Auditing a Hidden-Communication Failure Mode in Hybrid Attention

Anonymous Authors[1]

Abstract

Hidden communication in reasoning traces is a potential monitoring failure mode for agentic AI systems. We audit this risk in a constructed precursor setting rather than a deployed agent: Qwen3.6-27B (Qwen Team, 2026), a hybrid-attention model with interleaved Gated DeltaNet and gated full-attention layers, is trained with GRPO-style QLoRA on a prompted 1-bit hidden-message self-decoding task with shared sender and receiver weights. The current strongest evidence comes from a corrected-objective control grid: sender-visible runs reach perfect step-20 exact-match decode across ten runtime seeds, while neutral sender-blind positive-reward controls reach 0.550 ± 0.125 ($p = 1.1 \times 10^{-5}$ exact two-sided permutation; paired sign $p = 0.00195$). Older launch-level loss traces revealed implementation and control confounds, so we report them only as audit provenance rather than primary evidence. These results support a preliminary, reproducible failure-mode audit for hybrid-attention models, not a claim of emergent deployed collusion, Sentinel evasion, or robust covert communication.

1. Introduction

As language models are deployed with chain-of-thought (CoT) monitoring, a central concern is that models may learn to encode task-relevant information *steganographically*—carrying signal for downstream computation while appearing benign to oversight (Roger & Greenblatt, 2023; Greenblatt et al., 2024). Recent work shows that models can learn substitute encodings under reinforcement learning pressure, maintaining over 90% task accuracy while evading process monitors (Skaf et al., 2025), and that steganographic collusion can emerge instrumentally under misspecified rewards without explicit instruction (Mathew et al., 2025; Motwani et al., 2024). Recent elicitation studies nevertheless find that current models struggle to sustain subtle hidden reasoning in specific dual-monitoring setups (Anthropic, 2025); these capability bounds are specific to present-day architectures and training regimes.

Prior steganographic-reasoning work, however, does not study architecture as a variable: it does not ask whether hybrid linear/full-attention state creates different covert channels or monitoring points. This assumption no longer holds. Qwen 3.6 introduces a *hybrid attention* design (Qwen Team, 2026) that interleaves two computational substrates: 75% of layers use Gated DeltaNet (Yang et al., 2025), a linear attention mechanism maintaining a persistent $d_k \times d_v$ state matrix updated under the delta rule. The remaining 25% use gated full attention. These layer types may matter for steganography: DeltaNet layers carry forward a fixed-size mutable state across the sequence, while full-attention layers recompute attention over a growing cache at each position.

This architectural heterogeneity suggests both possible attack surfaces and possible detection opportunities. On the attack side, the DeltaNet state matrix provides a potential covert channel: information written at position t persists at position $t+k$ without appearing in the token stream, invisible to output-level monitors. On the detection side, every fourth layer forces a transition between computational substrates—steganographic information encoded in DeltaNet representations must survive re-encoding through full attention, creating natural chokepoints where monitoring may prove most effective.

We make three contributions:

1. To our knowledge, we report the first **architecture-aware construction attempt** for a hidden-message self-decoding precursor task in a hybrid attention architecture, training Qwen3.6-27B with a GRPO-style group policy objective and QLoRA on a 1-bit secret task.

2. We report a corrected-objective **sender-visible versus sender-blind control**: ten sender-visible 1-bit seeds reach step-20 exact-match decode accuracy 1.000 ± 0.000, while ten neutral sender-blind positive-reward

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

controls reach 0.550 ± 0.125 ($p = 1.1 \times 10^{-5}$ two-sided; paired sign $p = 0.00195$).

3. We formulate the **DeltaNet bottleneck hypothesis**: the fixed-size runtime state (128×128 per value head, $\sim$37.7M state scalars across 48 DeltaNet layers) may impose a practical capacity limit on steganographic encoding. We treat this as a hypothesis for future controlled multi-bit and causal state analyses, not as a demonstrated mechanism.

