What Do Latent Agents Actually Represent? Interpreting Hidden-State Communication in Multi-Agent Systems

Wenhao Lu [1] **Spursh Deshpande** [2]

Abstract

Latent multi-agent systems replace text communication between agents with direct hidden-state transfer, promising richer inter-agent collaboration. Yet what information actually flows through these latent channels remains poorly understood. We present a systematic mechanistic analysis of a four-agent system evaluating on mathematical reasoning, science multiple-choice, and code generation benchmarks. We find that task identity is nearly perfectly linearly separable in the latent space, that individual agent roles are recoverable with near-perfect accuracy from hidden states alone, and that the vast majority of the representation space is effectively unused. Correctness information peaks at mid-network layers, yet this signal exists equally in single-agent baselines, revealing it as a base-model property rather than an emergent collaborative signal. Agents are highly informationally redundant with one another despite exhibiting distinct role signatures. Taken together, our results suggest that the apparent richness of latent communication reflects base-model geometry rather than learned compositional coordination.

1. Introduction

Multi-agent systems have long promised benefits from decomposing complex tasks among specialised agents (Wooldridge and Jennings, 1995). Recent work has explored replacing natural language communication between LLM agents (Du et al., 2023; Liang et al., 2023) with direct transfer of latent hidden states (Zou et al., 2025), motivated by the hypothesis that continuous representations carry richer inter-agent signals than discrete tokens. In such *latent multi-agent systems* (LatentMAS), a learned alignment matrix $\mathbf{W}_a$ transforms hidden states from one agent into the key-value cache of a receiving agent.

Despite growing interest in latent-space agent communication (Zou et al., 2025; Hong et al., 2023), fundamental questions about what these systems actually represent remain open. Does the latent channel encode task-relevant abstractions that go beyond what a single model already knows? Do agents develop complementary representations that support genuine compositionality? Or does the apparent structure reflect properties of the underlying language model, present before any inter-agent communication takes place?

We address these questions through a battery of mechanistic experiments on a four-agent LatentMAS (Planner, Critic, Refiner, Judger) built on Qwen3-4B (Qwen Team, 2025), operating on three diverse benchmarks: GSM8K (Cobbe et al., 2021) for mathematical reasoning, ARC-Challenge (Clark et al., 2018) for science multiple-choice, and MBPP+ (Austin et al., 2021) for code generation. Our analysis spans more than a thousand examples across thirteen experimental conditions.

Our main contributions are:

- We show that task-domain geometry in LatentMAS achieves near-perfect linear separability, but this property is equally present in single-agent baselines, ruling out collaboration as the source.
- We demonstrate that agent roles are highly identifiable from hidden states alone while agents remain highly informationally redundant, revealing a tension between structural differentiation and semantic overlap.
- We map the layer-by-layer emergence of correctness information, showing it peaks at mid-network layers and is diffuse across dimensions rather than concentrated in a learned direction.
- We show that the vast majority of representation dimensions are inactive, with effective communication concentrated in a small number of principal directions.

[1]Millburn High School, Millburn, NJ, USA [2]Crescenta Valley High School, La Crescenta, CA, USA. Correspondence to: Wenhao Lu <wenhao.lu.2020@gmail.com>.

Accepted to the 2nd Workshop on Compositional Learning at ICML 2026, Seoul, South Korea. *Copyright 2026 by the author(s).*

- We characterise round-to-round information propagation, finding that mutual information decays rapidly and is higher for correct examples than incorrect ones.

2. Background and Related Work

Multi-Agent LLM Systems. Frameworks such as AutoGen (Wu et al., 2023), MetaGPT (Hong et al., 2023), and multi-agent debate (Du et al., 2023) arrange LLM instances as communicating agents that tackle tasks through structured interaction. Communication typically occurs via text. Du et al. (Du et al., 2023) show that multi-agent debate can improve factual accuracy over single-model baselines; Liang et al. (Liang et al., 2023) find that encouraging divergent thinking improves downstream reasoning. The common thread is that multiple agents checking each other's work should produce more reliable outputs than any single agent alone.

Latent Communication. Zou et al. (Zou et al., 2025) propose training-free latent-space communication between agents, arguing that continuous representations carry more information than discrete token sequences. Related ideas appear in neural module networks and multi-agent reinforcement learning with learned communication. Our work is the first systematic mechanistic analysis of what latent LLM communication channels actually encode.

