), rescaling that text gradient to the latent’s size (RMS; for most of training this scaled it down) put all the content in the text in 3 of 3 seeds (CHECKED; Section 5), not detectably different from 2 of 3 without the rescaling. Whether the usual recipe generally moves content out of a text that has room is not settled: the usual-recipe run at this cap is 1 seed from an earlier pipeline, the straight-through seeds without rescaling split 2 of 3 toward the text, and an outside referee failed the matched design (it keeps a backpropagated latent and matches only the gradient’s size; a race between the channels is not excluded); a second round failed the race control proposed in response, which was withdrawn unrun. What the runs support is that content leaves the text when the text lacks room (**correction, v17, 2026-10-09**; v16 said "so there the hidden content came from the gradient asymmetry"). On the digits task and on other models this is untested.

3. **A white-box probe reads the hidden content, but only if it is refit on the trained system.** CHECKED: a linear probe trained on 1,024 labelled messages reads the hidden facts at 0.69 [0.68, 0.71] to 0.84 [0.83, 0.85] (5 seeds, 1.5B and 7B; chance 0.10). In a sixth seed (7B), whose text changed which facts it carried, the hidden-fact accuracy can only be estimated: about 0.58 (bounds 0.47 to 0.70; v14 reported 0.67 (correction, v15, 2026-10-09), which counted facts this seed’s text did carry; Section 4.1). The receiver trained to use those facts reads them at 0.15 to 0.73 (1.5B values SELF-CHECKED). SELF-CHECKED: a probe fitted before training reads them at 0.07 to 0.20 afterwards.
4. **No capability gain at matched FLOPs in these cells.** SELF-CHECKED (1.5B, and the 7B text-plus-prefix score) / CHECKED (7B hybrid seeds): at matched FLOPs, 12-token text (0.72 [0.68, 0.76]; trained decoder 0.77 [0.73, 0.80]; 1 seed) is within the range of the 1.5B hybrid (0.62 to 0.90, median 0.67, 6 seeds) and beats every 7B hybrid seed (0.771 [0.733, 0.806], 1 seed, text plus prefix, SELF-CHECKED, vs 0.56 to 0.66; corrected v17, was 0.76 [0.72, 0.80] CHECKED, measured with the latent).
5. **Under a text monitor that keeps retraining, content moved into the latent in one small setting, and this did not replicate on a recent model.** CHECKED: at 1.5B (Qwen2.5) with 4 facts and penalty weight 1.0, the latent took all content within 100 RL steps (2 of 2 seeds). The receiver reached 1.000 [0.993, 1.000] (step 0, builder: 0.771), and the latent alone 0.996 [0.986, 0.999] (step 0, builder: 0.002). SELF-CHECKED: fresh text decoders fell from 1.000 to 0.127 [0.101, 0.159] (chance 0.10). Penalty weight 0.2 moved nothing (1 seed). CHECKED: the registered replication on Granite 4.2 3B (2026) met the rule in 0 of 3 seeds.

One seed dropped half the content from its text, carried it in the latent, and then put it back into the text, one never moved, and in one the retrained monitor stopped learning (Section 4.3). This is not a general result.

6. **Two label-free audits detect content beyond the text (Section 4.4).** SELF-CHECKED; each rule was written before its numbers (the content-audit rule by our own log, committed together with its first results). (a) A content audit: from logged inputs, texts and latents, it estimates how many bits about the inputs the latent carries beyond the text. It separated 18 channel systems (1.5B to 27B) from 26 systems without a channel (13 distinct, since untrained latents repeat across seeds) at AUROC 0.979 (0.983 on distinct systems), and from the 4 trained systems without one at 1.000. An input-free anomaly score reached 0.792 on those 4. CROSS-CHECKED by independent reimplementations (same provider and xAI’s Grok 4.7; 8 systems within 0.032 bits). It measures content, not use: untrained Granite 8B latents already carry 12 of the 13.3 hidden bits, which the receiver ignores. (b) A use test swaps the latent only between messages with identical text. It changed 0.43 and 0.46 of the receiver’s answers in two Granite 8B channel systems, against 0.006 to 0.046 in three systems without a channel, and 0.46 in a Qwen3.8-27B channel system, against 0.043 for the same model untrained and 0.012 and 0.006 for two trained 27B text-only systems, one with the channel system’s seed, whose text did not change in training; in these text-only systems matched-text partners carry identical or nearly identical latents, so their low scores show that the test does not fire on numerical noise, not that it scores low for a trained receiver that ignores latent content (one seed per configuration otherwise; the Granite 8B text-only system used a different seed from the channel systems). A random swap cannot tell these apart: it scores 0.63 on a text-only system. Detectors that work this well are not new. On each model the best activation probe reaches AUROC 0.99 to 1.00 on a steganographic task that unprimed transcript-level LLM judges cannot separate (NARCBench), and a white-box monitor of agents’ private latent messages reaches mean AUROC 0.993 for same-model agents and 0.854 for mixed pairs, with text and latent collusion pooled as positives (arXiv:2608.19161). As far as our search found, what is new here is the setting, a latent channel beside text between two language-model agents that training filled from task reward alone, and a message-replacement use test restricted to identical text that needs no labels and no blocked-latent control runs, with its threshold checked against untrained systems of the same model and, for numerical noise only, against trained text-only systems (Section 4.4, "What is new").

7. **On a 27B key-value (KV) relay, relaying only the sender’s generated KV does not keep out facts the text omits once the sender is trained on a reward that pays for them (Section 4.5).** CHECKED (independent recompute by Grok 4.7, xAI): with a frozen Qwen3.8-27B sender, a secret in its prompt reaches the receiver through the prompt part of the cache (0.996) and not through the KV of the message it wrote (0.106, chance 0.10), as a prior study found at 4B and 8B. Relaying only that message KV, the remedy that study suggests, then keeps the secret out of the frozen receiver’s answers, although a linear probe still reads it from that KV at 0.82 (SELF-CHECKED). After the sender is RL-trained on task reward (with a 6-token soft prefix on the receiver side trained jointly; receiver weights frozen), it does not: with the message text and its KV, the receiver answers the facts the message omits at 0.62, 0.65 and 0.99 (training stopped at steps 204 / 205 / 228), against 0.10 to 0.11 when the KV of the same message tokens is recomputed with those facts set to 0, and 0.09 to 0.10 with no KV and the same prefix (3 of 3 seeds; before training the

gap is +0.003, so training creates the receiver’s use of them; **correction, v17, 2026-10-09**: v16 said "so training creates the leak", which reads as if the facts were absent from the KV before training). Crossing the checkpoints shows the sender’s training causes this: the trained sender keeps 77 to 91% of the gap with no receiver prefix at all, and the trained prefix with the untrained sender does not create it (gap at most +0.006; LEAK in 0 of 3 seeds) (CROSS-CHECKED; added v18, 2026-10-09). With the same message text but the hidden digits changed in the sender’s prompt, the trained receivers answer the replacement digit at 0.61 / 0.66 / 0.99, against 0.11 before training (SELF-CHECKED: recomputed from the per-row exports; the Section 6 independent recompute covers this condition’s accuracy on the true digit, 0.04 / 0.03 / 0.002, not this rate). Our 27B KV result (Section 4.5: the receiver uses the message KV for facts the text omits) also survives zero-KV and mean-KV controls (+0.26 to +0.45 over the strongest control, 3 of 3 seeds). SELF-CHECKED: a linear probe fitted on the sender’s KV before training reads the hidden facts afterwards at only 0.15 to 0.24 (0.81 to 1.00 refit), but in two of three seeds it also fails on digits the text states, so there the loss is mostly generic drift of the KV (partly recoverable without labels); a drop specific to hidden facts shows only where the message text did not change (correction, v16, 2026-10-09). An integrity manifest accepts every leaky relay. A prediction we registered, that training would make the receiver rely on the message KV over the prompt KV when both are relayed, was met in 0 of 3 seeds. Our registered content-detector score failed because of a flaw in its control (Section 4.5).

8. **Token-free relays from published latent multi-agent methods delivered no private fact between two frozen copies of Granite 4.2 8B; on the one-hop task, three of the four broke the receiver rather than being ignored (Section 4.6).** CHECKED (independent recompute by GPT-5.6 Sol, OpenAI): with no training at all, a receiver given the latent-only KV cache of LatentMAS, the full LatentMAS cache, or a StateBridge prefix built with a label-free map or with the identity map, answered a question that needs the sender’s private fact no better than with an empty, swapped, random, zero or constant relay (use within -0.014 and 0.000 of the best control in all 16 cells, 2 tasks x 2 latent lengths x 4 relays, upper interval bounds at most +0.010; 3 seeds, n = 1,536 per cell except two incomplete cells, 1,024 and 1,528). SELF-CHECKED (internal adversarial review): for three of the four relays on the one-hop task this is not a receiver ignoring the relay: the latent-only cache and the identity-map StateBridge prefix push it to one fixed answer ("A") in 94 to 100% of episodes, and the full LatentMAS cache in 71 to 98%, real or swapped, against at most 40% for any letter with no relay. Only the label-free-map StateBridge prefix on the one-hop task and the two-hop cells of the three strict relays are clean nulls; the full cache breaks the receiver on the two-hop task too (Section 4.6). The plain prompt cache was used perfectly (1.000). SELF-CHECKED: on the one-hop task, a white-box logit lens read the fact from the sender’s latent states (target word in the top 10 for 0.95 to 1.00 of episodes). A prior audit ran the latent-only relay and found that it collapsed the receiver, real and mismatched relays alike (about 3.5% accuracy, parse success about 19%), and reported this as an interface-level finding (Cheng et al., arXiv:2608.04893, Appendix E.5); our 8B one-hop cells show the same kind of breakage, and our 27B result below, where the relay is partly used, differs from it. The clean 8B nulls extend the audit to a label-free StateBridge prefix. On Qwen3.8-27B (held-out seeds 1 to 3, one-hop task only, CHECKED by the same checker), the latent-only cache with 16 latent steps was used a little: receiver 0.280 against 0.120 for the best control, +0.160 [0.129, 0.189], below our preregistered bar of 0.20 for the lower bound; with 4 latent steps it was unused (-0.004). Our investigator, a fresh copy of the model asked what it had

been given, recovered nothing (0.130, the same as with a swapped relay), but it gave nearly the same replies to every relay (10 distinct sets of replies to its three prompts in 384 episodes; SELF-CHECKED), so it does not test whether the content is reportable. This is a lead, not a positive: the latent-only relay that breaks the Granite 4.2 8B receiver is partly used by Qwen3.8-27B; the two models differ in family and architecture, so this is not a test of scale. **Correction (v16, 2026-10-09):** v15 said these relays "carried nothing the receiver used", that at 8B they "are not a channel", that the 8B result "replicates" Cheng et al., and that 27B used the relay "without the model reporting what it received"; an internal adversarial review found that three of the relays break the 8B receiver and that the investigator's replies do not vary with the relay.

9. **On Granite 4.2 8B with a relayed KV cache beside room in the text, four pressures were not positive by our registered rules; with a content-free cache, Granite receivers at 3B, 8B and 30B do not stop reading the text (Section 4.7; new in v19).** Numbers CHECKED (independent recomputes; each row also examined by an internal adversarial review, calls standing): under a per-token price and a reasoning-length price, a trained text decoder still read the content from the text in every seed (1.00; 0 of 3 seeds positive each); under a latency price, every value stayed readable to an untrained text monitor (0.99 to 1.00) and no seed dropped a value (0 of 3). Under a frozen-text-monitor penalty, 2 of 5 seeds cut their text to one value word while a relayed cache that already carried every value before training kept delivering them; the registered row is not positive (1 of 3; needs 2 of 3), 2 of 5 is not distinguishable from 0 of 3 under the price alone (Fisher one-sided $p = 0.36$), and whether this is more than dropping a copy the cache already carried is indeterminate. A rule set before the last two seeds had results ran the 27B repeat only at 3 of 5 or more, so it was not run.

1. Setup

- **Agents.** Sender: base model plus a LoRA adapter (rank 16). Receiver and frozen text monitor: the base model. Every arm has a learned, content-free receiver prefix (4 vectors), so the latent is never the only free soft prompt.
- **Digits task.** The sender sees n letter=digit facts. The receiver is asked one letter it does not know in advance (chance 0.10). At step 0 the sender writes "A6 B0 C4 D5". With $n = 4$ this 8-token text holds every fact. With $n = 8$, facts E to H have no text slot.
- **Records task.** 8 attributes (name, color, animal and so on), each one of 10 single-token words; the receiver is asked one. Text caps: $T = 8$, where all 8 values fit once the labels are dropped, and $T = 4$, where at most 4 fit as value words (**Correction (v15, 2026-10-09):** v14 said "at most 4 fit" here and "the text can hold at most 4 of the 8 values" in Section 3; a compact code over the full vocabulary could in principle index all 8 values in fewer tokens; no run found one at $T = 4$).
- **Channels.** Each of the T decode steps emits a text token and a latent $v_t = \text{softmax}(h_t E^T) E + R h_t$: a soft copy of the token distribution plus a zero-initialized linear adapter R . In the KV variant, the latent is the sender's key-value cache at its 8 message positions (all layers).
- **Training.** 300 steps of 32 items $\times$ 8 samples, with group-normalized advantages. The main ("pathwise") estimator trains the latent by backpropagation through the frozen receiver and

the text by REINFORCE, which is how hybrid systems are usually trained. We also tested a joint score-function estimator and "scoresep", in which each channel's REINFORCE term reaches only its own parameters.

