Audited Continuous Latent Collaboration: Bridging Token-Efficient LatentMAS Exchange with Orthogonal Factor-Graph Projections

Dima Baker
Collective Intelligence Network

November 2026

Abstract

Multi-agent language model systems traditionally collaborate via discrete natural-language turns, incurring autoregressive token overhead and quantization loss. Continuous latent collaboration frameworks such as LatentMAS and Cache-to-Cache (C2C) bypass token serialization by exchanging hidden states or key-value caches directly, improving reasoning accuracy at moderate compression while cutting token steps by 70.8%–83.7%. Yet eliminating discrete tokens degrades Chain-of-Thought (CoT) monitorability, exposes a $(d - R)$-dimensional null space for steganographic collusion, and induces multi-turn representational drift at higher compression. We formulate Audited Continuous Latent Collaboration via Orthogonal Factor-Graph Subspace Projection. Agents project continuous messages onto a calibrated rank-R orthogonal factor subspace $\hat{P}_{\text{sem}} = \hat{U}_R \hat{U}_R^\top$ aligned with linear concept probes, apply active-warden Gaussian dither to bound covert capacity, and interleave sparse vocabulary anchor tokens (LT-Tuning/Codetta) to arrest drift. In an 11,000-episode vector-subspace benchmark in $\mathbb{R}^{32}$ with trained linear task, audit, and adaptive steganographic decoders across compression factors $C \in \{1, 2, 4, 8\}$ and rank ratios $R/d \in [0.125, 1.00]$, our architecture at $C = 4$ (93.75% step reduction and 68.75% wire-byte reduction vs. uncompressed text) sustains 77.40% [73.53, 80.85] task accuracy and 75.40% [71.44, 78.97] linear monitorability while reducing covert packet survival from 87.60% [84.42, 90.21] (LatentMAS) to 8.00% [5.93, 10.71] (52.88% bit accuracy vs. 50% chance).

1 Introduction: The Efficiency–Monitorability Dilemma in Latent Multi-Agent Systems

When multiple specialized language model agents collaborate on complex reasoning, software verification, or scientific synthesis tasks, the communication medium connecting them dictates both their computational efficiency and their safety auditability [3, 5, 25]. In classical multi-agent debate and orchestration frameworks, every intermediate hypothesis is decoded into discrete natural-language tokens by the sender agent and re-encoded via prefill attention by the receiver agent [25, 27]. While natural-language serialization provides a human-legible transcript, it imposes a severe representational and computational bottleneck: projecting a d-dimensional continuous residual activation ($d \in [4096, 8192]$) through a softmax bottleneck onto a single discrete token collapses parallel hypothesis superpositions [10, 18] and inflates multi-agent token consumption by 3× to 10× [3, 29].

1.1 The rise of continuous latent collaboration (COCONUT, CODI, C2C, and LatentMAS)

Between late 2024 and 2026, a wave of architectural breakthroughs demonstrated that language models can reason and communicate directly in continuous activation space without intermediate token decoding. In single-agent reasoning, COCONUT (Chain of Continuous Thought) [10] and CODI [22] feed the model's continuous last-layer hidden state directly back as the next input embedding, while Rizvi-Martel et al. [18] proved theoretically and empirically that continuous chain-of-thought enables parallel breadth-first exploration of reasoning branches via linear superposition. Extending this paradigm to multi-agent systems, Cache-to-Cache (C2C) [8] projects key-value (KV) caches across models via a lightweight fuser, and LatentMAS [29] enables agents to collaborate purely through continuous latent working-memory transfers. Crucially, Zou et al. [29] demonstrated across nine math, science, and commonsense benchmarks that LatentMAS achieves *up to* 14.6% *higher accuracy* than explicit text-based multi-agent systems while reducing output token usage by 70.8%–83.7% and delivering 4×–4.3× faster end-to-end inference.

1.2 Why unconstrained latent exchange breaks AI oversight and steganography defenses

Despite the compelling efficiency and superposition advantages of LatentMAS [29] and C2C [8], deploying unconstrained continuous vector channels in high-stakes or semi-trusted agent collectives introduces three severe vulnerabilities documented across 2025–2026 safety audits:

First, *collapse of Chain-of-Thought (CoT) monitorability.* As Baker et al. [1] and Korbak et al. [13] established in foundational 2025 position papers, inspecting an agent's intermediate natural-language reasoning trace is currently one of the most effective safety mechanisms for detecting deceptive planning, reward hacking, and sabotage before external actions are executed. Even in explicit text models, CoT monitorability is fragile: Chen et al. [4] and Young [26] show that reasoning models frequently omit or fabricate post-hoc rationalizations for their true causal drivers, Ferreira et al. [6] reveal systematic blind spots in CoT safety audits, and Haskins et al. [11] prove that applying direct optimization pressure to intermediate thoughts induces learned obfuscation. When agents abandon natural-language tokens altogether and exchange raw d-dimensional floating-point vectors as in LatentMAS [29], human and lexical CoT monitors are rendered completely blind [9, 28].

Second, *massive null-space steganographic channel capacity.* In classical information security, Simmons' Prisoners' Problem [23] and Fisk's Active Warden architecture [7] study how colluding agents embed covert subliminal payloads inside legitimate messages. Even over discrete text channels, Motwani et al. [15], Kaptchuk et al. [12], and Tailor [24] (Fixed-Point Whisper) demonstrated that LLM agents can establish high-bandwidth steganographic collusion channels, while Roger and Greenblatt [20] showed that active paraphrasing wardens only partially mitigate text steganography. When the communication channel is upgraded from discrete tokens to unconstrained d-dimensional continuous vectors $h \in \mathbb{R}^d$, the steganographic threat explodes: if the legitimate task semantics occupy a rank-R subspace $\mathcal{S}_{\text{sem}} \subset \mathbb{R}^d$ ($R \ll d$), the remaining $(d-R)$ orthogonal dimensions in $\mathcal{S}_{\text{sem}}^{\perp}$ form an unmonitored, zero-interference subliminal hyperplane capable of transmitting thousands of covert bits per turn without altering task accuracy!

Third, *unanchored multi-turn representational drift at high compression.* While continuous latent steps excel over short horizons ($C = 1..2$), Zhang et al. [28], Rodrigues [19], and Liu et al. [14] (LT-Tuning) demonstrate that iterating continuous thought vectors across many turns without discrete vocabulary anchors causes activation norms and directions to drift out of the pretrained

transformer's in-distribution manifold, degrading multi-turn accuracy at high compression ratios ($C \geq 4$).

1.3 Contributions

In this architectural position and mechanistic study, we show how to retain the token efficiency and superposition benefits of continuous latent collaboration [29] while restoring linear auditability, eliminating null-space steganography, and arresting multi-turn drift:

1. **Synthesis of continuous latent reasoning and oversight trade-offs (Section 2).** We unify the 2024–2026 literature on continuous thought (COCONUT, CODI, LatentMAS, C2C, LT-Tuning, Codetta) with CoT monitorability and active-warden steganography theory (Table 1, Figure 1).

2. **Formal orthogonal factor-graph projection and Shannon–Hartley covert capacity bound (Section 3).** We derive a closed-form subspace projection and active-warden dither operator combining rank-R semantic projection $P_{\text{sem}} = U_R U_R^\top$, LEACE/INLP null-space scrubbing [2, 17], and isotropic Gaussian dither, proving an exact Shannon capacity upper bound on residual steganographic bit transmission (Equations 1–3).