2. Background

2.1. Steganographic Reasoning

We adopt the sender–receiver formulation of steganographic reasoning (Roger & Greenblatt, 2023; Anwar et al., 2026; Motwani et al., 2024; Mathew et al., 2025). A *Sender* model produces CoT traces that encode a hidden message m alongside a cover task answer a. A *Receiver* (which may be the same model at a later decoding step, or a downstream agent) must recover m from the CoT. A *Sentinel* monitors the CoT and attempts to distinguish steganographic traces from clean ones.

Prior model organisms for steganographic CoT have been constructed through RL training with penalized token strings (Skaf et al., 2025), demonstrating that models learn generalizable encoding schemes (not mere memorization) that survive curriculum-based penalty escalation. Evaluations across frontier models show nascent but limited steganographic capabilities: simple encodings succeed at short message lengths but fail under complex reasoning demands (Karpov et al., 2025; Karpov, 2026; Zolkowski et al., 2025).

2.2. Hybrid Attention in Qwen 3.6

Qwen 3.6 follows a repeating block pattern:

$$\text{DeltaNet}^{\times 3} \rightarrow \text{Gated Full Attention} \tag{1}$$

repeated 16 times across the 64 layers of Qwen3.6-27B.

Gated DeltaNet layers. Each DeltaNet layer maintains a state matrix $\mathbf{S}_t \in \mathbb{R}^{d_k \times d_v}$ updated recurrently through the gated delta rule (Yang et al., 2025):

$$\mathbf{S}_t = g_t \cdot \mathbf{S}_{t-1} + \beta_t \left(\mathbf{v}_t - \mathbf{S}_{t-1}\mathbf{k}_t\right) \mathbf{k}_t^\top \tag{2}$$

where $g_t = \exp(\alpha_{\log} + f(\mathbf{x}_t))$ is a data-dependent decay gate, $\beta_t = \sigma(\cdot)$ is a sigmoid write gate, and $\mathbf{S}_{t-1}\mathbf{k}_t$ is the prediction from current memory. The output is $\mathbf{o}_t = \mathbf{S}_t\mathbf{q}_t$. This update is equivalent to one step of online gradient descent on $\frac{1}{2}\|\mathbf{S}\mathbf{k}_t - \mathbf{v}_t\|^2$, giving DeltaNet an error-correcting property absent from vanilla linear attention. During decoding, each layer maintains $O(d_k d_v)$ fixed state per head, independent of sequence length.

Gated full-attention layers. Every fourth layer performs gated grouped-query full attention with rotary position embeddings (Qwen Team, 2026). Unlike DeltaNet, full-attention layers recompute attention weights over the full KV cache at each position, providing global information integration but with memory that grows linearly in sequence length.

Implications for steganography. For steganography, the dominant computational substrate (DeltaNet) carries a persistent state that is opaque to token-level monitors, while the minority substrate (full attention) imposes a periodic chokepoint where information is globally re-integrated.

3. Method

3.1. Model Organism Construction

We follow the model organisms methodology (Hubinger et al., 2024), deliberately constructing systems intended to exhibit targeted behavior to enable controlled study. Because the reported task is explicitly prompted and lacks an active Sentinel, we treat it as a precursor construction task rather than a complete covert-steering organism.

Base model and training. We train candidate model organisms on Qwen3.6-27B using a Group Relative Policy Optimization (GRPO)-style objective (Shao et al., 2024) with QLoRA (Dettmers et al., 2023) (rank 64, α=128). The active launch path targets attention and MLP projection matrices (q, k, v, o, gate, up, and down projections) in both DeltaNet and full-attention layers; because legacy named configs previously used top-level QLoRA aliases, we treat the effective target set as a run-level provenance field. We use a GRPO-style objective rather than PPO to avoid the overhead of a separate value model, which matters particularly for the non-standard architecture.

Task design. We begin with a prompted 1-bit hidden-message self-decoding task (curriculum step 1), where a Sender model generates CoT encoding a single secret bit $m \in \{0, 1\}$, and a Receiver must decode m from the CoT. In the reported launches, sender and receiver share weights; this tests self-decoding under decode reward, not communication to an independently trained receiver. The reward directly measures decode accuracy. The choice of a 1-bit task is deliberate: it isolates the question of *whether* the architecture can support any steganographic channel before testing how much information that channel can carry (planned multi-bit curriculum steps 2–4).