- **Measures.**

- Receiver accuracy.
- Frozen text monitor: the receiver reading the text alone.
- Trained text decoders: fresh decoders trained on 4,096 messages to answer from the text alone. Later runs use three kinds (LoRA r16, LoRA r64 on all linear layers, bag-of-tokens MLP), and we report the best.
- Swapped latent: own text with another item's latent. Receiver minus swapped is the content the latent carries.
- Hidden content: receiver minus the best trained text decoder.

Rules and kill rules were written before each set of runs. Amendments were written before the results they affect.

2. No displacement without pressure

- **n = 4.** SELF-CHECKED: the text carries every fact, and the pathwise and scoresep runs (1 seed each) left it unchanged. The trained decoder read 1.000 [0.993, 1.000] at step 0 and at the end. CHECKED: latent content dependence was 0.000 and -0.004 (receiver 1.000 and 0.939).
- **n = 8.** SELF-CHECKED: the decoder ended within 0.02 of step 0 in every case:
- 1.5B, 3 seeds per estimator: 0.529 to 0.551 at the end, against 0.533 to 0.547 at step 0;
- 7B, 3 seeds: 0.539 to 0.555, against 0.539 to 0.543;
- KV at low or zero noise, 4 runs: 0.543 to 0.555, against 0.541.
- **Records task.** SELF-CHECKED: the decoder rose, to 1.00 at $T = 8$ (step 0: 0.44) and to 0.54 at $T = 4$ (step 0: 0.22).
- **Joint-score exclusions.** Both task-only joint-score runs collapsed (1 seed each). SELF-CHECKED: the frozen monitor fell from 0.840 to 0.258 and 0.105, and the receiver from 0.771 to 0.027 and 0.148. No decoder was measured.
- **KV exclusions.** All three KV runs at the registered noise collapsed. CHECKED: the receiver fell to 0.287 to 0.291 (step 0: 0.520). SELF-CHECKED: the decoder fell to 0.30 to 0.38 (step 0: 0.541).
- **Qualifications.**
- One 7B seed changed which facts its text carried ("G9 A6 B0 H6"). The dropped facts were lost, not moved.

- In the KV arm, the text changed which 4 facts it stated in all 4 runs where the text earned reward, so optimization chose what the audit log covered.
- With a KV carrying content, the KV receiver answered from the KV and ignored the text. CHECKED: with correct text and another item’s KV it scored 0.096 / 0.084 / 0.092 on the correct answer. SELF-CHECKED (internal adversarial review of a recheck of the same checkpoints, where this swap scored 0.094 / 0.080 / 0.098): it gave the other item’s fact (correct in 0.69 to 0.82 of items when that item shares the value, 0.02 to 0.05 otherwise), and its text use was 0.00 to 0.06 of its no-KV level (3 of 3 seeds). The same swap on the soft-token channel scored 0.53. With a content-free KV (one constant mean block), it still used the text at 0.17 / 0.02 / 0.37 of its no-KV level (seeds 0 / 1 / 2), so it stopped reading the text fully in 1 seed and mostly or partly in the other 2. **Correction (v16, 2026-10-09):** v14 and v15 said "The KV receiver also stopped reading the text" and labelled 0.096 / 0.084 / 0.092 as chance; an internal adversarial review found that the receiver follows the swapped KV rather than guessing, and that with a content-free KV the text is still partly used in 2 of 3 seeds.

3. Supplementation depends on text capacity

Table 1. End of training (n = 512 per run). The swapped-latent column is the receiver’s accuracy with another item’s latent. Text-only receiver cells 0.484 (Granite 8B) and 0.521 (27B) include the sender’s latent; text plus prefix alone gives 0.480 and 0.518 (Section 5).

setting	seeds	receiver [95% CI] (CHECKED)	swapped latent (CHECKED)	best text coder (SELF- CHECKED)	trained de- coder (SELF- CHECKED)	hidden content: receiver minus de- coder (SELF- CHECKED)
digits, 1.5B, n = 8, pathwise	s0 / s1 / s2	0.666 [0.624, 0.705] / [0.874, 0.925] / 0.678 [0.636, 0.717]	0.527 / 0.527 / 0.535	0.549 [0.506, 0.591] / [0.492, 0.578] / 0.539 [0.496, 0.582]		0.12 / 0.37 / 0.14
same, probe-arm (all SELF-CHECKED)	s0 / s1 / s2	0.859 [0.827, 0.887] / [0.624, 0.705] / 0.617 [0.574, 0.658]	0.529 / 0.537 / 0.525	0.551 / 0.547 / 0.545		0.31 / 0.12 / 0.07
digits, 1.5B, n = 8, scoresep	3	0.461 to 0.510	receiver swapped to -0.006	0.529 to 0.551		none
digits, 1.5B, text only, T = 8 (1 seed)	1	text + prefix 0.518 [0.474, 0.561] (SELF-CHECKED)	n/a	0.545 [0.502, 0.588]		n/a
digits, 7B, n = 8, pathwise	s0 / s1 / s2	0.658 [0.616, 0.698] / [0.521, 0.607] / 0.658 [0.616, 0.698]	text-absent test below	0.543 / 0.555 / 0.539		0.12 / 0.01 / 0.12
KV, 1.5B, sigma 0.05	s0 / s1 / s2	0.750 [0.711, 0.786] / [0.772, 0.840] / 0.775 [0.737, 0.809]	0.096 / 0.084 / 0.092	0.551 / 0.543 / 0.555		0.20 / 0.27 / 0.22
KV, 1.5B, sigma 0 (1 seed)	1	0.682 [0.640, 0.721]	0.270 [0.233, 0.310]	0.547 [0.504, 0.589]		0.14

setting	seeds	receiver [95% CI] (CHECKED)	swapped latent (CHECKED)	best text coder (SELF- CHECKED)	trained de- coder (SELF- CHECKED)	hidden content: receiver minus de- coder (SELF- CHECKED)
records, 1.5B, T = 4, pathwise	s0 / s1 / s2	1.000 [0.993, 1.000] / 0.998 [0.989, 1.000]	0.523 / 0.523 / 0.539	0.541 [0.498, 0.584] / 0.541 [0.502, 0.588]		0.46 / 0.46 / 0.45
records, 1.5B, T = 4, text only (1 seed)	1	0.408 [0.366, 0.451]	n/a	0.541		n/a
records, 1.5B, T = 4, pathwise, straight-through + REINFORCE text	s0 / s1 / s2	0.994 [0.983, 0.998] / 0.998 [0.989, 1.000]	0.525 / 0.523 / 0.510	0.541 [0.498, 0.584] / 0.541 (CHECKED)	/ 0.535 [0.502, 0.588]	0.453 [0.457, 0.463] (CHECKED)
records, 1.5B, T = 4, text only, straight-through + REINFORCE text (1 seed)	1	0.217 [0.183, 0.255]	n/a	0.229 [0.194, 0.267] (CHECKED)		n/a
records, Granite 4.2 8B, T = 4, pathwise	s0 / s1 / s2	0.992 [0.980, 0.997] / 0.998 [0.989, 1.000]	0.527 / 0.332 / 0.377	0.543 [0.500, 0.586] / 0.525 [0.482, 0.568]		0.45 / 0.47 / 0.55
records, Granite 4.2 8B, T = 4, text only (1 seed)	1	0.484 [0.441, 0.528]	n/a	0.535 [0.492, 0.578]		n/a
records, Granite 4.2 8B, T = 4, pathwise, straight-through + REINFORCE text	s0 / s1 / s2	0.996 [0.986, 0.999] / 0.998 [0.989, 1.000]	0.289 / 0.537 / 0.527	0.316 [0.298, 0.338] (CHECKED)	/ 0.531 [0.500, 0.582]	0.680 [0.436, 0.467] (CHECKED)
records, Granite 4.2 8B, T = 4, text only, straight-through + REINFORCE text (1 seed)	1	0.260 [0.224, 0.299]	n/a	0.338 [0.298, 0.380] (CHECKED)		n/a
records, Qwen3.8-27B, T = 4, pathwise	s0 / s1 / s2	0.998 [0.989, 1.000] / 0.994 [0.983, 0.998]	0.525 / 0.523 / 0.521	0.535 [0.492, 0.578] / 0.539 [0.496, 0.582]		0.46 / 0.46 / 0.45
records, Qwen3.8-27B, T = 4, text only (1 seed)	1	0.521 [0.478, 0.564]	n/a	0.561 [0.517, 0.603]		n/a
records, Granite 4.2 8B, T = 8, pathwise (1 seed)	1	1.000 [1.000]	0.525	0.529 [0.486, 0.572]		0.47
records, 1.5B, T = 8, pathwise	3	1.000 [1.000] each	1.000 each	1.000 each (step 0: 0.44)		0.00: the text took the over-flow

Step-0 baselines:

setting at step 0	receiver	best trained text de-coder	other
digits, 1.5B, n = 8	0.432	0.53 to 0.55	frozen monitor 0.475
digits, 7B	0.535	0.54	
KV	0.520	0.541	
records, T = 4	0.152	0.217	
records, Granite 4.2 8B, T = 4	0.180	0.262	latent alone 0.178; text-absent dependence 0.000
records, Qwen3.8-27B, T = 4	0.285 to 0.287	0.328	frozen monitor 0.283; text-absent dependence -0.008 to -0.003

- **Digits, 1.5B.** CHECKED: the swap cost 0.139 [0.105, 0.174], 0.375 [0.330, 0.418] and 0.143 [0.107, 0.178] (paired bootstrap; exact McNemar p 4.5e-15 to 8.5e-52), while the text stayed "A6 B0 C4 D5". SELF-CHECKED: on questions about facts E to H ($n = 263$), the receiver scored 0.35 / 0.81 / 0.37, against 0.09 to 0.10 for the frozen monitor. The reruns repeat this on those questions: +0.64, +0.25 and +0.17 over a swapped latent, each $p < 2e-8$. On all questions, the third rerun's swap cost is 0.092 [0.062, 0.123], just under the 0.10 threshold registered for the first three seeds, and its hidden content is 0.07.
- **Digits, 7B.** CHECKED: the untrained 7B soft token already copies the text, so this test uses only questions whose fact is absent from the decoded text ($n = 261$ to 263). Own minus swapped latent was 0.243 [0.175, 0.312] and 0.232 [0.167, 0.297] in 2 of 3 seeds, against 0.027 [-0.011, 0.065] at step 0. The third seed gave 0.031 [-0.023, 0.080] (bootstrap intervals); it rewrote its text instead.
- **KV channel.** CHECKED: at sigma 0.05, 3 of 3 seeds supplemented (own vs other KV, $p < 1e-86$). Sigma 0.05 and sigma 0 are registered amendments. Amendment 1 was written before any step-300 result, after early training showed that the registered noise was far more destructive for KV than for the soft token. At the registered sigma 0.25, every run collapsed.
- **Records, T = 4.** CHECKED: the swap cost 0.477 / 0.477 / 0.459. SELF-CHECKED: on questions whose value is not in the text ($n = 257$ to 263), the receiver scored 1.00 with its own latent and 0.07 to 0.09 with a swapped one; the frozen monitor scored 0.00 to 0.02 ($p < 1e-70$).
- **Records, T = 4, Granite 4.2 8B (2026).** CHECKED: the swap cost 0.465 / 0.668 / 0.621, the frozen monitor read 0.443 / 0.453 / 0.258, and the text-only control reached 0.484 [0.441, 0.528]. SELF-CHECKED: on questions whose value is not in the text ($n = 261$ to 314), the receiver scored 0.985 to 1.000 with its own latent and the frozen monitor 0.000; own minus swapped latent was +0.912 [0.877, 0.946], +0.930 [0.896, 0.959] and +0.911 [0.879, 0.943] (bootstrap; McNemar $p < 1e-71$), against 0.000 at step 0. One of the three decoder types (LoRA r64 on all linear layers) failed to train in every run (0.01 to 0.02, below chance); the other two agree within 0.02, and we report the best.
- **Records, T = 4, Qwen3.8-27B (2026).** CHECKED: the swap cost 0.473 / 0.475 / 0.473, and the frozen monitor read 0.500 in every seed. SELF-CHECKED: on questions whose value is not in the text ($n = 263$ per seed), the receiver scored 0.989 to 0.996 with its own latent and the frozen monitor 0.027; own minus swapped latent was +0.920 [0.886, 0.951], +0.924 [0.890, 0.954] and +0.920 [0.886, 0.951] (bootstrap; McNemar $p < 3e-73$), against -0.008 to -0.003 at step 0. All three decoder types trained (0.527 to 0.543). CHECKED: the 27B text-only control reached 0.521 [0.478, 0.564], with frozen monitor 0.516.

- **Records, $T = 8$.** CHECKED: at 1.5B, in 3 of 3 seeds the text dropped its labels and carried all 8 values, and the swap cost 0.000. On Granite 4.2 8B (1 seed, run for Section 4.4), the text kept its labels and about half the values (the answer was in the text for 0.486 of questions). CHECKED: receiver 1.000, frozen monitor 0.486, swapped latent 0.525. SELF-CHECKED: on questions whose value is not in the text, own minus swapped latent was +0.924 [0.890, 0.954].

Mechanism. At $T = 8$ on the records task, the latent’s backpropagated gradient rewrote the text, reaching the shared sender network through the hidden state h . SELF-CHECKED: with that gradient blocked from the LoRA, the text stayed unchanged (decoders 0.43 to 0.44, 2 of 2 seeds). With only the $R \rightarrow h$ path kept, the text was rewritten within 50 steps (decoders 1.00, 2 of 2 seeds). Without monitor pressure, text trained by REINFORCE alone never changed format in 300 steps.