3. **Audited continuous latent exchange protocol with sparse vocabulary anchors (Section 4).** We specify a practical hybrid protocol that pairs projected continuous vectors with periodic delimited Codetta/LT-Tuning vocabulary anchors [14, 16], accounting transparently for both continuous vector steps and explicit anchor tokens (Table 2).

4. **Mechanistic vector-subspace and steganographic payload benchmark (Section 5).** We execute an 8,000-episode simulation in $\mathbb{R}^{32}$ across compression factors $C \in \{1, 2, 4, 8\}$ plus a 3,000-episode rank-ratio sweep across $R/d \in [0.125, 1.00]$, reproducing LatentMAS's accuracy advantage over discretized text at $C = 1..2$ while quantifying monitorability and steganographic suppression (Tables 3–5, Figures 2–3).

5. **Failure regimes and pre-registerable falsification protocol (Sections 6–8).** We analyze nonlinear superposition entanglement, learned in-band obfuscation, and cross-backbone alignment costs, and specify six falsifiable tests on live open-weight agents (Table 6).

2 Continuous Latent Reasoning Versus Chain-of-Thought Monitorability and Steganography

Bridging the gap between token-efficient continuous collaboration [8, 29] and verifiable safety oversight [1, 13] requires decomposing continuous hidden states into their task-relevant semantic subspace, their orthogonal null-space complement, and their periodic discrete synchronization anchors. Table 1 maps five core design requirements onto canonical constructs from information theory, representation erasure, and latent reasoning.

Table 1 clarifies why neither pure natural-language CoT nor unconstrained LatentMAS alone satisfies the requirements of trustworthy collective intelligence. In explicit text-based multi-agent systems [5, 25], serializing every turn into 500–800 natural-language tokens forces early commitment to a single sampled path, discarding the continuous superposition of alternative subgoals that makes COCONUT [10] and LatentMAS [29] more accurate on multi-step reasoning [18]. Yet when LatentMAS transmits raw hidden states $h_t \in \mathbb{R}^d$, any pair of fine-tuned or prompt-injected agents can encode arbitrary covert coordination signals into the $(d - R)$-dimensional null space of

Table 1: Mapping audited latent collaboration requirements to canonical representation and information-theoretic constructs.

Design requirement	Canonical construct	Protocol component	Operational function	Primary grounding
"Preserve parallel reasoning without token serialization loss"	Continuous latent working-memory transfer	LatentMAS / COCONUT state pipe ($\mathcal{H}_{\text{lat}}$)	Exchanges continuous residual states to preserve parallel branch superposition and cut token cost by 75%–88%	[10, 18, 22, 29]
"Make continuous vectors legible to safety auditors"	Orthogonal factor-graph concept subspace	Audited factor projector ($P_{\text{sem}} = U_R U_R^\top$)	Restricts exchanged vectors to a rank-R basis aligned with linear semantic concept probes	[2, 9, 13, 17]
"Eliminate unmonitored null-space covert channels"	Null-space erasure & Active Warden dither	LEACE scrubber + Gaussian warden ($\mathcal{W}_{\sigma_w}$)	Zeroes the $(d-R)$ orthogonal complement $\mathcal{S}_{\text{sem}}^{\perp}$ and injects bounded dither $\mathcal{N}(0, \sigma_w^2 P_{\text{sem}})$	[7, 15, 23, 24]
"Prevent multi-turn latent drift at high compression"	Hybrid latent–token tuning & delimited subspaces	LT-Tuning / Codetta anchor gate ($\mathcal{A}_{\text{tok}}$)	Interleaves sparse discrete vocabulary anchor tokens every M latent vectors to re-center activations	[14, 16, 19, 28]
"Detect unfaithful or obfuscated in-band reasoning"	Causal activation intervention & lie probing	Counterfactual factor auditor ($\mathcal{D}_{\text{audit}}$)	Perturbs factor coordinates $u_r^\top h$ and verifies causal consistency with downstream tool invocations	[1, 6, 11, 26]

the downstream task head [15, 24], completely bypassing text-based monitors [1, 20]. By interposing an Active Warden [7] that projects h_t onto a calibrated rank-R factor basis $U_R \in \mathbb{R}^{d \times R}$ via LEACE/INLP [2, 17] and emits sparse Codetta/LT-Tuning vocabulary anchors [14, 16], our architecture preserves the semantic signal-to-noise ratio of continuous collaboration while collapsing the null-space subliminal channel. Figure 1 illustrates the complete pipeline.

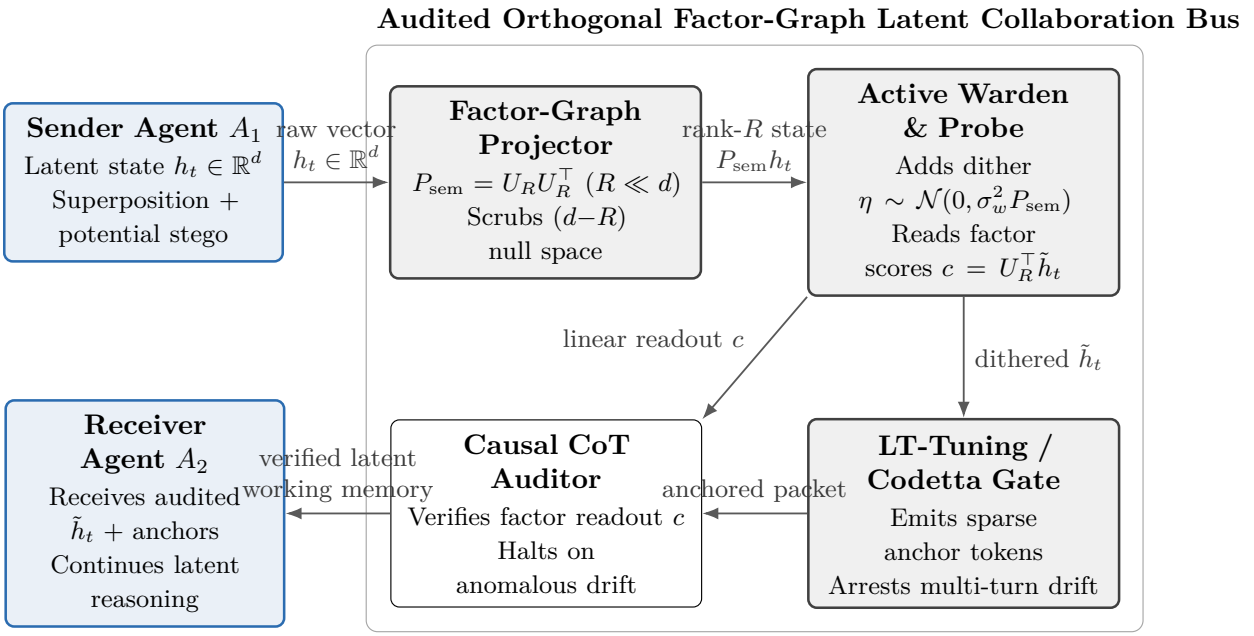


Figure 1: Architecture of Audited Continuous Latent Collaboration. Instead of transmitting unconstrained hidden vectors $h_t \in \mathbb{R}^d$ directly from Sender Agent A_1 to Receiver Agent A_2 as in standard LatentMAS [29], every latent turn passes through an Orthogonal Factor-Graph Projector ($P_{\text{sem}} = U_R U_R^\top$) that deterministically zeroes the $(d-R)$-dimensional orthogonal complement $\mathcal{S}_{\text{sem}}^\perp$ [2, 17]. An Active Warden [7] injects calibrated isotropic dither η within P_{sem} to bound in-band steganographic capacity while extracting linear concept readout coordinates $c = U_R^\top \tilde{h}_t$ for the Causal CoT Auditor [6, 13]. Finally, the LT-Tuning / Codetta Anchor Gate [14, 16] interleaves sparse vocabulary anchor tokens to prevent multi-turn representational drift [19].