Mechanistic Interpretability. Probing classifiers and the linear representation hypothesis have revealed rich geometric structure in transformer hidden states. Logit lens and activation patching have been used to localise factual information within language models. We apply this toolkit to multi-agent representations, asking whether inter-agent communication produces structure beyond what the base model already encodes.

Compositionality in Neural Networks. Whether neural networks learn compositional representations is a central open question. Lake et al. argue that human-like generalisation requires compositional structure; Hupkes et al. decompose compositionality into systematicity, productivity, and substitutivity; Dziri et al. show that transformers struggle with compositional tasks requiring structured reasoning. Our work examines compositionality through the lens of how agents internally represent and partition task knowledge, asking whether the multi-agent architecture produces the kind of modular, complementary representations that compositional learning requires.

Redundancy and Modularity. Mixture-of-experts models achieve specialisation through complementary rather than redundant experts. We measure redundancy at the agent level and ask whether LatentMAS achieves the complementarity that modular compositional architectures require.

3. System Description

LatentMAS consists of four agents instantiated from the same base model (Qwen3-4B (Qwen Team, 2025)), each given a distinct system prompt encoding its role: Planner (A0), Critic (A1), Refiner (A2), and Judger (A3). Let $h_t^{(i)} \in \mathbb{R}^D$ denote the post-generation hidden state of agent i at round t, where $D = 2560$. Communication from agent i to agent j proceeds via:

$$\tilde{h}_t^{(j)} = \mathbf{W}_a \, h_t^{(i)}, \tag{1}$$

where $\mathbf{W}_a \in \mathbb{R}^{D \times D}$ is a learned alignment matrix. The transformed vector $\tilde{h}_t^{(j)}$ is injected into the key-value cache of agent j before its next generation step, allowing agent j to condition its generation on the latent state of agent i without any explicit token-level communication.

Agents operate in sequence over up to four communication rounds. The Planner generates an initial response, the Critic evaluates it, the Refiner proposes improvements, and the Judger produces the final answer. At each round, each agent's hidden state is recorded immediately after generation, yielding a collection of $3 \times 4 = 12$ (agent, round) positions per example (excluding the Judger, whose state is not transmitted to a subsequent agent).

4. Experimental Setup

Data. We evaluate on GSM8K (Cobbe et al., 2021) (n=500), ARC-Challenge (Clark et al., 2018) (n=500), and MBPP+ (Austin et al., 2021) (n=378), yielding a total of 1,378 examples per condition. All experiments use the same random seed and question ordering across conditions to ensure paired comparability.

Baseline conditions. We compare against: single-agent latent with sampled decoding, single-agent latent with greedy decoding, text-based MAS (TextMAS), KV-blocked (no hidden-state injection), no-transfer (agents run independently without cross-agent hidden states), identity $\mathbf{W}_a$, and spectrum-matched random $\mathbf{W}_a$. Full condition list and hyperparameters are provided in Appendix D.

Analysis methods. We use linear SVM and logistic regression probes, the Kraskov kNN mutual information estimator, PCA-based subspace analysis, CKA, and ridge regression for cross-agent R^2 computation. All probe evaluations use stratified 5-fold cross-validation. Statistical significance is assessed via McNemar's test for accuracy comparisons and permutation tests for probe AUC values.

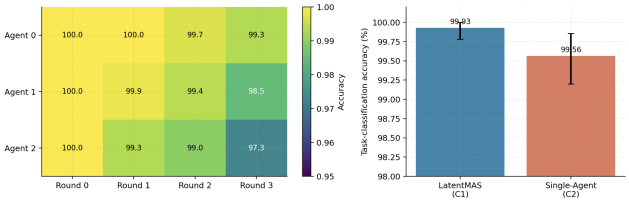

Figure 1. Task-domain linear separability. Left: SVM confusion matrix for LatentMAS hidden states (99.9% accuracy across three tasks). Right: Per-(agent, round) accuracy confirming all positions achieve $\geq 99\%$, identical to single-agent baselines.

5. Results

5.1. Task-Domain Geometry Is Dominated by the Base Model

A linear SVM trained to classify the task domain (GSM8K vs. ARC vs. MBPP+) from post-$\mathbf{W}_a$ latent hidden states achieves **99.9% accuracy** (5-fold CV, n=1,378). This result holds at every (agent, round) position, with all positions achieving between 99% and 100% accuracy. Crucially, the same classifier trained on single-agent hidden states achieves 99.6%.