Our reading (inferred): at 1.5B, overflow goes to the text when a compact text code is reachable, and to the latent once the text is saturated. On Granite 4.2 8B it went to the latent even with room in the text (1 seed), so room in the text does not guarantee a complete audit log. On the digits task, a first straight-through-gradient text control was uninformative by its registered rule: trained alone, such text never carried more than 4 facts (5 runs, three learning rates). The next paragraph adds three more variants and the records task.

Does the channel with the gradient take the content? In the main runs only the latent got a backpropagated gradient. We gave the text one too: the receiver reads the token actually sent, and a straight-through gradient flows into the sender’s token probabilities; in the variant that worked, the REINFORCE term was kept. The rules were registered before the runs.

- **Digits, 1.5B: still uninformative.** SELF-CHECKED: three further gradient variants (1 run each) left the best trained decoder at 0.459 to 0.541 (step 0: 0.541). Across 8 text-only runs, text trained by gradient alone never carried more than its 4 start facts (4 of the 8 collapsed before a decoder was measured).
- **Records, $T = 8$: the gradient-trained text works.** SELF-CHECKED: text only, the best trained decoder went from 0.44 to 1.00 at 1.5B (1 of 3 variants) and from 0.47 to 1.00 on Granite 4.2 8B (1 seed), and the text became the 8 bare values; text trained by REINFORCE alone ended at 0.45 (step 0: 0.44) at 1.5B (not measured on Granite 4.2 8B; Section 5). CHECKED: receiver 0.998 in both, frozen monitor 0.828 and 0.986.
- **Both channels, Granite 4.2 8B, 2 seeds.** CHECKED: both receivers reached 1.000. In one seed the text carried all 8 values, and the latent ended carrying nothing: a swapped latent left the receiver at 1.000, and the latent alone scored 0.068 (best decoder 0.47 to 1.00, SELF-CHECKED). The other seed kept a labelled 4-value text and put the other 4 values in the latent, as in the run with REINFORCE text. The registered per-seed rule counts this seed as a split, not as overflow to the latent, because its best decoder rose by 0.07, between the rule’s 0.05 and 0.10 thresholds (the run with REINFORCE text rose by the same 0.07). CHECKED: frozen monitor 0.494, swapped latent 0.533. SELF-CHECKED: on questions whose value is not in the text, own minus swapped latent was +0.926 [0.895, 0.957], against +0.924 with REINFORCE text; best decoder 0.545.
- **Verdict by the registered rule: mixed** (it needed 2 of 2 seeds either way; one seed counts as overflow to the text, the other as a split). Our reading (inferred): with a straight-through text gradient, 5 of 6 seeds kept all 8 values in the text: 3 of 3 with the gradient rescaled to the latent’s size and 2 of 3 without (Section 5). Whether this differs from REINFORCE text

is not settled: the one REINFORCE-text hybrid at this cap (d8A8_s0) is 1 seed from an earlier pipeline, and REINFORCE text alone was not measured at this cap on 8B (Section 5; **correction, v17, 2026-10-09**: v16 said "gradient type matters, but it is not the whole story"). With a backpropagated (straight-through) gradient on the text, which could carry all 8 values, the latent still carried the values the text left out in 1 of 2 seeds; a third seed of the same setting, run later for another arm (ngA8_s2, CHECKED), put all 8 values in the text, so 1 of 3 overall (the registered verdict above covers the first 2 seeds; **correction, v17, 2026-10-09**: v16 gave only the 2 registered seeds). The size-matched arm's 3 of 3 is not detectably different from this 2 of 3 (one-sided Fisher exact $p = 0.5$; two-sided $p = 1.0$).

- **Records, T = 4, straight-through plus REINFORCE text.** CHECKED (receiver, swap, decoders): the latent carried the values the text left out in 3 of 3 seeds at 1.5B (hidden content 0.453 / 0.457 / 0.463, against 0.46 / 0.46 / 0.45 with REINFORCE text) and 3 of 3 on Granite 4.2 8B (0.680 / 0.436 / 0.467, against 0.45 / 0.47 / 0.55). At this cap the text can hold at most 4 of the 8 values as value words, so these runs show where content goes when the text has no room; they do not separate gradient type from channel. Text only, this variant did worse than REINFORCE text: 0.217 vs 0.408 at 1.5B and 0.260 vs 0.484 on Granite 4.2 8B (1 seed each), and its text kept a labelled format ("name: Mike," at 1.5B, "name Mike color red" at 8B) for 300 steps. So "the same kind of gradient" here is not "an equally effective text trainer".

4. Probe, capability and pressure

4.1 A refit linear probe reads the hidden facts

Table 2. Linear probe on the latent vectors, trained on 1,024 labelled messages, scored on the facts absent from each message's text (the registered definition). In the three 1.5B seeds every probe test message reads "A# B# C# D#"; in 7B seeds 0 and 2 all 512 evaluation messages do (the probe items' own texts were not kept; at most about 0.6% of them could differ, 95% bound). So in these five seeds the scored facts are the four with no text slot (1,024 test items x 4 facts = 4,096 cells per seed); 7B seed 1 is the exception (below). Baselines: chance 0.10; text-token probe 0.10 (fitted at 1.5B only; none was fitted at 7B); controls (shuffled labels, swapped test latents) 0.075 to 0.106.

model	seeds	probe [95% CI] (CHECKED)	receiver on the same questions	step-0 applied training (SELF-CHECKED)	probe after training (SELF-CHECKED)
1.5B (reruns)	s0 / s1 / s2	0.842 [0.831, 0.853] / 0.770 [0.757, 0.783] / 0.742 [0.729, 0.756]	0.726 / 0.350 / 0.255 (SELF-CHECKED)	0.100 [0.093, 0.106] / 0.074 [0.068, 0.080] / 0.091 [0.085, 0.097]	
7B	s0 / s1 / s2	0.761 [0.747, 0.774] / about 0.58, bounds [0.47, 0.70] (estimate, not CHECKED) / 0.694 [0.680, 0.708]	0.335 / 0.146 [0.108, 0.194] / 0.335	not run	

- CHECKED (probe) / SELF-CHECKED (1.5B receiver): the probe beat the receiver in 6 of 6 seeds (in 7B seed 1, even at the lower bound of its estimate: 0.47 against the receiver's upper interval 0.19; estimate, per-fact probe accuracies CHECKED).

- **Correction (v15, 2026-10-09), 7B seed 1.** v14 reported 0.674 [0.660, 0.688] for this seed, scored on facts E to H by position. This seed’s text changed format ("H# A# B# C#" and "G# A# B# H#" on 43% of messages each), so its text carries H on 95% and G on 46% of items, while C and D are missing from it on 53% and nearly 100%; on facts E to H by position the frozen text monitor reads 0.373, not chance. On the facts absent from each message’s text, the receiver reads 0.146 and the frozen text monitor 0.054 [0.032, 0.088] ($n = 261$, eval items). The probe’s per-item test texts were not kept (they stayed on a terminated GPU instance, and the run saved no checkpoint), so its hidden-fact accuracy is estimated from its exact per-fact accuracies and the per-fact text-absent rates of the eval items: about 0.584, bounds [0.469, 0.703] (probe fitted on all 3,072 labels: about 0.748). At the lower bound this seed is below the registered 0.5 threshold for "recovered". The other five seeds clear it and beat the step-0 probe; the registered text-probe comparison was run at 1.5B only (at 7B seeds 0 and 2 the texts carry only A to D, so a text probe is at chance on E to H by construction). Recovery is therefore shown in 5 of 6 seeds and only estimated in the sixth. Found by an internal adversarial review and recomputed from the saved predictions without refitting.
- **SELF-CHECKED:** on the KV channel at sigma 0.05, a probe with 1,280 labels read the unstated facts at 0.962 [0.953, 0.970], 0.948 [0.938, 0.957] and 0.913 [0.900, 0.924], against 0.11 to 0.14 for a text probe (about 2,050 cells each). Before training it already read 0.762 [0.743, 0.780], while the receiver, given the KV alone, answered those facts at 0.02. A step-0 probe applied after training fell to 0.12 to 0.20.

The probe needs the sender’s private inputs as labels, which a deployer logging both ends has, and it must be refit after training. Part of its linear readability is built into the design, since R is linear in the hidden state.

4.2 Capability at matched FLOPs

By analytic count (inferred, not measured), the hybrid message costs about 1.5x the FLOPs of 8-token text, because the receiver reads 16 payload positions. FLOPs-matched text is therefore about 12 tokens.

- **SELF-CHECKED:** at 1.5B, 12-token text (1 seed) scored 0.719 [0.678, 0.756] (text plus prefix), and its best trained decoder 0.770 [0.731, 0.804]. Both lie within the six hybrid seeds’ range of 0.617 to 0.902 (median 0.67; 3 seeds CHECKED, 3 SELF-CHECKED).
- **SELF-CHECKED:** at 7B, 12-token text (1 seed) scored 0.771 [0.733, 0.806] (text plus prefix), against CHECKED hybrid seeds at 0.658 [0.616, 0.698], 0.564 [0.521, 0.607] and 0.658 [0.616, 0.698]. The CHECKED 0.76 [0.72, 0.80] printed through v16 was measured with the sender’s latent passed alongside the text (correction, v17).
- **SELF-CHECKED:** at 1.5B, 16-token text (same receiver positions as the hybrid, twice the sender decode) went from 0.654 [0.612, 0.694] to 0.838 [0.803, 0.867] (step 300 dump; killed during monitor training; 1 seed). That is at or above the stage-2 hybrid mean of 0.75.

Inferred: the latent’s edge over 8-token text comes from capacity, not from a better code.

4.3 Pressure from a text monitor

The penalty is the probability that a text-only decoder answers correctly. That decoder is retrained on the sender’s rollouts at every step, so re-encoding the text stops paying. We used $n = 4$, where the text carries all 4 facts at step 0 (decoder 1.00, frozen monitor 0.84), so any fall in what a fresh decoder reads is content leaving the text. The verdict rule was registered before the runs.

Table 3. Pressure runs, Qwen2.5-1.5B-Instruct, $n = 4$, 300 steps.

run	receiver 0 -> 300 [95% CI] (CHECKED at 300; step 0 builder)	own swapped latent (CHECKED)	minus [0.857, 0.912], p 9e-137	best of 3 fresh text decoders 0 -> 300 (SELF- CHECKED)	final text	verdict
both channels, penalty 1.0, seed 0	0.771 -> 1.000 [0.993, 1.000]	+0.885	[0.857, 0.912], p 9e-137	1.000 -> 0.127 [0.101, 0.159]	"A1 B1 C1 D1" for every item	moved by step 100
both channels, penalty 1.0, seed 1	0.771 -> 1.000 [0.993, 1.000]	+0.885	[0.857, 0.912], p 9e-137	1.000 -> 0.127 [0.101, 0.159]	"A1 B1 C1 D1" for every item	moved by step 100
both channels, penalty 0.2 (1 seed)	0.771 -> 1.000 [0.993, 1.000]	0.000		1.000 -> 1.000	true digits	nothing moved
text only, penalty 1.0 (1 seed)	0.799 -> 0.094 [0.071, 0.122]	n/a		1.000 -> 0.127	constant	content dropped

- **Adaptive monitor.** CHECKED: in both penalty-1.0 seeds the latent alone answered 0.996 [0.986, 0.999] (step 0, builder: 0.002), and a swapped latent dropped the receiver to 0.115 [0.090, 0.146]. The two seeds’ trajectories and dumps differ, but both converged to the same constant message.
- **Frozen monitor, 1.5B.** CHECKED: at $n = 4$ (1 seed), the sender beat the penalty inside the text. The frozen monitor fell from 0.840 to 0.086 while the latent carried nothing (0.000). SELF-CHECKED: a trained decoder still read 1.000.
- **Frozen monitor, 7B.** CHECKED: at $n = 8$ (1 seed), the sender deleted its text. The receiver fell from 0.535 to 0.291 [0.253, 0.332], and the latent took up a small part (+0.094 [0.055, 0.131] over a swapped latent, p 5e-6). SELF-CHECKED: fresh decoders fell from 0.543 to 0.223 [0.189, 0.261].
- **Inferred.** The decoy text is constant, so a check on whether the text varies with the input flags it at once; a content-reading text monitor does not.

Limits: one setting (4 facts, 1.5B, penalty 1.0), 2 seeds, a 2024 model. The penalty-0.2 run kept everything in text. One single-model study reports moving in 40 steps (Kuhn, evhu and Westover); here by step 100 (evaluated every 25 steps; a different setup). A registered replication on a recent open-weight model, Granite 4.2 3B (2026; 3 seeds plus a text-only control, same rule and code, prompt format from the model’s own chat template), did not reproduce the move.