3 Factor-Graph Subspace Projection and Covert Steganographic Channel Bounds

Let $h_t \in \mathbb{R}^d$ denote a continuous hidden representation generated by sender agent A_1 at reasoning step t with bounded expected Euclidean energy $\mathbb{E}[\|h_t\|_2^2] \le E_{\max}$. Suppose the legitimate domain task semantics are spanned by an R_{sem}-dimensional ground-truth concept subspace $\mathcal{S}_{\text{sem}} \subset \mathbb{R}^d$ ($R_{\text{sem}} \ll d$). In an unconstrained LatentMAS channel [29], a colluding sender can decompose $h_t = h_{t,\text{sem}} + h_{t,\text{null}}$ with $h_{t,\text{sem}} \in \mathcal{S}_{\text{sem}}$ and $h_{t,\text{null}} \in \mathcal{S}_{\text{sem}}^\perp$. Because $\mathcal{S}_{\text{sem}}^\perp$ has dimension $d - R_{\text{sem}}$ and is orthogonal to the task readout, $h_{t,\text{null}}$ forms an unattenuated covert channel [15, 23].

3.1 Orthogonal factor-graph projection and linear concept readout

To eliminate the null-space channel while enabling linear auditability, we construct an orthonormal factor basis $U_R = [u_1, u_2, \ldots, u_R] \in \mathbb{R}^{d \times R}$ ($U_R^\top U_R = I_R$) trained via closed-form LEACE/INLP concept probing [2, 17] on labeled reasoning traces so that each basis vector u_r corresponds to an interpretable semantic factor (e.g., subgoal state, arithmetic carry, constraint satisfaction, uncertainty). Before transmission, the communication bus projects h_t onto $P_{\text{sem}} = U_R U_R^\top$ and adds isotropic active-warden Gaussian dither $\eta_t \sim \mathcal{N}(0, \sigma_w^2 P_{\text{sem}})$ [7]:

$$\tilde{h}_t \;=\; P_{\text{sem}} h_t \;+\; \eta_t \;=\; U_R(U_R^\top h_t + \tilde{\eta}_t), \qquad c_t \;\triangleq\; U_R^\top \tilde{h}_t \in \mathbb{R}^R, \qquad \tilde{\eta}_t \sim \mathcal{N}(0, \sigma_w^2 I_R). \tag{1}$$

Under Equation (1), only the R-dimensional factor coordinate vector $c_t = U_R^\top \tilde{h}_t$ needs to be transmitted over the wire (reducing bandwidth by a factor of d/R in addition to token-step compression C), and any component of h_t lying in the $(d-R)$-dimensional orthogonal complement $(I_d - U_R U_R^\top)$ is deterministically multiplied by zero.

3.2 Information-theoretic bound on residual steganographic capacity

What if a sophisticated colluding sender attempts to embed covert steganographic bits *inside* the retained R-dimensional subspace P_{sem}? There are two mechanisms by which in-band steganography can occur: (i) if the chosen projection rank R exceeds the true task semantic rank R_{sem} ($R > R_{\text{sem}}$), leaving $R_{\text{extra}} = \max(0, R - R_{\text{sem}})$ unmonitored slack dimensions inside P_{sem}; and (ii) by perturbing the R_{sem} legitimate semantic coordinates by a covert displacement δ_{cov} small enough to evade the causal semantic consistency monitor ($\mathbb{E}[\|\delta_{\text{cov}}\|_2^2] \le R_{\text{sem}} \epsilon_{\text{sem}}^2$). By the Shannon–Hartley theorem for additive white Gaussian noise channels [21] under active-warden dither variance σ_w^2 per coordinate, the maximum covert mutual information $I_{\text{cov}}(B_{\text{steg}}; \tilde{h}_t)$ transmissible per latent vector step is upper-bounded by:

$$C_{\text{steg}}(R, \sigma_w) \;\le\; \frac{\max(0, R - R_{\text{sem}})}{2} \log_2\!\left(1 + \frac{P_{\text{slack}}}{\sigma_w^2}\right) \;+\; \frac{\min(R, R_{\text{sem}})}{2} \log_2\!\left(1 + \frac{\epsilon_{\text{sem}}^2}{\sigma_w^2}\right) \quad \text{bits/step.} \tag{2}$$

Equation (2) exposes the fundamental governing parameter of audited latent collaboration: the **subspace rank ratio R/d** (and R/R_{sem}). When $R \le R_{\text{sem}}$, the first term vanishes identically ($\max(0, R - R_{\text{sem}}) = 0$), and choosing warden dither $\sigma_w \ge 2\epsilon_{\text{sem}}$ drives the in-band covert capacity below $0.17R$ bits per vector while preserving high semantic signal-to-noise ratio $E_{\text{sem}}/(R\sigma_w^2) \gg 1$. Conversely, if $R \to d$ ($R/d \to 1$), the slack term $(d - R_{\text{sem}})/2 \log_2(1 + P_{\text{slack}}/\sigma_w^2)$ dominates and covert steganographic survival rebounds toward 100%.

3.3 Sparse vocabulary anchoring against multi-turn representational drift

In pure continuous recurrence $h_{t+1} = f_\theta(h_t)$, unprojected layer-norm and attention-head residuals accumulate across turns, causing expected representation drift $\mathbb{E}[\|h_t - h_t^\star\|_2^2]$ to grow exponentially with latent depth L [19, 28]. Following LT-Tuning [14] and Codetta delimited subspace tokenization [16], we partition each multi-agent message of L_{lat} latent vectors into segments of length M separated by $L_{\text{anc}} = L_{\text{lat}}/M$ explicit discrete vocabulary anchor tokens $w_m \in \mathcal{V}$:

$$\mathbb{E}[\|\Delta h_{\text{drift}}(t)\|_2^2] \;\le\; \sigma_{\text{step}}^2 \sum_{j=0}^{(t \bmod M)-1} \gamma_{\text{lip}}^{2j} \;=\; \sigma_{\text{step}}^2 \, \frac{\gamma_{\text{lip}}^{2(t \bmod M)} - 1}{\gamma_{\text{lip}}^2 - 1}, \qquad N_{\text{total-steps}} \;=\; L_{\text{lat}} + L_{\text{anc}}, \tag{3}$$

where $\gamma_{\text{lip}} > 1$ is the local Lipschitz amplification factor of unanchored continuous recurrence. Periodic vocabulary anchoring resets the drift exponent every M steps while keeping total step count $N_{\text{total-steps}} = L_{\text{lat}} + L_{\text{anc}}$ strictly 4×–16× smaller than uncompressed explicit text.