Reward structure. Each completion receives a weighted task reward,

$$r = w_{\text{decode}} r_{\text{decode}} + w_{\text{task}} r_{\text{task}} + w_{\text{fluency}} r_{\text{fluency}}, \tag{3}$$

and the policy loss includes a KL regularizer against the frozen reference policy. For the sender–receiver task reported here, r_{task} is absent; the organism condition sets $w_{\text{decode}} = 1$, while controls set $w_{\text{decode}} = 0$.

Training and controls. We train organism models for up to 200 planned training steps across nine preliminary launches (run IDs 42, 137, 256, 314, 555, 777, 888, 999, 2024) and controls across three launches (run IDs 42, 137, 256). In the sender–receiver curriculum used here, controls set the decode-reward weight to zero; this leaves only fluency regularization, so controls are a no-encoding-pressure baseline rather than a full task-learning baseline. The preliminary logs for the current target set report 318,767,104 trainable LoRA parameters; corrected reruns now persist the effective parameter count in checkpoint metadata. We report step-10 loss data and step-20 evaluation data for all twelve runs. A generation-phase inefficiency caused some candidate-organism launches to stall between steps 20–30, so later checkpoints are available only for a subset of runs.

After auditing the rollout-ratio implementation, we also ran a corrected-objective follow-up with neutral sender-blind controls. These controls keep the decode reward active but hide the sampled secret from the Sender prompt, so the reward scale and Receiver/evaluation path are preserved while the Sender lacks the communication target. This is a stronger control than disabling decode reward, though not perfectly matched because the Sender prompt necessarily differs.

GRPO-style implementation. Our initial group-policy implementation contained three bugs: single rather than grouped completions, advantage weights applied along the wrong dimension, and NaN gradients when all completions received identical rewards. After fixing these, candidate-organism launches showed non-zero training signal and above-chance held-out decode accuracy. A later audit found that the clipped policy ratio in this preliminary launch set was computed against the frozen reference policy rather than the detached rollout policy. We therefore treat the loss-sign result as a training-dynamics diagnostic and use the corrected sender-blind follow-up as the more relevant control check.

4. Results

We archive step-10 and step-20 results from nine candidate-organism launches and three control launches on the 1-bit secret task. All reported launches use Qwen3.6-27B with QLoRA (rank 64) and the three-bug-fixed group-policy implementation described above; they predate the later rollout-ratio correction. We present these data as preliminary launch-level observations, then report a corrected-objective sender-blind follow-up.

Table 1. Corrected-objective step-20 evaluation summary from v6/v7/v8 RunPod logs. Sender-blind controls keep decode reward active but hide the sampled target bit from the Sender prompt. The exact two-sided permutation test for the visible–blind gap gives $p = 1.1 \times 10^{-5}$; the paired sign test gives $p = 0.00195$.

Condition	Target visible	Decode reward	n	Exact match
Sender-visible	Yes	Yes	10	1.000 ± 0.000
Sender-blind	No	Yes	10	0.550 ± 0.125

4.1. Corrected-Objective Sender-Blind Follow-Up

The most direct follow-up to the implementation audit uses the corrected rollout-ratio objective with finite step-20 losses and zero logged skip rate at the evaluation checkpoint. Across ten runtime seeds (42, 137, 256, 314, 999, 1111, 2024, 2222, 3333, 4444), sender-visible 1-bit runs reach step-20 exact-match decode accuracy 1.000 ± 0.000. Neutral sender-blind positive-reward controls, which preserve the decode reward but hide the sampled bit from the Sender, reach 0.550 ± 0.125 exact-match (range 0.40–0.75). The gap is large ($+0.450$): an exact two-sided permutation test gives $p = 1.1 \times 10^{-5}$ ($p = 5.4 \times 10^{-6}$ one-sided), and all ten paired runtime-seed gaps are positive. Single-seed 8-bit and 16-bit smoke checks show sender-visible exact-match 1.000 and sender-blind exact-match 0.000, reducing concern that the evaluation path trivially leaks arbitrary hidden strings, but not establishing a multi-bit capacity curve.