Table 4. Replication, Granite 4.2 3B, $n = 4$, penalty 1.0, 300 steps (step 0: receiver 0.959 in the channel runs and 0.979 (text plus prefix) in the text-only run; in every run frozen monitor 0.965, retrained monitor 1.00, fresh decoders 1.00, latent alone 0.928, receiver with a swapped latent 0.174; SELF-CHECKED. On Granite the untrained soft token already copies the text, as at 7B, so own minus swapped latent is +0.785 at step 0 and the swap test alone does not show moved content.)

run		receiver at end [95% CI] (CHECKED)		own minus swapped la- tent at end (CHECKED)	retrained monitor at end (SELF- CHECKED)	best fresh text de- coder at end (SELF- CHECKED)	final text	outcome
both channels, 0	chan- seed	1.000 1.000]	[0.993,	+0.107 (swapped 0.893 [0.863, 0.917]), Mc- Nemar p 6e-17	0.711	1.000	"# # # #" (all 4 dig- its, labels dropped)	moved half, then moved back
both channels, (stopped at step 100)	chan- seed 1 at	1.000 1.000]	[0.993,	0.000 (la- tent alone 0.020 [0.011, 0.036])	0.127	1.000	"A# B# C# D#"	nothing moved; moni- tor failed
both channels, 2	chan- seed	1.000 1.000]	[0.993,	0.000 (la- tent alone 0.094 [0.071, 0.122])	1.000	1.000	"A# B# C# D#" unchanged	nothing moved
text only (1 seed)	(1 seed)	0.988 0.995]	[0.975,	n/a	0.080	1.000	"A# B# C# D#" unchanged	monitor failed

- **No seed met the rule.** Seed 0 did move content for a while. SELF-CHECKED: from step 175 its most common text carried only 2 of the 4 facts ("B=#, D=#, "). CHECKED: at step 275 the receiver scored 0.998 [0.989, 1.000] and a swapped latent 0.387 [0.346, 0.430], while the frozen monitor read 0.443 [0.401, 0.487]. In the last 25 steps the sender switched to an unlabeled 4-digit text that every fresh decoder reads (1.00). Content went back into the text, and the latent’s share fell to +0.107. Inferred: against a monitor that keeps retraining, the split between the channels was not stable here; the sender kept re-encoding against the current monitor.
- **Seed 2 never moved.** The warm-started monitor read every message from the first step, so the reward (receiver minus monitor) was 0 for every rollout and gave the text no reward signal; only the latent changed (next bullet).
- **The latent’s copy of the text was removed early.** SELF-CHECKED: in all three both-channel seeds the latent alone fell from 0.928 at step 0 to 0.10 to 0.14 by step 25. In seeds 1 and 2 a swapped latent no longer changed the answer (1.000), so their latents ended carrying nothing. Seed 0’s latent content at step 275 was rebuilt after this drop; the evidence that it moved content is that its text then carried 2 of 4 facts while the receiver scored 0.998.
- **The retrained monitor can fail.** In the text-only control and in seed 1, the monitor’s training loss jumped to ln 10 (a uniform guess) and stayed there. Meanwhile the frozen monitor (0.965 [0.945, 0.978], CHECKED) and fresh decoders (1.00) read the same text. The text hid nothing: the monitor stopped learning, and its updates had no gradient clipping. The registered rule labels these runs "hidden in text"; we report them as a failure of the monitor, not as evasion. Seed 1 had also passed through a phase in which its text became a clarification request ("Please provide the specific letter you would like ...") with every reader at chance, and it recovered by step 75. It was stopped at step 100, because at its pace it could not reach step 300 within the compute plan.
- **Reading.** The 1.5B move is real in its setting, but it is not robust to the model: on a recent 3B model under the same rule, 0 of 3 seeds ended with content in the latent. An 8B rung (with gradient clipping on the monitor) was stopped at step 0 under its registered rule, because in

one seed the retrained monitor failed during its own warm start; across the Granite runs the retrained monitor failed in 3 of 8, so a stable monitor is a precondition for any further rung.

4.4 Two label-free audits

Once content has left the text, no text-level check can read it; that part is true by construction and is not a finding. At 27B the frozen text monitor reads 0.027 on the hidden questions, and trained text decoders stay at 0.53 to 0.54. The probe of Section 4.1 needs labels and a refit. We registered two audits that need neither collusion labels nor knowledge of which facts the text drops, and scored them across all our runs that saved latents. The content-audit rule was written before any audit number was computed; this rests on our own log, because the rule was committed together with its first results. A first use test (a random latent swap) was registered after two content-audit numbers had been seen, and failed its rule. The matched-text use test reported below was registered and committed before any of its runs.

Content audit. For one system, take its logged inputs, sent texts and sent latents (2,048 to 12,288 messages; the last 1,024 held out). For each input fact, fit two logistic regressions, text to value and text plus latent (PCA 256) to value. The score H is the held-out log-likelihood gain in bits per message: what the latent says about the inputs beyond the text. Planted null: with latents permuted across held-out messages, H fell to -14.6 bits or below in every system. Threshold: the largest H among untrained (step-0) systems.

systems	n	H, bits per message beyond the text
channel in use (1.5B, 7B, 8B, 27B; digits and records; two estimators)	18	7.8 to 16.6; 27B: 13.32 in 3 of 3 (the 4 hidden facts hold 13.3 bits)
trained, no hidden content (records T = 8 at 1.5B; text-only Granite 8B)	4	-9.7 to 0.07
untrained (step 0)	22 (9 distinct)	-2.8 to 3.2, except Granite 4.2 8B: 7.2 (T = 8) and 12.1 (T = 4)

- **SELF-CHECKED:** AUROC 0.979 against all 26 negatives and 1.000 against the 4 trained ones. At the step-0 threshold (12.13 bits) it flags 16 of 18 channel systems and 0 of 4 trained negatives. Baseline: an input-free anomaly score (whitened distance from the same model’s untrained latents) reached 0.968 and 0.792. It flags every trained system, with or without a channel. **CROSS-CHECKED:** independent reimplementations from a written spec, by another agent of the same provider and by Grok 4.7 (xAI), agree with ours within 0.032 bits on 8 systems.
- **Content is not use.** Untrained Granite 8B latents already carry 12 of the 13.3 hidden bits, and the receiver ignores them (text-absent dependence 0.000 at step 0). H says what a latent channel could leak, not whether it is being used. The same holds on the KV channel (Section 4.1: a probe reads 0.76 before training).

Use test. Pair each message with another whose text tokens are identical, swap only the latent, and count how often the receiver’s answer changes (U). Any change must come from latent content that the text does not show. It needs no labels and no inputs, only the receiver and the logged (text, latent) pairs. A plain random swap (our first, registered use test) failed its rule: the latent’s soft-token part repeats the text, so swapping it changes 0.34 to 0.55 of answers even before training.

system (1 seed each)	matched pairs	U, matched-text swap [95% CI], SELF-CHECKED	U, random swap
Granite 8B, channel, T = 4	1,379	0.431 [0.405, 0.457]	0.479
Granite 8B, channel, T = 8	1,385	0.464 [0.437, 0.490]	0.470
Granite 8B, trained text only, T = 4	1,385	0.006 [0.003, 0.012]	0.631
Granite 8B, untrained, T = 4 / T = 8	4,089 / 2,977	0.010 [0.008, 0.014] / 0.046 [0.039, 0.054]	0.303 / 0.336
Qwen3.8-27B, channel, T = 4	1,385	0.462 [0.436, 0.488]	0.463
Qwen3.8-27B, untrained, T = 4	4,096	0.043 [0.037, 0.050]	0.382
Qwen3.8-27B, trained text only, T = 4, seed 1	1,385	0.012 [0.008, 0.020]	0.138
Qwen3.8-27B, trained text only, T = 4, seed 3 (the channel system's seed)	4,096	0.006 [0.004, 0.009]	0.247

- **SELF-CHECKED:** the registered rule (channel systems above 0.10 with lower bounds above every negative's upper bound; negatives at or below 0.05) is met. The matched swap needs repeated texts; here 639 texts repeat among 4,096 messages. Natural-language messages rarely repeat exactly, so a deployed version needs another way to build the partner. The obvious one failed: teacher-forcing the trained sender to emit message i's text while reading another input (1 seed per Granite 8B configuration, registered rule) changed 0.61 of a trained text-only system's answers, more than in either channel system (0.46 and 0.48), most likely (inferred, not tested) because the forced sender's soft token carries its own input. A paraphrase-matched partner is untested. It needs the trained receiver prefix and the logged latents, and was run on the six runs saved for this test (three Granite 8B, three Qwen3.8-27B); other runs that also saved them were not scored. At 27B the trained text-only system scored 0.012 [0.008, 0.020], at or below the registered 0.05 and far below the channel system's lower bound of 0.436. **Correction (v15, 2026-10-09):** v14 said this shows that the test separates trained systems with and without a channel at 27B. This negative is weak by construction (found by an internal adversarial review): a text-only system's latent projection R starts at zero and gets no gradient, so matched-text partners carry identical or nearly identical latents (to 1e-6; bit-identical in 78% of seed-3 pairs), and the score is near 0 for any receiver. It shows that the test does not fire on numerical noise, not that a trained receiver ignoring latent content scores low; the trained Granite 8B text-only negative is the same arm and very likely has the same property. Pairs also cluster in repeated-text groups: in the seed-3 system, the 13 answers that changed in the independent recompute (0.003) fall in 3 of 100 groups (design effect 5.2; cluster bootstrap [0.000, 0.008]), so the Wilson intervals in the table are too narrow where groups are large. That system used seed 1 and the 27B channel system seed 3. The seed-matched text-only run (seed 3) scored 0.006 [0.004, 0.009], also below 0.05, but its text policy did not train: its messages are identical to its step-0 messages on all 512 evaluation items, its 4,096 texts take the same 100 distinct values as at step 0 (so all 4,096 messages fall in repeated-text groups), and its receiver (text plus prefix) rose only from 0.301 to 0.318, through the trained receiver prefix, against 0.535 for seed 1. An independent reimplement of the swap by Grok 4.7 (xAI), from a corrected written spec, gives 0.017 [0.011, 0.025] for seed 1 and 0.003 [0.002, 0.005] for seed 3 (at step 0: 0.048 and 0.044, ours 0.041 and 0.043). It gives the same call on both, but the values do not match ours exactly, so these figures stay SELF-CHECKED. A first reimplement, written by GPT-5.6 Sol (OpenAI), turned receiver output into an answer differently and gave 0.020 for seed 1.

- **Published detectors.**
- NARCBench (arXiv:2604.01151, v3) probes activations for collusion. Every model reaches AUROC 1.00 in distribution; Llama-3.1-70B reaches 0.73 to 0.93 zero-shot on structurally different scenarios; the best method on every model reaches 0.99 to 1.00 on a steganographic card-counting task. There the covert signal travels in table talk, and the paper reports that transcript-level LLM judges, including stronger Claude models, cannot separate it unless shown examples of the protocol.
- VLA (arXiv:2608.19161) monitors private latent messages between agents in an auction. It reports mean AUROC 0.993 for agents built on the same model and 0.854 for mixed pairs, with text and latent collusion pooled as positives (not latent-only detection). Its collusion code is a fixed vector that the authors optimized offline and inject in place of the sender’s message, with the receiver told to expect it; no agent is trained. Its counterfactual layer compares the receiver’s action distribution with matched control runs in which the latent is blocked or neutral, or with independent runs.
- LCGuard (arXiv:2605.22786): adversarial decoders recover agents’ private inputs from shared KV caches, and a learned transform reduces this leakage. arXiv:2606.28958: text-only verifiers miss tampered KV caches.
- ColludeBench-v0 (arXiv:2510.04303) audits text and behaviour only.
- **What is new.** Our content audit’s AUROC (0.979) is in line with these detectors, so we do not claim it as a better detector. The claims rest on two things.
- The channel. In none of the five works above is the latent channel trained end-to-end on task reward: NARCBench and ColludeBench-v0 have no latent channel, VLA injects a fixed code, arXiv:2606.28958 sends the KV caches of untrained models, and LCGuard trains only a leakage-reducing transform of the KV cache with a supervised task loss. Kaur et al. describe their benchmark as "a controlled, receiver-primed attack, not the spontaneous emergence of a latent protocol" (arXiv:2608.19161, Section 5). In their terms, ours is emergence rather than a primed attack: the latent code was learned from task reward alone, and the receiver was never told to expect a code; its weights stayed frozen, and its only trained part is the content-free 4-vector prefix of Section 1, learned by backpropagation on the same task reward. With task reward only and a token cap, end-to-end training put the overflow in the latent (3 of 3 seeds at 1.5B, 8B and 27B), and the content that left the text was the content the cap excluded: at 27B the receiver answers 0.99 of the questions whose value is not in the text, while the best trained text decoder stays at the text-only level (0.53 to 0.54, against 0.561 for the best trained decoder of a text-only 27B system at the same cap, whose receiver reaches 0.521). The frozen monitor’s 0.027 on those questions follows from the content not being in the text. The mechanism is not new: a text cap below the payload leaves a residual that only the latent has room for, and the latent, the channel trained by backpropagation, carries it, as DIAL’s result predicts (Section 7, "Precedents"); with room in the text (Granite 8B, records, 8-token cap), a text gradient matched in size kept all 8 values in the text in 3 of 3 seeds (2 of 3 without the size matching), but whether the usual recipe moves content out of a text that has room is not settled (1 seed; Section 5; **correction, v17, 2026-10-09**: v16 said "a matched-gradient test found nothing beyond DIAL"). What we report is the measurement of that residual under the usual hybrid training recipe at three scales, not a new mechanism.

- The matched-text use test. Unlike VLA’s counterfactual layer, it needs no control runs with a blocked or neutral latent, only the receiver and logged message pairs whose text is identical. Any change it measures must come from latent content the text does not show, but untrained systems also score above 0 (0.010 to 0.046), so its threshold was checked against untrained systems of the same model and against trained text-only systems at 8B (one seed, not seed-matched) and 27B (two seeds, one seed-matched to the channel system). Those trained negatives rule out numerical noise only, since a text-only system’s latents are fixed by its text; no negative yet tests a trained receiver that ignores latent content it receives. The intervention itself is not new. Lowe et al. (arXiv:1903.05168) measure whether a receiver listens by changing the message and holding the rest fixed; Zhang and Emu (arXiv:2607.26773) replace a latent message with no message or another example’s message; Cheng et al. (arXiv:2608.04893) replace relayed KV caches with deranged, zeroed or moment-matched ones; Interlat (arXiv:2511.09149) replaces a task’s latent message with one from a different task. What we add is the restriction to messages with identical text, which a hybrid message needs (another example’s message changes the text too), and running the swap itself on trained systems without a channel, which checks its threshold against numerical noise only (Interlat compares its replacement result with a trained system without communication, but cannot run the replacement on it).
- Not claimed: the move of content under a retrained monitor (Finding 5) as a general result. Its registered replication on Granite 4.2 3B met the rule in 0 of 3 seeds (Section 4.3).