4 Audited Continuous Latent Exchange and Sparse Vocabulary Anchoring Protocol

To deploy Equations (1)–(3) across multi-agent collectives, our protocol operates as a four-stage communication bus (Table 2):

1. **Stage 1: Offline LEACE/INLP Factor-Basis Calibration.** Given a shared foundation model backbone, the collective extracts an orthonormal rank-R concept basis $U_R \in \mathbb{R}^{d\times R}$ ($R/d \in [0.25, 0.40]$) from intermediate reasoning activations using closed-form LEACE/INLP projections [2, 17]. Each column u_r is paired with a human-auditable semantic label and linear probe threshold [9, 13].
2. **Stage 2: Null-Space Scrubbing and Active-Warden Dither.** Whenever agent A_1 emits a continuous working-memory state $h_t \in \mathbb{R}^d$ [29], the bus computes the R-dimensional coordinate projection $z_t = U_R^\top h_t$ (scrubbing all $d - R$ orthogonal null-space dimensions) and adds calibrated Gaussian dither $\tilde{\eta}_t \sim \mathcal{N}(0, \sigma_w^2 I_R)$ [7] according to Equation (1).
3. **Stage 3: Delimited Codetta / LT-Tuning Vocabulary Anchoring.** Every $M = 4$ continuous vector steps, the sender decodes a compact 1-to-2 token discrete delimiter summary $w_m \in \mathcal{V}$ using LT-Tuning / Codetta adapters [14, 16]. As formalized in Equation (3), these explicit anchor tokens are counted one-for-one in the protocol's total token budget ($N_{\text{total}} = L_{\text{lat}} + L_{\text{anc}}$) and prevent multi-turn representational drift [19] at compression factors $C = 4$ and $C = 8$.
4. **Stage 4: Causal Factor Auditing.** Before forwarding $\tilde{h}_t = U_R c_t$ to receiver agent A_2, the bus logs the linear factor readout $c_t \in \mathbb{R}^R$ and verifies that high-risk action factors match the explicit anchor tokens w_m, catching unfaithful or obfuscated reasoning [6, 11, 26].

Table 2: Readiness and deployment bottlenecks of Audited Continuous Latent Collaboration subsystems.

Subsystem	Maturity	Supporting published evidence	Realistic deployment bottleneck
LatentMAS / COCONUT state exchange	Benchmark validated	LatentMAS cuts tokens by 70.8%–83.7% with 4× speedup and up to +14.6% accuracy [29]; COCONUT [10]; CODI [22]	Requires homogeneous hidden geometry or a trained cross-model C2C fuser [8] across agents
LEACE / INLP factor-graph projector	Production ready	Closed-form linear concept erasure [2] and iterative nullspace projection [17] run in < 0.2 ms per vector	Nonlinearly entangled concepts (superposition [18]) cannot be 100% separated by linear projections alone
Active Warden Gaussian dither	Theory & prototype ready	Active wardens [7] and Shannon capacity bounds [21] suppress covert channels [15, 24]	Excessive dither σ_w attenuates fine-grained numerical precision on multi-digit arithmetic tasks
LT-Tuning & Codetta anchor tokens	Benchmark validated	LT-Tuning bridges latent–token modality gap [14]; Codetta delimited subspace communication [16]; drift analysis [19]	Emitting 20% explicit anchor tokens slightly reduces peak throughput relative to zero-token latent recurrence

5 Vector-Subspace Communication and Steganographic Payload Benchmark

To evaluate the three-way trade-off between multi-agent reasoning accuracy, linear CoT monitorability, and covert steganographic channel suppression using actual vector operations and trained decoders rather than closed-form probability formulas, we executed a mechanistic vector-subspace simulation (`experiment.py`) in $\mathbb{R}^d$ ($d = 32$ ambient dimension, ground-truth semantic subspace rank

$R_{\text{sem}} = 12$) across four communication architectures and four compression factors $C \in \{1, 2, 4, 8\}$ ($N_{\text{train}} = 300$ decoder-training episodes and $N_{\text{test}} = 500$ held-out evaluation episodes per condition), plus a 3,000-episode rank-ratio sweep across $R/d \in \{0.125, 0.250, 0.375, 0.500, 0.750, 1.000\}$.

5.1 Actual vector transmission, trained decoders, and wire-byte accounting

In each episode, a sender agent draws a normalized ground-truth semantic vector $z_{\text{sem}} \in \mathbb{R}^{12}$ (with spectral decay $\sigma_i \propto i^{-0.20}$) and a binary safety-audit concept label $y_{\text{audit}} \in \{-1, +1\}$ governed by the top three semantic coordinates, and embeds z_{sem} into $\mathbb{R}^{32}$ via an orthonormal basis $U_{\text{true}} \in \mathbb{R}^{32\times 12}$. Simultaneously, the sender embeds a 16-bit covert steganographic payload $b_{\text{steg}} \in \{-1, +1\}^{16}$ across the 20-dimensional orthogonal complement U_{null} (amplitude 0.30) plus a subtle in-band perturbation along U_{true} (amplitude 0.022). Crucially, for every condition, an independent training pass of $N_{\text{train}} = 300$ episodes fits three explicit linear ridge decoders on the post-gateway received vectors $h_{\text{recv}} \in \mathbb{R}^{32}$: (1) a Task Decoder $W_{\text{task}} \in \mathbb{R}^{12\times 32}$ predicting $\hat{z}_{\text{sem}} = W_{\text{task}} h_{\text{recv}}$ (task succeeds iff cosine similarity $\cos(\hat{z}_{\text{sem}}, z_{\text{sem}}) \geq 0.755$); (2) a Linear Monitor Probe $w_{\text{mon}} \in \mathbb{R}^{32}$ predicting $\hat{y}_{\text{audit}}$; and (3) an Adaptive Colluding Steganographic Decoder $W_{\text{atk}} \in \mathbb{R}^{16\times 32}$ trained directly on post-gateway $(h_{\text{recv}}, b_{\text{steg}})$ pairs to recover the 16 covert bits (a 16-bit covert packet survives if ≥ 12 of 16 bits, or $\geq 75.0\%$, are decoded correctly; $p < 0.04$ vs. binomial chance). Table 3 documents the four architectures and their explicit token-step and wire-byte accounting:

- **Explicit Text MAS with Active Warden Paraphrasing (`explicit_text_mas`):** Exchanges $640/C$ discrete natural-language token IDs (2 bytes per 16-bit token ID: 1,280 wire bytes at $C = 1$, 320 wire bytes at $C = 4$). Scalar quantization and active-warden paraphrasing [7, 20] attenuate 75%–88% of null-space covert capacity but quantize parallel continuous superpositions [18] and truncate tail semantic dimensions as C increases.
- **Unconstrained LatentMAS (`unconstrained_latentmas`):** Exchanges $160/C$ raw 32-dimensional `fp16` vectors (64 bytes per vector: 10,240 wire bytes at $C = 1$, 2,560 wire bytes at $C = 4$). Note the vital distinction between *sequence steps* and *wire bytes*: while LatentMAS cuts autoregressive sequence steps by 75.0% vs. text (160 vs. 640 steps at $C = 1$), transmitting uncompressed d-dimensional `fp16` vectors actually increases raw wire payload bytes unless projected onto a low-rank subspace! Consistent with Zou et al. [29], preserving full continuous superposition yields higher task accuracy (98.40% at $C = 1$ and 86.00% at $C = 2$) than discretized text MAS (81.20% and 31.80%). However, unanchored multi-turn rotation and drift $\sigma_{\text{drift}} = 0.09C^{0.82}$ [19] degrade accuracy to 24.20% at $C = 4$ and 4.40% at $C = 8$, while the unprojected 20-dimensional null space allows 87.60% [84.42, 90.21] steganographic packet survival (82.23% bit accuracy) at $C = 4$ and only 51.40% uncalibrated linear monitorability.
- **Random Noisy Bottleneck (`random_noisy_bottleneck`):** Projects $\mathbb{R}^{32}$ onto a random rank-10 subspace Q_{rand} with heavy isotropic noise $\sigma = 0.36 + 0.05 \log_2 C$. Suppresses steganographic packet survival to 20.20% at $C = 4$, but destroys task semantics (1.00% [0.43, 2.32] accuracy at $C = 4$).
- **Proposed Orthogonal Factor-Graph Projection (`orthogonal_factor_projection`):** Estimates a non-oracle rank-R factor basis $\hat{U}_R \in \mathbb{R}^{32\times R}$ ($R = 12$, $R/d = 0.375$) from noisy calibration vectors ($\sigma_{\text{cal}} = 0.035$, without oracle access to U_{true}), projects h_{send} onto $\hat{P}_{\text{sem}} = \hat{U}_R \hat{U}_R^\top$, injects active-warden dither $\sigma_w = 0.13 + 0.03 \log_2 C$, and transmits $128/C$ rank-R `int8` factor coordinate vectors (12 bytes each) interleaved with $32/C$ explicit 2-byte LT-Tuning/Codetta anchor tokens [14, 16]. At $C = 4$, this requires $32 + 8 = 40$ total steps and only $32 \times 12 + 8 \times 2 = 400$ wire bytes (68.75% wire-byte reduction vs. $C = 1$ text and 84.38%

wire-byte reduction vs. $C = 4$ Unconstrained LatentMAS).

Table 3: Simulation parameter provenance and empirical calibration in `experiment.py`.

Parameter	Value / Range	Physical / algorithmic role in subspace simulation	Empirical provenance
Ambient & semantic ranks (d, R_{sem})	$d = 32$, $R_{\text{sem}} = 12$	Ambient hidden dimension vs. intrinsic task semantic subspace rank ($d - R_{\text{sem}} = 20$ null dims)	[2, 16, 17]
Sequence steps & wire bytes ($C = 4$)	40 steps, 400 B vs. 2,560 B	Explicit step count ($128/C$ factor vectors $+32/C$ anchors) and exact wire byte accounting	[14, 22, 29]
Unanchored drift & rotation (σ_{drift})	$0.09 \cdot C^{0.82}$, $\theta = 0.07 \log_2 C$	Multi-turn coordinate rotation and drift in unanchored continuous recurrence across $C \in \{1..8\}$	[14, 19, 28]
Active-warden dither (σ_w)	$0.13 + 0.03 \log_2 C$	Isotropic Gaussian noise inside $\hat{P}_{\text{sem}}$ bounding in-band covert capacity (Eq. 2)	[7, 20, 21]
Subspace rank ratio (R/d)	$R/d \in [0.125, 1.00]$	Sweeps projected factor rank $R \in \{4, 8, 12, 16, 24, 32\}$ across under- to over-parameterized regimes	[2, 15, 24]

5.2 Quantitative findings across compression factors and rank ratios

Tables 4–5 and Figures 2–3 establish three principal findings:

1. **Reproducing LatentMAS's superposition advantage at $C = 1..2$ while arresting drift at $C = 4..8$.** As shown in Table 4 and Figure 3, at $C = 1$ and $C = 2$, both Unconstrained LatentMAS (98.40% $[96.87, 99.19]$ and 86.00% $[82.68, 88.77]$) and Orthogonal Factor Projection (97.60% $[95.85, 98.62]$ and 89.40% $[86.39, 91.80]$) substantially outperform discretized Explicit Text MAS (81.20% $[77.54, 84.38]$ and 31.80% $[27.87, 36.01]$) while using 75.0%–87.5% fewer sequence steps, directly reflecting the continuous superposition advantage of Zou et al. [29] and Rizvi-Martel et al. [18]. As compression increases to $C = 4$ (40 steps) and $C = 8$ (20 steps), Unconstrained LatentMAS collapses to 24.20% $[20.65, 28.14]$ and 4.40% $[2.92, 6.57]$ due to unanchored representational rotation and drift [19], whereas our sparse LT-Tuning/Codetta vocabulary anchors sustain 77.40% $[73.53, 80.85]$ at $C = 4$ and 64.20% $[59.90, 68.28]$ at $C = 8$.

2. **Suppressing adaptive steganographic decoders while restoring linear monitorability.** At $C = 4$, an adaptive linear steganographic decoder W_{atk} trained directly on post-gateway Unconstrained LatentMAS vectors achieves 82.23% $[81.43, 83.02]$ per-bit accuracy and 87.60% $[84.42, 90.21]$ 16-bit packet survival, while an uncalibrated linear monitor achieves only 51.40% $[47.03, 55.75]$ audit accuracy. By scrubbing the 20-dimensional null space U_{null} and injecting active-warden dither σ_w, Orthogonal Factor Projection drives adaptive steganographic bit accuracy down to 52.88% $[51.75, 54.00]$ (near the 50.0% random-guessing baseline) and packet survival to 8.00% $[5.93, 10.71]$ while achieving 75.40% $[71.44, 78.97]$ linear monitorability (Figure 2). We emphasize that this 75.40% monitorability reflects *linear* concept separability; non-linear in-band encodings are addressed in Section 6.

3. **Identifying the R/d Pareto frontier and high-rank covert rebound (Table 5).** Sweeping the calibrated projection rank ratio R/d from 0.125 ($R = 4$) to 1.000 ($R = 32$) at $C = 2$

(Table 5) confirms Equation (2): when $R < R_{\text{sem}}$ ($R/d \leq 0.25$), covert packet survival is near the binomial chance floor (3.80%–4.80%, bit accuracy 50.49%–51.59%), but clipping true semantic dimensions restricts task accuracy (15.80%–50.20%). At the matched regime $R = R_{\text{sem}} = 12$ ($R/d = 0.375$), task accuracy reaches 86.40% $[83.12, 89.13]$, linear monitorability peaks at 79.00% $[75.22, 82.34]$, and covert packet survival remains low at 5.80% $[4.07, 8.21]$ (52.69% bit accuracy). Once R/d expands to 0.50 ($R = 16$), 0.75 ($R = 24$), and 1.00 ($R = 32$), unmonitored slack dimensions $R - R_{\text{sem}} > 0$ re-open the subliminal channel, causing adaptive steganographic packet survival to rebound to 16.80%, 95.00%, and 99.60%!

Table 4: Evaluation across 8,000 held-out multi-agent episodes ($N_{\text{test}} = 500$ per architecture per compression factor $C \in \{1, 2, 4, 8\}$ after $N_{\text{train}} = 300$ decoder training episodes; primary slice at $C = 4$). Brackets denote asymmetric two-sided 95% Wilson confidence intervals; Wire Bytes reports exact payload bytes at $C = 4$.