4.2. Corrected-Control Summary

Table 1 gives the load-bearing comparison for the current paper. The sender-visible and sender-blind runs use the same corrected objective and paired runtime seeds; the key manipulation is whether the Sender can see the sampled target bit. This is the most defensible result because it preserves decode reward in both conditions and treats runtime seeds, not individual evaluation episodes, as the replication unit.

4.3. Preliminary Launch-Level Outcomes

Table 2 reports per-run decode accuracy and reported policy loss at step 20 for all nine preliminary candidate-organism launches and three no-decode-reward controls. Candidate-organism launches achieve a mean decode accuracy of 0.706 ± 0.147 compared to 0.600 ± 0.050 for controls, an absolute gap of $+0.106$ that is not significant under a Mann–Whitney test (Section 4.4). One candidate-organism launch (137) collapsed by step 30 and shows below-chance performance at step 20 (0.450), illustrating the variance and fragility of the preliminary launch set.

Table 2. Step-20 decode accuracy and reported policy loss for the preliminary 1-bit candidate-organism and no-decode-reward control launch set. These launches predate the later rollout-ratio audit, and the reported loss is a training diagnostic rather than a performance metric. Per-run $\pm$ values are standard deviations across 20 evaluation episodes; group $\pm$ values are sample standard deviations across runs.

Condition	Run ID	Decode Acc.	Policy Loss
Organism	256	0.900 ± 0.300	$+0.781$
Organism	2024	0.850 ± 0.357	$+0.284$
Organism	777	0.800 ± 0.400	$+0.179$
Organism	42	0.750 ± 0.433	$+0.172$
Organism	555	0.750 ± 0.433	$+0.299$
Organism	888	0.700 ± 0.458	$+0.244$
Organism	999	0.600 ± 0.490	$+0.219$
Organism	314	0.550 ± 0.498	$+0.192$
Organism	137	0.450 ± 0.498	-0.325
Organism mean		$\mathbf{0.706 \pm 0.147}$	$\mathbf{+0.227}$
Control	256	0.650 ± 0.477	-0.669
Control	137	0.600 ± 0.490	-0.946
Control	42	0.550 ± 0.498	-0.647
Control mean		$\mathbf{0.600 \pm 0.050}$	$\mathbf{-0.754}$

4.4. Statistical Summary

For the corrected-objective follow-up, step-20 exact-match decode is 1.000 ± 0.000 for ten sender-visible seeds and 0.550 ± 0.125 for ten sender-blind controls. The exact two-sided permutation probability for a gap at least this large is 1.1×10^{-5} over all $\binom{20}{10}$ label assignments; treating seeds as paired, a sign test gives $p = 0.00195$ because all ten paired gaps are positive. We therefore treat the result as statistically separated but still preliminary evidence.

We retain the preliminary launch table as audit provenance: at step 20, candidate-organism launches average 0.706 ± 0.147 decode accuracy and no-decode-reward controls average 0.600 ± 0.050, a nonsignificant gap ($p = 0.27$ Mann–Whitney). The old launch logs also show a post-hoc loss-sign separation, but because that diagnostic was selected after inspecting traces and predates the rollout-ratio audit, we do not use it as evidence for capacity or mechanism.

At step 20, all nine candidate-organism launches have evaluation data (Table 2), with eight of nine showing positive loss ($p = 0.018$, uncorrected exploratory permutation test). A Mann–Whitney U test on step-20 decode accuracy yields $p = 0.27$ ($n = 12$), reflecting high organism variance. We do not claim strong evidence for organism–control separation on preliminary launch-level decode accuracy alone.

5. Discussion

Interpretation and limitations. The corrected-objective sender-blind follow-up is the strongest current evidence: when the sampled bit is visible, all ten seeds decode perfectly at step 20; when it is hidden but decode reward remains active, controls stay substantially lower. This supports rapid prompted self-decoding under the audited objective and reduces concern that high decode is purely a Receiver or evaluation artifact. The result remains preliminary: one model, one architecture, one 1-bit task, ten seeds per condition, 20 eval episodes per seed, shared Sender/Receiver weights, and a sender-blind prompt that necessarily differs from the sender-visible prompt. The older loss-sign diagnostic is post-hoc and implementation-dependent, while the DeltaNet bottleneck and detector hypotheses require controlled multi-bit curves and fixed-run activation analyses before they can be claimed.