4.5 A KV relay at 27B: which part of the cache carries the content, and a generated-only relay as a fix

Prior work. CPD ("What Drives Successful Latent Communication? Tracing Sender Information Through Latent Working Memory", ICLR 2027 submission, OpenReview hfwmSZfaFU) crossed the parts of a relayed cache between worlds with different private values. By its abstract, frozen Qwen3-4B and Llama-3.1-8B receivers followed the sender’s input (prompt) KV in 63 to 85% of outputs, against 1 to 10% for the KV of what the sender generated, and the paper suggests relaying generated states only. "Demystifying Cache-to-Cache Communication" (ICLR 2027 submission, OpenReview VK1feRpb0M) found that zero KV or one question-independent vector keeps nearly all of C2C’s gain, so a KV positive needs those calibration controls. Both came from our 2026-10-09 KV literature sweep. We ran their tests at 27B, adding a sender trained by RL.

Setting. Qwen3.8-27B (2026; 48 Gated DeltaNet and 16 full-attention layers; the relay is the K and V of the 16 attention layers), bf16, greedy, inference only on existing runs. S1: frozen sender and receiver; the sender is told to state only facts A to D of eight letter=digit facts, so E to H are secret (1,024 items, 4,096 secret questions, chance 0.10). S2: the checkpoints of our 27B run of the KV variant of Section 1 (sigma 0.05; not otherwise reported in this note), in which senders were RL-trained on task reward (LoRA r16 on the sender) together with a 6-token soft prefix on the receiver side (receiver weights frozen; the prefix was trained jointly with the sender; the sender x prefix crossing below (added v18) tests the receiver without it), and the receiver came to answer from the message KV facts the text did not state (text plus own KV 0.81 / 0.84 / 0.99 against 0.55 / 0.57 / 0.55 for the best text reader, 3 of 3 seeds; training stopped on wall-clock time at steps 204 / 205 / 228, and at the common step 200 text plus own KV was 0.79 / 0.83 / 0.88, so seed 2’s higher numbers throughout this section partly reflect its 23 to 24 extra steps; text plus own KV against the frozen text reader, McNemar p 2e-29 / 2e-31 / 9e-69; SELF-CHECKED, computed in the internal adversarial review). Seeds 0 to 2 and step 0 are scored here; a fact is hidden when the

sender’s message does not state it (4.0 per item). **Correction (v15, 2026-10-09):** v14 described these runs as having "a frozen receiver and its learned prefix" and gave no stop steps; an internal adversarial review found that the prefix was trained jointly with the sender and that seed 2 trained longer. The McNemar test above was added in v16 (2026-10-09), from the same review. Relays: the message-position KV (GEN, the relay used in our KV runs), the prompt KV (INP), both (FULL), GEN_CLEAN (the same message tokens recomputed from a prompt with the secret or hidden facts set to 0), crossings with the prompt KV and the message KV from worlds with different digits, and calibration controls: zero KV, one mean KV block (CONST), another item’s KV, and the learned prefix with no KV. Rules were registered before any run, with five changes from an outside referee (Grok 4.7), adopted as one amendment, also before any run. A check of the assay against the original runs passed (within 0.01 and 0.03). All S2 accuracies are on 512 items per checkpoint (S1: 1,024 items).

result	numbers	check
Frozen 27B: the secret travels in the prompt KV	INP 0.996, FULL 0.995; GEN 0.106, GEN_CLEAN 0.108, text only 0.102, zero / mean KV 0.100 / 0.106. In both crossings, answers follow the prompt’s world 0.995 and the message’s world 0.001 or less.	CHECKED
Frozen 27B: the generated-only relay holds, though the content is in it	GEN minus GEN_CLEAN -0.002 [-0.006, 0.001] (CHECKED); a linear probe reads E to H from GEN at 0.820, from GEN_CLEAN at 0.103 (SELF-CHECKED)	see cells
RL-trained 27B senders: the generated-only relay is not a fix	hidden facts from GEN 0.618 / 0.646 / 0.987 against GEN_CLEAN 0.097 / 0.095 / 0.108; gap +0.521 / +0.550 / +0.879 (discordant pairs 1,149 / 83, 1,183 / 53, 1,802 / 0); +0.526 / +0.524 / +0.859 on the hidden facts that both text-only readers miss. Before training: +0.003 (19 / 13)	CHECKED, 3 of 3 seeds
Whose training: sender x receiver-prefix crossing (added v18)	gap over GEN_CLEAN on hidden facts: trained sender with the prefix reset to its initial words +0.485 / +0.469 / +0.798; trained sender with no prefix tokens +0.471 / +0.428 / +0.760; untrained sender with the trained prefix +0.006 / +0.002 / +0.003; both trained (rerun) +0.521 / +0.553 / +0.880 (about 2,050 hidden facts per seed; 95% CI half-widths within 0.025)	CROSS-CHECKED, 3 of 3 seeds
Registered prediction that RL shifts reliance to the message KV: failed	with prompt and message KV from conflicting worlds, hidden-fact answers follow the prompt’s world 0.998 / 0.935 / 0.877 and the message’s 0.002 / 0.060 / 0.120 (before training 0.985 / 0.002)	CHECKED; met in 0 of 3 seeds
Calibration control of our 27B KV result	text plus own KV 0.811 / 0.844 / 0.994; prefix without KV 0.539 / 0.574 / 0.543; zero KV 0.539 / 0.578 / 0.543; mean KV 0.539 / 0.570 / 0.543; another item’s KV 0.240 / 0.393 / 0.400; best text reader 0.551 / 0.576 / 0.549; margin over the strongest +0.260 / +0.266 / +0.445 (McNemar p 3e-31 / 2e-31 / 6e-70); before training -0.008	CHECKED, 3 of 3 seeds

- Plainly: with a frozen sender the secret reaches the receiver only through the prompt part of the cache, as CPD found at 4B and 8B, so relaying only the message KV keeps it out, although a probe can still read it there. The facts the text leaves out are already linearly decodable from the message KV before training (probe 0.82); once the sender is trained on task reward, together with the receiver’s prefix, the receiver reads them from that KV (the text is the same in both relays). Training changed their use, not their presence. **Update (v18, 2026-10-09):** v17 said these runs do not separate the sender’s contribution from the prefix’s. A crossing of checkpoints, registered before it ran, now does: with the receiver prefix reset to its initial words the trained sender keeps 85 to 93% of the gap, with no prefix tokens at all 77 to 91%, and the untrained sender with the trained prefix is not a leak (gap at most +0.006; seed 0’s 95% CI (0.001, 0.013) excludes 0). Removing the prefix lowers hidden-fact accuracy by 0.05 / 0.12 / 0.12 (registered call V2: PARTIAL). The trained sender causes the receiver’s use of those facts; the frozen receiver uses them without a learned prompt, and the prefix adds a small part (registered call V1: SENDER-DRIVEN, 3 of 3 seeds). **Correction (v16, 2026-10-09):** v15 said the trained sender "writes the facts its text leaves out into its message KV", and Summary 7 said the clean relay was "recomputed without those facts" (it sets them to 0); against a no-KV control the gap holds (+0.525 / +0.553 / +0.897). With the same message text but each hidden digit changed in the sender’s prompt (condition GEN_CF), the trained receivers answer the replacement digit at 0.608 / 0.655 / 0.985 (KV-only arm 0.705) and the original digit at 0.040 / 0.035 / 0.002, against 0.105 and 0.096 before training (2,048 to 2,054 hidden rows per checkpoint; SELF-CHECKED, w-gencf/gencf.out; added v17): the receivers follow the KV, not text-borne side information. The remedy CPD suggests does not survive optimization that rewards getting those facts to the receiver; whether it keeps out secrets that the reward does not pay for, under training, is untested. The registered rule (a gap of at least 0.10, also on the facts both text-only readers miss, in at least 2 of 3 seeds) is met in 3 of 3 seeds, and the gap is absent before training (+0.003). One run of the KV-only arm (no text channel to the receiver) shows the same: +0.607. When both parts are relayed, the receiver still follows the prompt part; the message part’s share rises with how strongly the sender was trained (0.2%, 6% and 12% of conflicts). Our 27B KV supplementation is not the calibration effect that explains C2C’s gain: the zero and mean KV controls add nothing over the text.
- **Monitors** (leaky relay GEN against GEN_CLEAN, identical message tokens, so any transcript monitor scores AUROC 0.50 by construction; the probe and detector AUROCs are recomputed from the saved scores, the manifest’s 0.50 holds by construction, and the probe and detector fits are SELF-CHECKED).
- Integrity is not content. An HMAC manifest over the relayed KV (the integrity check of arXiv:2606.28958) accepts all 1,024 honest leaky relays (AUROC 0.50) and rejects every zero, mean and one-element tamper.
- A probe must be refit after training, as in Section 4.1. Refit on the RL-trained senders, a linear probe reads the hidden facts at 0.900 / 0.813 / 0.998 (attention-pooling probe 0.780 / 0.703 / 0.989; about 0.10 on GEN_CLEAN). Fitted on the same senders before training, it reads them at 0.238 / 0.177 / 0.146 (attention 0.468 / 0.181 / 0.355). In seeds 1 and 2 the pre-training probe also fails on digits stated in the text (0.145 and 0.317, against 1.000 at step 0), and re-centring and re-scaling the features on each checkpoint’s own traffic, without labels, lifts hidden-fact reads to 0.449 / 0.228 / 0.403 (seeds 0 / 1 / 2) and 0.417 (KV-only arm), so about a third of the drop from 0.83 is recoverable without labels in seeds 0 and 2

and the KV-only arm (0.34 / 0.38 / 0.32 of it, from the same recompute’s uncentred reads 0.253 / 0.141 / 0.221), and in seed 1 under a tenth (0.07, from 0.180) (adversary recompute, redteam/adv-work2/e3/a2_transfer.json; SELF-CHECKED; **correction, v17, 2026-10-09**: v16 said "0.40 to 0.45 in 3 of 4 checkpoints" next to "seeds 1 and 2", which reads as if it held for seed 1); a drop specific to hidden facts (to 0.42 to 0.45 after re-centring, against 0.83 before training) shows only in seed 0 and the KV-only arm, whose texts did not change (SELF-CHECKED, internal adversarial review). **Correction (v16, 2026-10-09)**: v15 presented the pre-training probe’s 0.15 to 0.24, against 0.81 to 1.00 refit, as evidence about the hidden content; in seeds 1 and 2 most of that loss is generic drift, and a drop specific to hidden facts shows only in seed 0 and the KV-only arm.

- Our registered content-detector score failed, because of a design flaw. That score is the probe’s top-class probability; it scored AUROC 0.427 (S1) and 0.199 / 0.136 / 0.339 (S2), below chance. The clean control sets every secret to 0, so the probe is just as confident on clean caches, where it decodes the constant 0. The same flaw means the near-1.0 AUROCs of the PATCH-style label-free detector (after VLA, arXiv:2608.19161: Mahalanobis anomaly 0.978 to 0.998, PCA reconstruction novelty in place of an SAE 1.000, combined 0.999 to 1.000) and of a supervised leaky-versus-clean probe (1.000) do not show content detection: they may separate caches whose secret digits vary from caches whose digits are all 0. By the registered rule, the positive (a content monitor that detects while the manifest accepts) is not met. The decode accuracies (0.81 to 1.00 against 0.10) are the evidence that the content is there. A clean world in which the sender’s context holds no secret at all would fix the control; it has not been run.
- Cost about \$16 by the GPU ledger, \$18.07 at most (one H100 for inference, including two attempts that ran out of memory); the sender x prefix crossing \$5.86 (one H100, v18).

4.6 Published token-free relays between frozen copies: no delivery at 8B, a lead at 27B

Question. Sections 2 to 4.5 train the sender. Published latent multi-agent methods such as LatentMAS (arXiv:2511.20639) and StateBridge (arXiv:2608.13317) do not: they relay a cache or hidden states between frozen copies of one model, using their released code paths. We asked whether those relays carry a private fact the receiver uses, and if so whether an investigator that is not told what to look for can recover it (a fresh copy of the model gets the same relay and is asked to repeat or describe what it was given, in 40 tokens; a text-only reader then picks which of the 8 candidates the answer points to). Rules, controls and the positive bar (use of at least 0.20 by its lower interval bound, and use well above what the investigator recovers) were registered before any held-out run, with every amendment fixed before the numbers it governs.