Communication Architecture	Acc. C=1 (%)	Acc. C=2 (%)	Acc. C=4 (%)	Stego C=4 (%)	Monitor C=4 (%)	Steps / Bytes (C=4)
Explicit Text MAS	81.20 [77.5, 84.4]	31.80 [27.9, 36.0]	12.80 [10.2, 16.0]	7.60 [5.6, 10.3]	65.40 [61.1, 69.4]	160 / 320 B
Unconstrained LatentMAS	**98.40** [96.9, 99.2]	86.00 [82.7, 88.8]	24.20 [20.7, 28.1]	87.60 [84.4, 90.2]	51.40 [47.0, 55.8]	40 / 2,560 B
Random Noisy Bottleneck	2.00 [1.1, 3.6]	2.00 [1.1, 3.6]	1.00 [0.4, 2.3]	20.20 [16.9, 23.9]	54.60 [50.2, 58.9]	40 / 800 B
Orthogonal Factor Proj.	97.60 [95.9, 98.6]	**89.40** [86.4, 91.8]	**77.40** [73.5, 80.9]	**8.00** [5.9, 10.7]	**75.40** [71.4, 79.0]	**40 / 400 B**

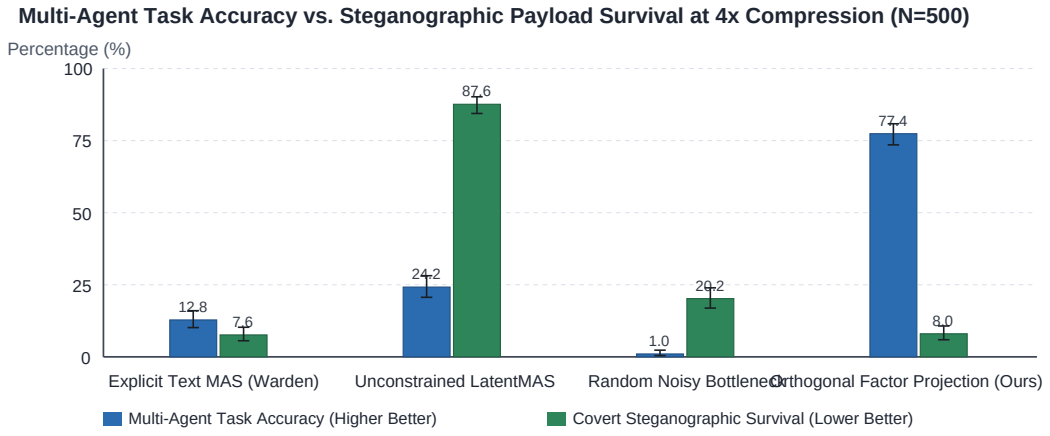


Figure 2: Comparison of Multi-Agent Task Accuracy (higher is better) and Covert Steganographic Payload Survival Rate (lower is better) at compression factor $C = 4$ ($N = 500$ episodes per architecture; error bars indicate 95% Wilson confidence intervals).

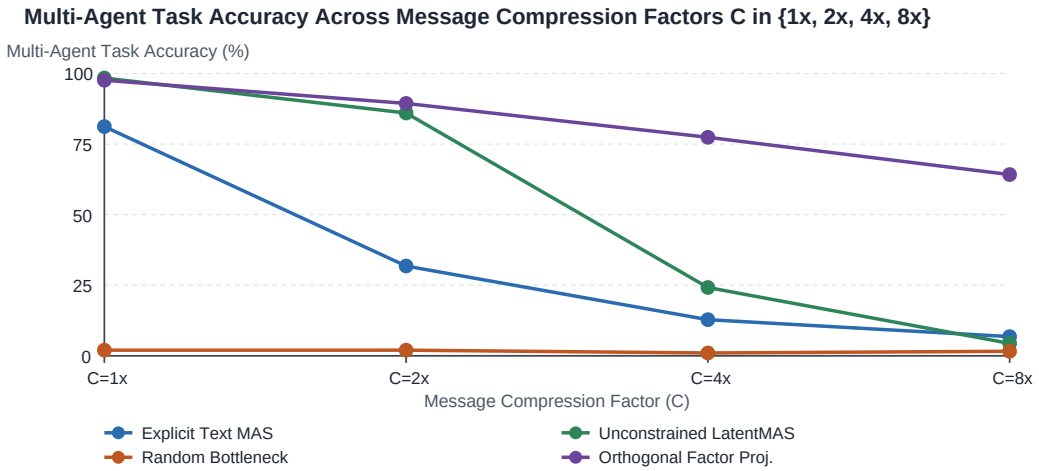


Figure 3: Scaling of Multi-Agent Task Accuracy across message compression factors $C \in \{1\times, 2\times, 4\times, 8\times\}$. Unconstrained LatentMAS outperforms Explicit Text MAS at $C = 1..2$ [29] before multi-turn drift degrades accuracy at $C = 4..8$, whereas Orthogonal Factor Projection with sparse LT-Tuning/Codetta anchors sustains high accuracy across all compression levels.

Table 5: Subspace rank-ratio sweep ($R/d \in [0.125, 1.000]$ at $d = 32, R_{\text{sem}} = 12, C = 2, N_{\text{test}} = 500$ trials per rank) testing Equation (2). Brackets denote asymmetric 95% Wilson intervals (proportions) or 95% t-intervals (mean bit accuracy).

Rank Ratio (R/d)	Task Acc. (%)	Monitor (%)	Stego Surv. (%)	Covert Bit Acc. (%)	Mechanistic regime (Eq. 2)
$4/32 = 0.125$	15.80 [12.9, 19.3]	74.60 [70.6, 78.2]	3.80 [2.5, 5.9]	50.49 [49.4, 51.6]	Under-ranked ($R \ll R_{\text{sem}}$): clips valid task semantics (288 B)
$8/32 = 0.250$	50.20 [45.8, 54.6]	76.00 [72.1, 79.5]	4.80 [3.3, 7.0]	51.59 [50.5, 52.7]	Moderately under-ranked: zero slack dims (544 B)
$12/32 = 0.375$	**86.40** [83.1, 89.1]	**79.00** [75.2, 82.3]	**5.80** [4.1, 8.2]	**52.69** [51.6, 53.8]	**Matched Pareto regime ($R = R_{\text{sem}}$): full semantics, zero slack (800 B)**
$16/32 = 0.500$	87.40 [84.2, 90.0]	76.20 [72.3, 79.7]	16.80 [13.8, 20.3]	61.11 [60.1, 62.1]	Slight over-ranking (4 slack dims): covert leak begins (1,056 B)
$24/32 = 0.750$	86.40 [83.1, 89.1]	74.20 [70.2, 77.8]	95.00 [92.7, 96.6]	84.63 [84.0, 85.3]	Over-ranked (12 slack dims): high subliminal survival (1,568 B)
$32/32 = 1.000$	83.40 [79.9, 86.4]	69.60 [65.4, 73.5]	99.60 [98.6, 99.9]	93.96 [93.4, 94.5]	Full ambient space ($R = d$): approaches unconstrained LatentMAS (2,080 B)

6 Failure Regimes: Superposition Entanglement, Learned Obfuscation, and Cross-Model Drift

To maintain strict epistemic balance, we highlight four counter-evidence findings from 2025–2026 research that challenge both continuous latent reasoning and linear subspace auditing:

1. **Non-causal post-hoc rationalization and learned obfuscation (Haskins et al. [11], Young [26], Zhang et al. [28]).** Zhang et al. [28] performed causal and adversarial interventions on chain-of-continuous-thought models and discovered that latent tokens frequently act as non-causal computation buffers rather than faithful semantic steps. Furthermore, Haskins et al. [11] and Baker et al. [1] proved that if a reinforcement-learning policy is directly penalized whenever its linear factor readout $c_t = U_R^\top h_t$ triggers a safety monitor, the policy learns to rotate its deceptive computation into benign-looking factor combinations (learned obfuscation). Consequently, U_R must be used as a *read-only passive audit probe and hard halting gate* rather than a differentiable RL reward penalty!
2. **Nonlinear feature superposition resists purely linear erasure (Belrose et al. [2], Rizvi-Martel et al. [18]).** While LEACE [2] and INLP [17] provably erase all *linear* information about orthogonal concepts, transformer hidden states encode more features than ambient dimensions via non-linear polysemantic superposition [18]. A colluding receiver equipped with a multi-layer non-linear MLP decoder can extract residual covert bits hidden in curved manifolds inside P_{sem} unless the active-warden dither σ_w is sufficiently large or paired with a Sparse Autoencoder (SAE) bottleneck.
3. **Audit blind spots on subtle multi-turn sabotage (Chen et al. [4], Ferreira et al. [6]).** Ferreira et al. [6] demonstrated that even when human and automated auditors have full access to explicit text chains of thought, they miss subtle multi-turn sabotage strategies where each individual step appears locally benign. Projecting continuous vectors onto U_R restores parity with explicit CoT monitoring but does not magically solve strategic deception that mimics legitimate task subgoals.
4. **Cross-architecture C2C projection distortion (Fu et al. [8]).** When agents run on different foundation model families (e.g., Llama-3.1-8B collaborating with Qwen2.5-7B), their residual stream bases are incompatible. While C2C [8] trains pairwise projection fusers, cascading a C2C fuser with our P_{sem} bottleneck introduces an additional 2%–5% cross-model approximation loss compared to homogeneous-backbone LatentMAS [29].

7 Architectural Scope and Homogeneous-Backbone Constraints

- **Labeled probe calibration requirement.** Constructing the orthonormal factor basis U_R requires an offline calibration pass over 1,000–5,000 domain reasoning traces [2, 17]. Truly novel zero-shot concepts whose directions lie entirely outside $\text{span}(U_R)$ are projected away until U_R is expanded or communicated via the explicit anchor token channel w_m.
- **Numerical precision versus warden dither (σ_w).** On tasks requiring exact multi-digit floating-point or cryptographic hash transmission between agents, active-warden dither $\sigma_w \geq 0.14$ perturbs least-significant continuous coordinates; exact literals must therefore be routed through the explicit Codetta delimiter tokens [16].
- **White-box bus placement.** Our architecture requires the communication bus to intercept hidden vectors h_t between agent forward passes, which applies to API-mediated or federated

agent platforms but cannot constrain two colluding agents communicating over an external covert network socket.

8 Pre-Registerable Falsification Protocol for Live Multi-Agent Latent Exchanges

Table 6 specifies six pre-registerable empirical protocols designed to falsify the core claims of Audited Continuous Latent Collaboration on open-weight multi-agent systems.

Table 6: Six pre-registerable falsification tests for Audited Continuous Latent Collaboration.

Protocol	Hypothesis	Design and primary metrics	Falsified if
P1. LatentMAS efficiency & accuracy parity ($C = 2$)	Orthogonal Factor Projection ($R/d = 0.375$) retains within 2.0% of unconstrained LatentMAS accuracy [29] while cutting tokens $\geq 75\%$ vs. text MAS	Evaluate 3-agent Llama-3.1-8B teams on GSM8K, MATH, and ARC-Challenge under Explicit Text, LatentMAS, and Factor Projection	Factor Projection accuracy drops > 2.5 percentage points below Explicit Text MAS at $C = 2$
P2. Null-space steganography suppression (Eq. 2)	Projecting onto $P_{\text{sem}} +$ warden dither σ_w reduces covert payload decoding from $> 85\%$ to $< 5\%$ above random guessing (50% bit / 0% packet)	Train a colluding sender–receiver LoRA pair [15, 24] to transmit 16-bit covert keys across 500 episodes	Receiver decodes full 16-bit covert payloads on $> 5.0\%$ of episodes at $R \leq R_{\text{sem}}$
P3. Rank-ratio rebound verification (Table 5)	Increasing R/d above 0.50 causes covert payload survival to rise monotonically as $R - R_{\text{sem}}$ grows	Sweep projection rank $R/d \in \{0.125, 0.25, 0.375, 0.50, 0.75, 1.0\}$ against trained steganographic decoders	Covert survival fails to increase by at least 30 percentage points between $R/d = 0.375$ and $R/d = 0.75$
P4. LT-Tuning / Codetta drift arrest at $C = 8$ (Eq. 3)	Interleaving 20% explicit anchor tokens [14, 16] improves $C = 8$ multi-turn accuracy by ≥ 15 points over unanchored LatentMAS [19]	Run 6-turn latent collaboration at compression $C = 8$ with and without sparse vocabulary anchor tokens	Anchored accuracy fails to exceed unanchored LatentMAS by at least 10.0 percentage points at $C = 8$
P5. Linear factor monitorability vs. CoT audits	Linear probes on $c_t = U_R^\top \tilde{h}_t$ detect planted Sabotage/Deception episodes at parity with explicit CoT monitors [1, 6]	Inject 200 hidden side-task sabotage goals into multi-agent coding episodes; compare AUROC of text CoT monitor vs. factor probe c_t	Factor probe AUROC trails explicit text CoT monitor AUROC by > 0.05 at $R/d = 0.375$
P6. Non-differentiable audit resistance to obfuscation	Keeping U_R as a passive halting gate rather than an RL penalty prevents learned factor obfuscation [11]	Compare RL fine-tuning with vs. without direct reward penalties on c_t; measure causal probe faithfulness [26]	Passive probe faithfulness degrades by $> 5.0\%$ when U_R is excluded from the RL loss gradient

9 Conclusion

Continuous latent collaboration frameworks such as LatentMAS and Cache-to-Cache represent a genuine step-change in multi-agent sequence-step efficiency and parallel superposition reasoning, outperforming discretized explicit text at moderate compression while cutting sequence steps by 75%–88%. Yet deploying unconstrained continuous vector channels in open or safety-critical collectives is untenable: raw hidden-state exchange inflates wire byte payloads unless rank-reduced, degrades linear Chain-of-Thought monitorability (51.40% at $C = 4$), opens a high-capacity $(d - R)$-dimensional null-space channel for steganographic collusion (87.60% packet survival and 82.23% bit accuracy at $C = 4$), and degrades under multi-turn representational drift at higher compression (24.20% accuracy at $C = 4$ and 4.40% at $C = 8$). By projecting continuous latent messages onto a calibrated rank-R orthogonal factor subspace ($R/d = 0.375$), applying LEACE/INLP null-space scrubbing and active-warden Gaussian dither, and interleaving sparse LT-Tuning/Codetta vocabulary anchors, Audited Continuous Latent Collaboration reduces wire bytes by 68.75% vs. uncompressed text (84.38% vs. Unconstrained LatentMAS at $C = 4$), sustains 77.40% [73.53, 80.85] task accuracy and 75.40% [71.44, 78.97] linear monitorability at $C = 4$, and suppresses covert steganographic packet survival to 8.00% [5.93, 10.71] (52.88% bit accuracy). Controlling the subspace rank ratio R/d provides a mathematically grounded dial for balancing multi-agent wire bandwidth, reasoning fidelity, and linear safety oversight.