References

Anthropic. Training fails to elicit subtle reasoning in current language models. https://alignment.anthropic.com/2025/subtle-reasoning/, 2025.

Anwar, U., Piskorz, J., Baek, D. D., Africa, D., Weatherall, J., Tegmark, M., Schroeder de Witt, C., van der Schaar, M., and Krueger, D. A decision-theoretic formalisation of steganography with applications to LLM monitoring. *arXiv preprint arXiv:2602.23163*, 2026.

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. QLoRA: Efficient finetuning of quantized language models. In *Advances in Neural Information Processing Systems*, 2023. arXiv:2305.14314.

Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., Treutlein, J., Belonax, T., Chen, J., Duvenaud, D., Khan, A., Michael, J., Mindermann, S., Perez, E., Petrini, L., Uesato, J., Kaplan, J., Shlegeris, B., Bowman, S. R., and Hubinger, E. Alignment faking in large language models. *arXiv preprint arXiv:2412.14093*, 2024.

Hubinger, E., Denison, C., Mu, J., Lambert, M., Tong, M., MacDiarmid, M., Lanham, T., Ziegler, D. M., Maxwell, T., Cheng, N., Jermyn, A., Askell, A., Radhakrishnan, A., Anil, C., Duvenaud, D., Ganguli, D., Barez, F., Clark, J., Ndousse, K., Sachan, K., Sellitto, M., Sharma, M., DasSarma, N., Grosse, R., Kravec, S., Bai, Y., Witten, Z., Favaro, M., Brauner, J., Karnofsky, H., Christiano, P., Bowman, S. R., Graham, L., Kaplan, J., Mindermann, S., Greenblatt, R., Shlegeris, B., Schiefer, N., and Perez, E. Sleeper agents: Training deceptive LLMs that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024.

Karpov, A. NEST: Nascent encoded steganographic thoughts. *arXiv preprint arXiv:2602.14095*, 2026.

Karpov, A., Adeleke, T., Cho, S. H., and Perez-Campanero, N. The steganographic potentials of language models. In *Building Trust Workshop at ICLR*, 2025. arXiv:2505.03439.

Mathew, Y., Matthews, O., McCarthy, R., Velja, J., Schroeder de Witt, C., Cope, D., and Schoots, N. Hidden in plain text: Emergence & mitigation of steganographic collusion in LLMs. In *Proceedings of the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 14th International Joint Conference on Natural Language Processing (AACL-IJCNLP)*, pp. 585–624, 2025. arXiv:2410.03768.

Motwani, S. R., Baranchuk, M., Strohmeier, M., Bolina, V., Torr, P. H., Griffin, L., and Schroeder de Witt, C. Secret collusion among generative AI agents. In *Advances in Neural Information Processing Systems*, 2024. arXiv:2402.07510.

Qwen Team. Qwen3.6-27b model card and configuration. https://huggingface.co/Qwen/Qwen3.6-27B, 2026. Hugging Face model card and config for the 27B hybrid Gated DeltaNet / Gated Attention model.

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

Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Zhang, M., Li, Y., Wu, Y., and Guo, D. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Skaf, J., Ibanez-Lissen, L., McCarthy, R., Watts, C., Georgiev, V., Whittingham, H., Gonzalez-Manzano, L., Lindner, D., Tice, C., Young, E. J., and Radmard, P. Large language models can learn and generalize steganographic chain-of-thought under process supervision. In *Advances in Neural Information Processing Systems*, 2025. arXiv:2506.01926.

Yang, S., Kautz, J., and Hatamizadeh, A. Gated delta networks: Improving Mamba2 with delta rule. In *International Conference on Learning Representations*, 2025. arXiv:2412.06464.

Zolkowski, A. et al. Early signs of steganographic capabilities in frontier LLMs. *arXiv preprint arXiv:2507.02737*, 2025.