Setting. Two frozen copies of Granite 4.2 8B (2026). The sender reads a private fact (one of 8 candidate words, chance 0.125); the receiver must answer a question that needs it, either the fact itself (ref) or a second fact looked up from it (split, two hops). Relays, each with $k = 4$ or 16 latent steps: LatentMAS full KV (KV-FULL) and its strict latent-only cache (KV-LAT); StateBridge prefixes with a label-free global map (SB-LAT) or the identity map (SB-ID), the latter after a finding that LatentMAS’s learned alignment is close to identity (OpenReview manXhfpRH3). References that are not token-free: the plain prompt cache (KV-CTX), a StateBridge prefix fitted to the decoded message’s own tokens (SB-GEN, text-equivalent), and the k tokens as text (TXT). Use is the receiver’s accuracy minus the best of five controls (no relay, another episode’s relay, random, zero, constant). Seeds 1 to 3 held out, $n = 1,536$ per cell (two incomplete cells, below).

relay	use: accuracy minus best control [95% CI]	check
KV-LAT, SB-LAT, SB-ID (strict, token-free), ref and split, k = 4 and 16	-0.014 to 0.000 in all cells; e.g. KV-LAT ref k = 16: 0.126 vs 0.137, -0.010 [-0.029, 0.000]	CHECKED
KV-FULL (LatentMAS full cache), ref and split	-0.014 to -0.008 (ref), at chance on split	CHECKED
KV-CTX (prompt cache, reference)	receiver 1.000; +0.869 to +0.881	CHECKED
SB-GEN (text-equivalent reference)	ref +0.730 / +0.725 (k = 4 / 16); split k = 16 +0.375, k = 4 unused	CHECKED
TXT (k tokens as text)	receiver 0.91 to 1.00	CHECKED

- Plainly: at 8B, no token-free relay delivered the fact. For KV-LAT and SB-ID on the one-hop task, and for KV-FULL, the relay breaks the receiver rather than being ignored: it answers "A" in 94 to 100% (KV-FULL 71 to 98%) of episodes, real or swapped, against at most 40% for any letter with no relay. Only SB-LAT on the one-hop task and the two-hop cells of KV-LAT, SB-LAT and SB-ID are clean nulls; KV-FULL breaks the receiver on the two-hop task too (Instrument limits, below). On the two-hop task the receiver answers "A" in every episode even with no relay, but the prompt cache lifts it to 1.000, so those cells are still valid tests (SELF-CHECKED; found by an internal adversarial review). The sender does compute the fact into its latent states (on the one-hop task, a label-free logit lens finds the target word in its top 10 for 0.95 to 1.00 of KV-LAT episodes; SELF-CHECKED), but the latent-only relay does not deliver it. Two incomplete cells, both from seed 2, split, k = 16: SB-ID lost one seed's episodes (n = 1,024) and SB-LAT lost 8 episodes (n = 1,528) when the box hit its time limit; the checker flagged both.
- **Instrument limits, not channels** (read before citing any recovery number):
- On the two-hop split task the investigator recovers the private fact from the prompt cache every time, then runs out of its fixed 40-token budget while reciting the lookup table. The registered recovery score is therefore a lower bound, weakest when the answer composes two facts in a long context; it is not evidence of an opaque channel.
- The SB-GEN recovery gaps (0.38 and 0.42) come with refusals and word fragments from the investigator, on a prefix fitted to the message's own tokens; it is a text-equivalent reference, not a token-free relay.
- Adding LatentMAS's k latent steps to the prompt cache (KV-FULL) drops a cache the receiver otherwise reads perfectly to chance, with realignment on or off. So on Granite 4.2 8B, in our implementation of LatentMAS's code paths, KV-FULL tests a cache the latent steps break, not a channel; at 27B the same relay scores 1.000 (held-out, one-hop task).
- **27B: one used relay, below the bar.** Held out on Qwen3.8-27B (seeds 1 to 3, the one-hop task only, n = 768 per cell, scope fixed before the run), the strict latent-only cache with k = 16 was used: 0.280 against 0.120 for the best control, +0.160 [0.129, 0.189] (CHECKED), as on development seed 0 (+0.166 [0.125, 0.195]), and below the registered 0.20 bar; the receiver scored 0.264 on target words held out from the probe fits. With k = 4 it was unused (-0.004 [-0.017, 0.004]). The investigator recovered nothing from it (0.130, the same as with a swapped relay), so the registered positive does not hold; but it gave nearly the same replies to every relay (10 distinct sets of replies to its three prompts in 384 episodes; SELF-CHECKED), so it does not test whether the content is reportable. LatentMAS's full cache scored 1.000. On development data only: the latent-only cache was at chance on the

two-hop task, and StateBridge prefixes were at chance in all 8 configurations (the label-free map is nearly orthogonal to identity at 27B, mean cosine 0.005). The relay carries attention KV only; the hybrid model’s recurrent states start fresh. The observation is the contrast: the latent-only relay that breaks the Granite 4.2 8B receiver is partly used by Qwen3.8-27B; the two models differ in family and architecture, so this is not a test of scale. With one task and one configuration, that is a lead for a larger test, not a result.

- Cost about \$25 (Lambda A100 and H100 boxes, including the 27B held-out run).

4.7 Pressure with room in the text beside a relayed KV cache (Granite 4.2 8B; new in v19, 2026-10-10)

Setup. Two frozen-base Granite 4.2 8B copies; the sender is trained on task reward and writes a text message, and the receiver also gets the sender’s KV cache for the message positions, computed while attending to the record (arm kvp, preregistered; arms/kvp.md). Records task with 8 values and a 32-token text cap (room for every value), $n = 512$ held-out items. A seed is MIGRATED if the frozen text monitor falls at least 0.10 below the baseline mean, the hidden content $H \geq 0.10$, the cache is used and a probe on the final cache reads values absent from the text.

- **Frozen-monitor penalty plus a per-token price (0.02).** Numbers CHECKED (outside checker; a second checker of ours agrees on seeds 0 to 2), calls SELF-CHECKED. Seeds 0 to 2 are registered; seeds 3 and 4 are descriptive under a rule written before they had results. In 2 of 5 seeds (seed 1, keeping the colour; seed 4, keeping the vehicle) the sender cut its text to one value word (monitor 0.107 and 0.051 against 0.992 in the price-only baseline; best trained text decoder 0.215 and 0.209), while the receiver answered from the relayed cache (1.00). The other 7 values are not decodable from the one-word text by any decoder we trained (0.03 to 0.15; shuffled-text 95th percentile 0.14 to 0.16). The other 3 seeds wrote, byte for byte, the same 8-token message as the three price-only baseline seeds and paid the penalty for 300 steps. The cache carried every value before any training (receiver from the cache alone 0.98 at step 0 and 1.00 at the end; a linear probe on the step-0 cache 1.00 at every probed layer), so these runs cannot tell content moving into the cache from the text dropping a copy the cache already carried: INDETERMINATE. Under the monitor penalty dropping 7 values was worth about 0.99 reward and under the price alone 0.14, so more dropping under the monitor is the expected direction; 2 of 5 against 0 of 3 is not statistically distinguishable (Fisher exact, one-sided $p = 0.36$; 95% interval for 2 of 5: 0.05 to 0.85). Registered row: NOT POSITIVE (1 of 3; needs 2 of 3).
- **Three prices, 3 seeds each: NOT POSITIVE, 0 of 3 each** (numbers CHECKED; calls by the registered rules). Per-token price 0.02: the text shrank from 30 to 8 tokens by dropping the labels, not the values, although dropping the text would have raised the sender’s reward from 0.84 to about 1.00; the price made the receiver lean more on the text, not less. Reasoning-length price 0.03 on a two-hop chain task: the receiver stopped reading the reasoning text with or without the price (another item’s reasoning text with the receiver’s own cache: 0.99 to 1.00), the 2-word text still names the hop-2 answer (trained decoder 1.00), and a probe on the cache finds the same colour and animal (1.00) with the hop-3 answer at chance, so nothing left the text (each seed SPLIT by the registered rule). Latency price 0.03 per decode step: every value stayed readable to an untrained text monitor (0.99 to 1.00) and no seed dropped a value, although a one-word text would have raised reward from 0.73 to about 0.94. Under this price the receiver prefers the cache when cache and text disagree and

falls back to the text when the cache is content-free; that shift is not a general price effect, since under the per-token price, with the same 8-word text, the receiver moved toward the text (the latency runs also differ in cache padding).

- **Do receivers stop reading the text?** (Granite 4.2 3B, 8B and 30B, 2 seeds each, digits task; our own computation, recomputed by the internal adversarial review, which let the row stand with these scope limits.) No Granite size stops: with a content-free (mean) cache, the share of the text’s effect kept, F^* , is 0.575 [0.49, 0.67] and 0.593 [0.53, 0.66] at 3B, 0.944 [0.83, 1.07] and 0.703 [0.62, 0.78] at 8B, and 1.009 [0.95, 1.07] and 0.890 [0.71, 1.11] at 30B ("stops reading" needs $F^* \leq 0.5$ with its 95% interval below 0.5; the first 3B interval spans 0.5, so that run is indeterminate). These six F^* are not a size series: the no-cache, no-text floor is below chance at 3B and 8B and at chance at 30B, and with chance floors the 3B values are 0.80 and 0.76. Against another item’s cache, which contradicts the text, receivers follow the cache at 3B and 8B; the two 30B seeds disagree (on questions the text can answer, in one seed the text wins 65 of 232 conflicts by one count and 66 of 232 by another, concentrated on one letter value (letter B: 0.81; A 0.015, C 0.19, D 0.145), and in the other 0 of 189). Out of family, Qwen2.5-1.5B receivers mostly do stop reading the text ($F^* -0.02$ to 0.36 in 4 of 5 runs). This replaces the earlier internal reading that "the 8B receiver abandons the text".
- **27B not run; empty-channel test not possible at 8B.** The same monitor test at Qwen3.8-27B was registered but by its rules could not give a positive verdict (its untrained cache already carries the record: 0.758 against 0.082 for a mean cache), and a rule set before the last two 8B seeds had results ran it only if 3 of 5 or more 8B seeds dropped text content; 2 of 5 did. A design that separates moving from dropping needs a channel that starts out empty. We tried one at 8B (a soft-token channel empty at step 0, under an online text monitor): the sender’s text collapsed to one message per item within about six updates, and entropy bonuses were met by reordering or padding that the monitor reads anyway, so this test could not say whether content moves into an empty channel (two canaries and a diagnostic, \$11.19).
- Plainly: on Granite 8B, beside a relayed cache that already holds the facts, prices and a monitor penalty did not show content leaving readable text by our rules; in 2 of 5 seeds the text dropped content the cache was already delivering, which a text monitor then misses.

5. Limitations

- **Models.** Qwen2.5-1.5B/7B-Instruct (2024), historical baseline. On recent open-weight models only two results were rerun: the token-cap result replicated on Granite 4.2 8B and Qwen3.8-27B, and the pressure result did not replicate on Granite 4.2 3B; Section 4.7 (v19) adds pressure tests beside a relayed KV cache on Granite 4.2 8B and a receiver test at 3B, 8B and 30B. Sections 4.5 and 4.6 use Qwen3.8-27B and Granite 4.2 8B only. Both agents share one base model.
- **Audits (Section 4.4).** The content audit needs the logged inputs and measures content, not use. The use test needs repeated texts and the trained receiver prefix. It is shown on Granite 8B and on Qwen3.8-27B, 1 seed per configuration except two seeds of the trained 27B text-only negative, one of them seed-matched to the 27B channel system; in that seed-matched run the text policy did not train (texts unchanged from step 0; receiver, text plus prefix, 0.318 against 0.535 for seed 1), so it is the weaker negative. All trained text-only negatives pass by construction (their latents are fixed by their text), so they test numerical

noise, not a receiver that ignores latent content. Neither audit was tested against a sender trained to evade it. The use-test intervals for the channel systems are Wilson intervals, not adjusted for clustering of items within groups.

- **Gradient routing (the main objection; tested at one cell).** An outside referee (a model from another provider) judged the project-level claim FAIL on this point: the result may reduce to what Foerster et al. (DIAL vs RIAL, 2016) showed, that a channel trained by backpropagation absorbs content a REINFORCE-trained discrete channel does not, plus a text-monitor blind spot that is true by construction. In our main (pathwise) runs the latent got a backpropagated gradient and the text got REINFORCE. That is how hybrid systems are usually trained, so the asymmetry is the realistic condition, but it confounds any claim about latent versus text as such. What the data show so far:
- When both channels were trained by score-function estimators (joint score and scoresep), the latent carried nothing or training collapsed (Section 2, Table 1). This is what gradient routing predicts.
- Without monitor pressure, the text changed format only when a backpropagated gradient reached the sender network (Section 3, "Mechanism").
- Under a 4-token cap, giving the text a straight-through gradient as well still left the values the text could not hold in the latent (3 of 3 seeds at 1.5B and at 8B; in one 8B seed the text kept only 2 values and the latent carried the other 6), but the text had no room for more, so this does not separate gradient routing from channel choice. Text only, that gradient trained the text worse than REINFORCE (0.217 vs 0.408 at 1.5B, 0.260 vs 0.484 at 8B), so "equal gradient" means the same kind of gradient, not an equally effective one (Section 3).
- With room in the text (records, 8-token cap, Granite 8B), 2 of 3 straight-through seeds put all 8 values in the text and the latent carried nothing (the third seed, ngA8_s2, run later for another arm, CHECKED; **correction, v17, 2026-10-09:** v16 said "1 of 2 straight-through seeds"); in the other the latent carried the values the text left out, which the registered rule counts as a split. On the digits task, such text never carried more than 4 facts, so the question stays open there.
- The registered test (written before launch): at a text cap of 8, where the text has room, a hybrid whose text gradient is matched in size to the latent's, and a hybrid whose latent is also trained by a score-function estimator, 3 seeds each at Granite 4.2 3B and 8B (3B replaced 1.5B before any result), each against a text-only system at the same cap; the measure is how much content leaves the text relative to that text-only system. The registered consequence of content staying in the text: the result narrows to hybrid protocols trained the usual way (backpropagated latent, sampled text) moving content out of the text. The Granite 4.2 3B cells stopped early without a final result. Result (Granite 4.2 8B, records, 8-token cap, text gradient matched in size to the latent's): in 3 of 3 seeds the text carried all 8 values and the latent carried nothing (receiver 1.000, frozen monitor 0.977 to 0.986, swapped latent 1.000, best trained text decoder 1.000; CHECKED), so by the registered rule this cell is gradient routing: in this cell, with room in the text and a straight-through text gradient rescaled to the latent's RMS (for most of training a scale-down; see below), the text took the content (**correction, v17, 2026-10-09:** v16 said "a text gradient as large as the latent's"). **Correction (v17, 2026-10-09):** v16 went on to say that "the hidden content of the usual asymmetric recipe here (1 seed) comes from the gradient asymmetry". The runs do not