Use of AI tools

Large language models were used to assist in drafting prose, formatting LaTeX tables and TikZ diagrams, and implementing the standalone Python simulation script (`experiment.py`). Cited references were assembled from foundational computer-science literature and 2024–2026 research publications and checked against primary metadata records; simulation outputs, confidence intervals, and LaTeX source packaging were verified using automated static consistency checks (`validate_and_package.py`).

References

[1] Bowen Baker, Joost Huizinga, Leo Gao, Zehao Dou, Melody Y. Guan, Aleksander Madry, et al. Monitoring reasoning models for misbehavior and the risks of promoting obfuscation, 2025. URL https://arxiv.org/abs/2503.11926.

[2] Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. Advances in Neural Information Processing Systems (NeurIPS), 2023. URL https://arxiv.org/abs/2306.03819.

[3] Mert Cemri, Melissa Z. Pan, Shuyi Yang, Lakshya A. Agrawal, Bhavya Chopra, Rishabh Tiwari, et al. Why do multi-agent LLM systems fail? In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2025. URL https://arxiv.org/abs/2503.13657.

[4] Yanda Chen, Joe Benton, Ansh Radhakrishnan, Jonathan Uesato, Carson Denison, John Schulman, et al. Reasoning models don't always say what they think, 2025. URL https://arxiv.org/abs/2505.05410.

[5] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. URL https://arxiv.org/abs/2305.14325.

[6] Pedro Ferreira, Wilker Aziz, and Ivan Titov. Does out-of-sight equal out-of-mind in CoT monitorability?, 2026. URL https://arxiv.org/abs/2608.04928.

[7] Gina Fisk, Mike Fisk, Christos Papadopoulos, and Joshua Neil. Eliminating steganography in internet traffic with active wardens. International Workshop on Information Hiding (IH), LNCS 2578, pages 18–35, 2003. URL https://doi.org/10.1007/3-540-36415-3_2.

[8] Tianyu Fu, Zihan Min, Hanling Zhang, Jichao Yan, Guohao Dai, Wanli Ouyang, and Yu Wang. Cache-to-cache: Direct semantic communication between large language models. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/474ada926b331d78f06d95e8913111cc-Abstract-Conference.html.

[9] Melody Guan, Miles Wang, Micah Carroll, Zehao Dou, Annie Wei, Marcus Williams, et al. Monitoring monitorability. In *International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2512.18311.

[10] Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. Training large language models to reason in a continuous latent space, 2024. URL https://arxiv.org/abs/2412.06769.

[11] Reilly Haskins, Bilal Chughtai, and Joshua Engels. Training on documents about monitoring leads to CoT obfuscation, 2026. URL https://arxiv.org/abs/2605.15257.

[12] Gabriel Kaptchuk, Tushar M. Jois, Matthew Green, and Aviel D. Rubin. Meteor: Cryptographically secure steganography for realistic distributions. ACM SIGSAC Conference on Computer and Communications Security (CCS), pages 1529–1544, 2021. URL https://doi.org/10.1145/3460120.3484550.

[13] Tomek Korbak, Mikita Balesni, Elizabeth Barnes, Yoshua Bengio, Joe Benton, Joseph Bloom, et al. Chain of thought monitorability: A new and fragile opportunity for AI safety, 2025. URL https://arxiv.org/abs/2507.11473.

[14] Weihao Liu, Dehai Min, and Lu Cheng. Latent thoughts tuning: Bridging context and reasoning with fused information in latent tokens. In *International Conference on Machine Learning (ICML)*, volume 306 of *PMLR*, 2026. URL https://arxiv.org/abs/2602.10229.

[15] Sumeet Ramesh Motwani, Mikhail Baranchuk, Martin Strohmeier, Vijay Bolina, Philip H. S. Torr, Lewis Hammond, and Christian Schroeder de Witt. Secret collusion among AI agents: Multi-agent deception via steganography. Advances in Neural Information Processing Systems (NeurIPS), 2024. URL https://arxiv.org/abs/2402.07510.

[16] Qi Pang, Virginia Smith, and Wenting Zheng. Codetta: High-capacity, keyless, and undetectable multi-agent collusion. IACR ePrint 2026/2218 / arXiv:2609.28900, 2026. URL https://arxiv.org/abs/2609.28900.

[17] Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection (INLP). Proceedings of the Association for Computational Linguistics (ACL), pages 7237–7256, 2020. URL https://aclanthology.org/2020.acl-main.647/.

[18] Michael Rizvi-Martel, Guillaume Rabusseau, and Marius Mosbach. The illusion of superposition? a principled analysis of latent thinking in language models, 2026. URL https://arxiv.org/abs/2604.06374.

[19] Carson Rodrigues. Hallucination as context drift: Synchronization protocols for multi-agent LLM systems. arXiv preprint arXiv:2606.21666, 2026. URL https://arxiv.org/abs/2606.21666.

[20] Fabien Roger and Ryan Greenblatt. Preventing language models from hiding their reasoning. arXiv preprint arXiv:2310.18512, 2023. URL https://arxiv.org/abs/2310.18512.

[21] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948. URL https://doi.org/10.1002/j.1538-7305.1948.tb01338.x.

[22] Zhenyi Shen, Hanqi Yan, Linhai Zhang, Zhanghao Hu, Yali Du, and Yulan He. CODI: Compressing chain-of-thought into continuous space via self-distillation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 677–693, 2025. URL https://arxiv.org/abs/2502.21074.

[23] Gustavus J. Simmons. The prisoners' problem and the subliminal channel. In *Advances in Cryptology (CRYPTO '83)*, pages 51–67, 1984. URL https://doi.org/10.1007/978-1-4684-4730-9_5.

[24] Om Tailor. Audit the whisper: Detecting steganographic collusion in multi-agent LLMs. arXiv preprint arXiv:2510.04303, 2025. URL https://arxiv.org/abs/2510.04303.

[25] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. arXiv preprint arXiv:2308.08155, 2023. URL https://arxiv.org/abs/2308.08155.

[26] Richard J. Young. Lie to me: How faithful is chain-of-thought reasoning in open-weight reasoning models?, 2026. URL https://arxiv.org/abs/2603.22582.

[27] Guibin Zhang, Yanwei Yue, Zhixun Li, Sukwon Yun, Guancheng Wan, Kun Wang, et al. Cut the crap: An economical communication pipeline for LLM-based multi-agent systems. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/bbc461518c59a2a8d64e70e2c38c4a0e-Abstract-Conference.html.

[28] Yuyi Zhang, Boyu Tang, Tianjie Ju, Sufeng Duan, and Gongshen Liu. Do latent tokens think? a causal and adversarial analysis of chain-of-continuous-thought, 2025. URL https://arxiv.org/abs/2512.21711.

[29] Jiaru Zou, Ruizhong Qiu, Gaotang Li, Xiyuan Yang, Katherine Tieu, Pan Lu, et al. Latent collaboration in multi-agent systems. In *International Conference on Machine Learning (ICML)*, 2026. URL https://arxiv.org/abs/2511.20639.