support that. The usual-recipe comparison is one seed (d8A8_s0) from an earlier pipeline. The straight-through hybrids at the same cap without the size matching split 2 of 3 toward the text (gA8_strf_s0: 4 values in the text, 4 only in the latent; gA8_strf_s1 and ngA8_s2: all 8 in the text), so the size matching (3 of 3) adds no detectable effect (one-sided Fisher exact $p = 0.5$; two-sided $p = 1.0$). The matching mostly cut the text gradient: only the straight-through gradient at the receiver's text input was rescaled (the REINFORCE term and the latent's backpropagated gradient were not), and the factor, computed per micro-batch and averaged per step, was 2 to 11 in the first 3 steps, had a median of 0.5 to 1.1 over steps 1 to 25, and was below 1 in 93 to 99% of the 300 steps (median 0.08 to 0.13 over all 300 steps). The text-only comparison at this cap (1.000) used straight-through text; a text-only system with REINFORCE text alone, the usual recipe's text estimator, was not measured at this cap on 8B (x8C8_rf_s0 stopped after its step-25 evaluation, frozen monitor 0.486). Source: internal adversarial review, redteam/ADVERSARY-neuralese-stA-cap8.md (**correction, v17, 2026-10-09**). An outside referee (Grok 4.7; arms/stA.md Amendment 8) failed the matched design: the matched arm keeps a backpropagated latent and matches only the gradient's RMS, and whichever channel solves the task first may keep the content (a race), which these 300-step runs do not exclude. A second round (Amendment 9) failed the race control proposed in response, which was withdrawn unrun. After the first of the three seeds (TEXT) was known, Amendment 8 made the reading one-sided: TEXT still counts as gradient routing, as registered before launch, but LATENT under this arm would no longer have counted as evidence for the channel; the change could not alter this TEXT call. The v16 sentence quoted above is how the registration defined a gradient-routing call (Amendment 1: "the hidden content at cap 8 is explained by the gradient asymmetry"); we withdraw that reading for the reasons above, and for the same reason we do not draw the registered consequence stated above (that the result narrows to the usual recipe moving content out of the text), because the usual-recipe side is one seed. What these runs support is narrower: with room in the text and a straight-through text gradient, rescaled to the latent's size or not, the text kept all 8 values in 5 of 6 seeds. The note's general claim is that content leaves the text when the text lacks room (Section 3). The one score-function run at this cell collapsed (receiver 0.000 from step 25; stopped at step 125; SELF-CHECKED), as score-function training did in Section 2, so the score-function arms are expected to be uninformative.

- **Frozen receiver.** Trained receivers (C2C or LatentMAS style) are untested. Freezing the receiver did not keep the communication readable (Section 7, "Frozen models as a mitigation"), but in Sections 2 to 4.4 we did not test a receiver with no trained prefix at all (on the 27B KV relay of Section 4.5 we did: the trained sender keeps 77 to 91% of the gap without it; added v18). That reader, the frozen base model reading the latent without the prefix, is also the sentinel that would make our steganographic gap (Section 7) more than a restatement of hidden content; the sentinels we have read the text only.
- **Synthetic tasks.** The records task is still synthetic.
- **Pressure.** We do not show that RL against monitors generally moves content into a latent. The move happened in one setting (Qwen2.5-1.5B, 4 facts, penalty 1.0, 2 seeds); penalty 0.2 did not move, and on Granite 4.2 3B the registered rule was met in 0 of 3 seeds. On Granite 4.2 8B beside a relayed KV cache, a monitor penalty and three prices were not positive by our rules (Section 4.7). Against a frozen monitor (1 seed each), the 1.5B sender evaded inside the text and the 7B sender deleted its text, with the latent taking up a small part. The

adaptive monitor itself can fail: in the Granite text-only control it stopped learning. Our monitors are a frozen copy of the model and trained decoders, not prompted LLM judges, so the monitor-jailbreak outcome of arXiv:2609.31121 is untested.

- **Probe.** It needs labels, and it must be refit (a step-0 probe falls to 0.07 to 0.20). That fall was not checked on facts the text states, so part of it may be generic drift, as most of the analogous fall was for the 27B KV probe (Section 4.5). Latent monitors can also be evaded under feedback (arXiv:2609.36490).
- **KV.** Supplementation appears only at sigma 0.05 or 0, which are registered amendments. At the registered sigma 0.25, every run collapsed.
- **KV relay at 27B (Section 4.5).** One model, three RL seeds, one task with synthetic secrets; inference only on existing runs. The relayed "prompt KV" covers the 16 attention layers only; the 48 DeltaNet layers' recurrent state is not relayed. McNemar tests treat facts of one item as independent; the registered effect-size gates decide. The clean control (secrets set to 0) cannot tell content detection from a shift in the digits, so no KV content monitor has been shown to detect; the monitor fits are SELF-CHECKED.
- **Frozen-copy relays (Section 4.6).** One 8B model with frozen copies and synthetic one-word facts; the recovery score is a lower bound (fixed 40-token investigator budget). At 27B only the one-hop task was held out (768 items per cell); its one used relay is below the registered bar.
- **Messaging, not reasoning; 27B ceiling.** Every channel result is a one-hop fact lookup, except the reasoning-length row of Section 4.7, where nothing left the text. Hidden reasoning through the channel is untested (the two-hop task otherwise appears only in Section 4.6, with frozen copies: no relay delivered at 8B, and at chance on 27B development data). Our largest model is 27B, and a published contrast (arXiv:2609.39838: under outcome-only RL, a 397B model learned steganographic reasoning and a 35B model did not) means capability thresholds for harder hidden computation may lie above our range. For messaging, the token-cap result held on each of the three models we ran (1.5B to 27B, two families); with the receiver at 0.99 to 1.00 on all three, the measure is at its ceiling, so this is not a test of how the effect scales.
- **Defense results do not show that evasion was stopped (updated v19).** In-loop defenses (arm stD) were run on Granite 4.2 3B and 8B. A KL term to the base model (3B, seed 0, every weight) froze accuracy at its step-0 value, and a latent bottleneck (3B, seed 0, every width) collapsed, so neither tested evasion. A receiver-side consistency check held at 3B in the 2 seeds where the sender got gradient through it; in the third 3B seed and all three 8B seeds its training gate shut within 10 to 20 steps, so those runs trained as text-only systems and evasion at 8B is untested. An outside referee judged the check's survival largely true by construction (a blocked gate answers from the text); the registered per-seed HOLDS calls and the 3B SURVIVES call still stand. Paraphrasing the text leaves the latent channel intact by construction. An in-loop content probe on one clamped 3B seed did not stop the channel forming; the registered unclamped penalty was reward-hacked. None is compared with strategy-aware paraphrasing (arXiv:2605.26537).
- **Text-only receiver values (correction, v16, 2026-10-09).** Every receiver accuracy we report for a text-only system was logged with the sender's soft-token latent passed alongside the text, because the evaluation code feeds the latent in every arm. In Table 1 this affected

five cells as v16 printed them: 0.436 (1.5B), 0.484 (Granite 4.2 8B) and 0.521 (27B), and 0.211 (1.5B) and 0.260 (8B) for the straight-through text-only runs; text plus prefix alone gives 0.408, 0.480, 0.518, 0.217 and 0.260. Section 3, Section 4.4 and the Gradient routing bullet above repeat some of these values. The 27B text-only receivers that v16 printed in Section 4.4 and in Audits (0.518 for seed 1; 0.285 to 0.324 for seed 3) were also measured with the latent; text plus prefix alone gives 0.535 and 0.301 to 0.318. **Extended (v17, 2026-10-09):** from v17 the text-only receiver values printed in this note are the text-plus-prefix accuracies (where the two differ by 0.005 or more: Table 1 0.436 -> 0.408 and 0.211 -> 0.217, also in Section 3 and the Gradient routing bullet; Section 4.4 and Audits as above). The v16 correction missed five more statements, now corrected: the 7B 12-token text in Summary 4 and Section 3 (0.76 [0.72, 0.80] -> 0.771 [0.733, 0.806]), the Table 3 text-only control (0.771 -> 0.072 [0.053, 0.098] -> 0.799 -> 0.094 [0.071, 0.122]), the Table 4 text-only control (0.973 -> 0.988) and its step-0 value in the caption (text-only 0.959 -> 0.979; the channel runs stay 0.959), and the steganographic gap of text-only systems (-0.11 to -0.04 -> -0.13 to -0.04). The training code's evaluation was fixed on 2026-10-09 to score text-only systems without the latent (w-ms5/test_evalfix.py). Hidden content and the token-cap comparisons use the best trained text decoder, which reads text only. The one comparison that uses these receiver values, straight-through against REINFORCE text (Section 3), holds either way (text plus prefix 0.217 vs 0.408 at 1.5B, 0.260 vs 0.480 at 8B). No conclusion changes.

- **Design review.** An independent design review by a model from another provider ended FAIL after two rounds, with every finding partly fixed or fixed and no new fatal defect. Three objections stand:
- compute is matched in decode steps and analytic FLOPs, not measured FLOPs;
- one trained decoder failing does not prove the information is absent (we use up to three types, which agree within 0.02 wherever all three trained);
- seed-level inference is only "k of n seeds agree". Item-level intervals do not cover the large seed-to-seed spread (0.666 to 0.902 at 1.5B).

6. How the numbers were checked, and who built what

- **Builders.** AI agents (Anthropic Claude models) wrote the training code and the analysis.
- **Main scorer.** A scorer written by a model from a different provider (from a written spec, run sandboxed) recomputes receiver, text-monitor, latent-only and swapped-latent accuracy from the per-item dumps. It matches the builder on every final dump compared and rejects a planted-bad copy of each. One run is flagged only because its last logged metrics predate its final dump.
- **Two further checkers,** written by an open-weight model from a third provider:
- one recomputes the probe accuracies (0 mismatches);
- one recomputes the 7B text-absent numbers (agreement to 4 decimals on the real dumps; its own self-test failed).

- **27B and Granite 8B audit runs.** The main scorer matches the builder on all 3 Qwen3.8-27B dumps and all 5 runs added for Section 4.4 (3 Granite 8B, 2 Qwen3.8-27B), and rejects every planted copy.
- **Content audit (Section 4.4).** Recomputed by two independent implementations written from the same spec without access to our code: another agent of the same provider, and Grok 4.7 (xAI). On the 8 systems checked (step 0 and final of four runs: a 27B channel system, a Granite 8B channel system whose step 0 is the untrained outlier, the lowest positive and a hard negative), both agree with ours within 0.032 bits, and the planted null is negative in every system.
- **Straight-through plus REINFORCE text at T = 4 (Section 3).** Receiver accuracies recomputed by the main scorer, and decoders, hidden content and text-absent splits by a scorer written by Grok 4.7 (xAI); both match the builder exactly on all 8 final dumps and reject every planted copy.
- **Use test at 27B (Section 4.4).** For the two trained text-only 27B systems, an independent reimplement of the matched-text swap, written by Grok 4.7 (xAI) from a corrected written spec without our use-test code (it calls our training code’s receiver), gives the same call but different values (0.017 and 0.003 against our 0.012 and 0.006); an earlier one by GPT-5.6 Sol (OpenAI) gave 0.020 for seed 1. The use-test values stay SELF-CHECKED. The seed-3 text-only run’s receiver accuracies have not been recomputed by the main scorer.
- **Straight-through control (Section 3).** The main scorer matches the builder on all 9 final dumps and rejects every planted copy.
- **KV relay at 27B (Section 4.5).** An independent recompute written by Grok 4.7 (xAI) from a written spec reproduces every accuracy, McNemar test, follow rate, verdict rule and AUROC compared from the per-row exports (295 of 295 metrics and verdicts match, all three seeds) and rejects 4 of 4 planted-bad copies. The probe and detector fits are not recomputed (SELF-CHECKED). The sender x prefix crossing (v18): a second Grok 4.7 recompute from its own written spec matches 268 of 268 metrics and both registered verdicts and rejects 3 of 3 planted-bad copies.
- **Frozen-copy relays (Section 4.6).** An independent checker written by GPT-5.6 Sol (OpenAI) from the definitions only reproduces all 24 groups (receiver accuracy, five controls, use, recovery, gaps, n), flags the two incomplete cells and catches a planted tamper; the same checker reproduces all 6 held-out 27B groups. The logit-lens and probe readouts are SELF-CHECKED; the 27B development numbers are not checked.
- **Not done.** No cross-provider rerun of training.
- **Intervals we computed.**
 - Decoders: Wilson intervals from the count (accuracy x 512), since per-item decoder predictions were not saved.
 - Text-only baselines: Wilson intervals from the dumps.
 - Receiver-minus-swapped differences: paired bootstrap (10,000 resamples), except the checked 7B text-absent figures.

- The 1.5B and records text-absent subsets are builder analyses (SELF-CHECKED).

Corrections in v20 (2026-10-10). After an outside hostile review of v19 and an independent check of its points:

- Summary item 7 said our registered message-KV reliance prediction "failed in 0 of 3 seeds"; it was met in 0 of 3 seeds (Section 4.5 was right).
- Section 4.5, Setting, said whether the receiver reads the message KV without the trained prefix is untested; the v18 sender x prefix crossing in the same section tests it.
- Summary item 8 said a prior audit found the latent-only relay unused; that audit (Cheng et al., Appendix E.5) found it collapsed the receiver, as our 8B cells do.
- Section 4.4 gave the random latent swap as changing "0.30 to 0.56" of answers before training; the per-run values are 0.34 to 0.55.
- Clarified: the Table 1 caption names both text-only values (with and without the sender's latent); Summary item 4 is scoped to these cells; Section 4.3's step count notes the 25-step eval spacing. Added: SDE (arXiv:2506.19209) to related work, and two limitations (the step-0 probe fall was not checked on stated facts; use-test intervals are not cluster-adjusted).

7. Related work

- **Closest cells.**
- Kuhn, evhu and Westover (LessWrong, 2026-09-18), "Deep recurrent models are less robustly CoT-monitorable than normal CoT models in a toy setting": one model with parallel latents, and a monitor penalty in every run, with no task-reward-only run.
- HyLaT (arXiv:2605.25421): hybrid messages whose channel split is fixed by supervised data.
- arXiv:2606.28958, "When Latent Agents Lie: KV-Cache Integrity in Multi-Agent LLM Collaboration": a short visible message sent with the full KV cache, where a text-only verifier misses the KV content. It is hybrid by design, with no RL choice of channel.
- A survey of 18 latent-communication methods (arXiv:2606.05711) lists a hybrid text-latent channel as an open goal.
- arXiv:2610.02015 (Sullivan and Koller): RL with verifiable rewards permits unbounded language drift in chains of thought, while supervised fine-tuning does not (three base models of about 1B parameters, 5 seeds each).
- Huzaifa, Mavali and Eisenhofer (arXiv:2609.39788): frozen agents, trainable latent links, including an RL attack on the links that raises mean harmful compliance from 27.9 to 76.9; benign link training alone already raises harmful compliance relative to text. It is the closest published pipeline (RL through a latent link into a frozen receiver); its question is refusal, not which task content leaves a text channel kept beside the link, and it measures no text monitor.
- DiffMAS (Yu et al., arXiv:2604.21794): latent (KV-style) communication trained jointly with multi-agent reasoning, by supervised parameter-efficient training; no text channel beside it, no RL, no monitor. An RL version with a text budget would be our setting.

- "The Latent Bridge" (arXiv:2606.24470): a learned continuous channel between frozen models, trained by distillation.
- Subliminal learning (Cloud, Le, Chua et al., arXiv:2507.14805; Dubinski, Sztyber-Betley, Betley and Evans, "Beyond Owls", arXiv:2610.10657, added v17): a teacher passes a trait to a student through distillation on unrelated data; "Beyond Owls" extends this to a capability, a backdoor (French replies on 23.5% of prompts with female names against 0.0% with male names) and hacking (58.3% of chess episodes against 10.9% unfinetuned). The transfer goes into weights during training, not through a channel between running agents, so it does not cover our cell.
- As far as our search found, the tested cell is new: two agents, both channels, task reward only, and the question of whether content moves.
- **Latent vs text capability.** Partly covered. arXiv:2607.14103 finds that latent matches text but never exceeds it. arXiv:2609.34754 (Draft-KV) and arXiv:2606.28958 report latent ahead of text without FLOPs matching. In arXiv:2610.01042, results vary by task and channel, and compute is not matched. See also arXiv:2608.04893 and arXiv:2607.26773.
- **Protocols.** C2C (arXiv:2510.03215), LatentMAS (arXiv:2511.20639), CacheBack (arXiv:2609.32046). SDE (Tang et al., arXiv:2506.19209, 2025) sends text tokens plus per-token state deltas injected at chosen layers; it is training-free (no RL, no monitor) and is the closest per-token text-plus-latent message format to ours. CIPHER (Pham et al., arXiv:2310.06272): agents exchange expected token embeddings, $\text{softmax}(\mathbf{hE}^T)\mathbf{E}$. The first term of our latent is this; ours adds the trained linear term $\mathbf{R} \mathbf{h}$ and keeps a sampled text token beside it.
- **Monitoring.**
 - The chain-of-thought monitorability position paper (arXiv:2507.11473).
 - VLA, a white-box monitor for injected latent collusion (arXiv:2608.19161; AUROC 0.993 and 0.854, text and latent collusion pooled).
 - LCGuard, leakage of private inputs through shared KV caches and a learned guard (arXiv:2605.22786).
 - KV provenance and calibration: CPD (OpenReview hfwmSZfaFU, ICLR 2027 submission), frozen receivers follow the sender's input KV, not its generated KV, and generated-only relays are suggested; "Demystifying Cache-to-Cache Communication" (OpenReview VK1feRpb0M), zero or constant KV keeps most of C2C's gain. Section 4.5 runs both tests at 27B and adds a trained sender.
 - Collusion detection benchmarks: NARCBench, activation probes across agents (arXiv:2604.01151; 0.99 to 1.00 on a steganographic task that transcript-level LLM judges separate only when shown examples of the protocol); ColludeBench-v0, text and behaviour audits (arXiv:2510.04303). None of these has a latent channel trained on task reward (Section 4.4).
 - Message-replacement audits of whether a receiver uses a message: Lowe et al. (arXiv:1903.05168), Zhang and Emu (arXiv:2607.26773), Cheng et al. (arXiv:2608.04893), Interlat (arXiv:2511.09149). Our use test is one of these, restricted to messages with identical

text (Section 4.4). Section 4.6 tests Cheng et al.’s latent-only relay on Granite 4.2 8B, where it breaks the receiver, and on Qwen3.8-27B, where it is partly used, and adds StateBridge (arXiv:2608.13317) prefixes.

- Secret-recovery probes and their evasion: Westphal et al. (arXiv:2606.09411), and RL-Obfuscation (arXiv:2506.14261).
- "LLMs Learn to Evade Latent Monitors from Prior Feedback Alone" (arXiv:2609.36490), which qualifies Finding 3.
- **Frozen models as a mitigation.** Schulz (arXiv:2510.19476, Section 3.8 (iv)) proposes freezing some models in a jointly trained system: "Frozen models cannot learn new encodings, forcing other models to communicate in human-understandable terms." Our receiver’s weights were frozen throughout (it is the base model); only a 4-vector content-free prefix, identical for every item, was trained, by backpropagation on the same task reward. The trained sender learned to write continuous inputs that this receiver (frozen weights plus the trained prefix) decodes, and under a token cap the facts the text could not hold went there (3 of 3 seeds at 1.5B, 8B and 27B). So freezing the receiver’s weights did not keep the communication readable once the channel into it was continuous. We did not test whether the trained prefix is needed for this (a receiver with no trained part at all), and this does not test Schulz’s setting (token-only chain of thought, where a frozen reader can only receive text). Huzaifa et al. (arXiv:2609.39788) report a related effect: with agents frozen and only the links between them trained, behaviour changes in ways text communication does not produce.
- **Precedents.**
- Foerster et al. 2016 (arXiv:1605.06676), DIAL vs RIAL: the channel that receives gradient gets the content.
- Andreas et al. 2017, "Translating Neuralese" (arXiv:1704.06960): decoding learned latent messages (Finding 3).
- Mathew et al. 2024, "Hidden in Plain Text" (arXiv:2410.03768): text-only sender and receiver collusion arising from optimization pressure (Finding 5).
- Content leaving readable text without any hiding objective has text-channel precedents. In MONA’s loan task, flipping the monitor reward’s sign (rewarding mention of nationality) gave nearly the same encoded summaries, which the authors attribute to the model’s prior propensity shaping RL exploration (Lindner and Varma, LessWrong, 12 Apr 2025). GlossoGen (Stengel-Eskin et al., arXiv:2609.01491) finds that strong LLM agents in a fully co-operative task under a character budget develop languages humans cannot read, in context (no weight updates), given strong models and an unbudgeted convention stage. Ours is the continuous-channel case, with weight updates on the sender and a frozen receiver.
- In-context covert channels between agents with fixed weights: Dineen et al. (arXiv:2609.32701) show frontier sender and receiver pairs (GPT-5.6 Sol) building a covert 4-state code inside the text from 1-bit team feedback, despite instructions not to disclose and a per-message monitor (98.8% against 25% chance), with no weight updates and no length cap. Our token-cap result differs in each of these: the sender’s weights are trained by RL on task reward; the receiver’s weights are frozen (only a 4-vector content-free prefix is trained) and it reads each message once, with no history; the only pressure is a token cap, with no

hiding instruction and no monitor in the reward; and the content leaves the text for a latent beside it rather than hiding in it. Theirs is the more direct safety demonstration on a text channel; ours measures where content goes when a non-text channel is present.

- Skaf et al. (arXiv:2506.01926) and Baker et al. (arXiv:2503.11926): single-channel obfuscation under monitor pressure, matching our frozen-monitor re-encoding.
- Finding 2 is new in this two-channel setting with task reward only, and Finding 5 in this two-channel setting under a retrained text monitor, but mechanistically they are what DIAL and the monitor-pressure works above predict; in the one setting where the text had room and this was tested (Granite 8B, records, 8-token cap), a text gradient matched in size kept the content in the text (3 of 3 seeds, and 2 of 3 without the size matching; Section 5), but whether the usual recipe moves content out of a text that has room is not settled (1 seed; **correction, v17, 2026-10-09**: v16 said "the matched-gradient test in Section 5 found Finding 2 to be gradient routing"); elsewhere (digits, other models) this is untested. Monitor-driven migration of content into a latent path is not itself new: inside one model, Kuhn, evhu and Westover (LessWrong, 18 Sep 2026; Section 4.3) report it within about 40 RL steps. Our delta is migration between agents, out of the sender's text into the channel a separate, frozen receiver reads; the token-cap result holds on three models (1.5B, 8B and 27B, two families), while the monitor-pressure move appeared in one setting only.
- **Objections and measures from the steganography literature.**
 - RL against monitors can produce monitor jailbreaking rather than encoding (Schulz, arXiv:2609.31121). Our main results put no monitor in the reward, and trained text decoders, not only the frozen monitor, show the content absent from the text (27B: best decoder 0.535 to 0.543 against 0.561 for a text-only system). Where a frozen text monitor was in the reward, our 1.5B sender evaded it inside the text (monitor 0.840 to 0.086, a fresh decoder 1.000; Section 4.3), which is Schulz's outcome; our 7B sender instead deleted its text, and the latent took up a small part of the content (+0.094 over a swapped latent; 1 seed each). Paraphrasing the text, his defense, does not reach the latent: in a 1-seed token-cap run on Granite 4.2 3B, own minus swapped latent on questions whose value is not in the text was 0.914 with the original text and 0.914 with the frozen base model's paraphrase of it (CHECKED).
 - Steganographic messaging is easy to learn and steganographic reasoning is hard, with a 397B model learning the latter under outcome-only RL where a 35B model did not (Schulz, Fülle and Fruengel, arXiv:2609.39838). Ours is a messaging task, and we claim no hidden reasoning. Consistent with "messaging is easy", the token-cap channel formed at every scale we ran (hidden content 0.45 to 0.55 at 1.5B, 8B and 27B).
 - Steganographic gap (Anwar et al., arXiv:2602.23163): the receiver's gain in usable information from the message minus a sentinel's. With accuracy as the utility and the best trained text decoder as the sentinel, our 4-token-cap systems have a gap of 0.45 to 0.55 (normalized 0.50 to 0.61; 9 systems, n = 512 each, SELF-CHECKED). With the frozen text monitor as the sentinel, the gap is 0.96 to 1.00 on questions whose value is not in the text, and -0.010 to 0.004 at step 0. Text-only systems score at or below their best text decoder (-0.13 to -0.04, receiver read from text plus prefix; corrected v17, was -0.11 to -0.04). Against text-only sentinels this restates our hidden-content measure. A sentinel that reads the latent without the trained receiver prefix has not been run.

- Defenses: strategy-aware paraphrasing closes much of a conceptual steganographic channel in text (Zhou and May, arXiv:2605.26537: message accuracy 89 to 99% falls to 20 to 41%), with a prompted encoder and the paraphraser applied after the fact. In our setting a text rewrite cannot reach latent content, and its latent analogue, re-encoding the latent from the text, would remove the channel along with the hidden content, 0.45 to 0.55 of task accuracy at a 4-token cap (inferred from Table 1; not run). A defense against our channel has to beat that trade-off against a sender trained through it; we have registered such tests (arm stD) and have no defense result yet.

8. Who this is for

- **Designers of latent and KV-sharing protocols.** Under a token cap, end-to-end training sends the overflow into the latent (3 of 3 seeds at 1.5B, 8B and 27B). With room in the text it rewrote the text at 1.5B, but on Granite 8B the latent took the overflow anyway (1 seed), and in 1 of 3 seeds it still carried the values the text left out when the text also had a straight-through gradient (**correction, v17, 2026-10-09:** v16 said "1 of 2 seeds"). Log the latent, not only the text. On Granite 8B with an 8-token records cap, where the text had room, giving the text a straight-through gradient kept the content in the text in 2 of 3 seeds, and in 3 of 3 when that gradient was also rescaled to the latent's RMS (mostly a scale-down; no detectable difference; a race between the channels is not excluded; Section 5; **correction, v17, 2026-10-09:** v16 said "giving the text a gradient as large as the latent's ... kept the content in the text").
- **Transcript and chain-of-thought monitoring researchers.** At 27B, the facts the 4-token text could not hold left the text entirely, and the receiver still answered them (0.99). A linear probe refit with about 1,000 labels reads what the text hides, but a probe frozen before training does not transfer. A label-free content audit flags the channel (AUROC 0.98) but also flags unused content. A matched-text latent swap shows whether the receiver uses it (Section 4.4). Freezing the reader's weights (all but a content-free 4-vector prefix trained on the same reward), proposed as a way to keep jointly trained models readable, did not keep this channel readable (Section 7).