The fact that single-agent and multi-agent hidden states are equally task-separable demonstrates that the high separability is a property of the base model's geometry, present from the very first forward pass, and not a product of the inter-agent communication protocol. The alignment matrix $\mathbf{W}_a$ and the multi-round communication structure neither create nor destroy task-domain separability.

Figure 1 shows the SVM confusion matrix and per-position accuracy heatmap. The absence of any off-diagonal entries in the confusion matrix is notable: the model never confuses a GSM8K hidden state for an ARC or MBPP+ state, or vice versa.

5.2. Correctness Encoding Is a Base-Model Property

We train logistic regression probes at each (agent, round) position to predict whether the final answer will be correct, based solely on the hidden state at that position. AUC values for LatentMAS range from 0.657 to 0.708 across agents and rounds (Table 1), with the highest AUC at agent 2 (Refiner), round 0.

Single-agent baselines show a comparable AUC range of 0.67 to 0.71 across 12 rounds with no significant trend. The absence of any improvement in probe AUC from the single-agent to the multi-agent condition indicates that correctness predictability does not increase with inter-agent communication. Whatever signal enables correctness prediction is present in the base model's representations independently of collaboration.

Table 1. Per-(agent, round) AUC of logistic regression probes predicting final answer correctness. Single-agent range (0.67 to 0.71) matches LatentMAS, indicating correctness encoding is a base-model property.

	Round			
Agent	0	1	2	3
Planner (A0)	0.682	0.657	0.695	0.671
Critic (A1)	0.685	0.658	0.691	0.658
Refiner (A2)	0.708	0.681	0.691	0.672

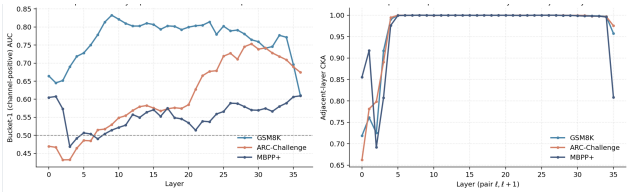

Figure 2. Per-layer correctness probe AUC on GSM8K (n=200). AUC rises progressively through the network, peaking at layer 9 (AUC = 0.833). The correctness signal crystallises in mid-to-late transformer layers. Full per-layer values in Appendix B.

Rate-distortion analysis provides a complementary picture: probe AUC stays approximately flat at 0.52 to 0.55 regardless of how many top principal components are used as input features (1 to 499 dimensions). This flatness indicates that correctness information is diffuse across many dimensions rather than concentrated in a learned low-dimensional direction. There is no "correctness subspace" that can be extracted via PCA.

5.3. Correctness Information Peaks at Mid-Network Layers

We train per-layer correctness probes on GSM8K (n=200, L=37 layers). As shown in Figure 2, AUC rises from 0.664 at layer 0 to a peak of **0.833 at layer 9**, then remains elevated at 0.80 to 0.82 through layers 10 and 11 before gradually declining. Early layers (0 to 2) have lower and sometimes non-significant AUC.

This result has direct implications for the LatentMAS communication mechanism. Hidden-state transfer via $\mathbf{W}_a$ injects the *final-layer* hidden state of one agent into the KV cache of the next. Our layer-wise analysis suggests that final-layer representations contain less correctness signal (AUC $\approx$ 0.77) than mid-layer representations (AUC $\approx$ 0.83 at layers 8 to 9). The communication protocol may therefore be transmitting a less informative representation than an alternative design that extracted mid-layer states.

5.4. Role Specialisation Is Structural, Not Semantic

Agent identity. A 3-way logistic regression classifier trained to identify which agent (A0, A1, or A2) produced

Table 2. Within-agent vs. cross-agent cosine similarity and agent-ID classifier accuracy. Agents are role-identifiable but geometrically similar, with high within-agent consistency across rounds.

Task	Within-agent	Cross-agent	ID Acc.
GSM8K	0.9999	0.8495	98.9%
ARC	0.9999	0.8632	99.5%
MBPP+	0.9999	0.8677	98.9%

a given hidden state achieves 98.9% accuracy on GSM8K, 99.5% on ARC, and 98.9% on MBPP+ (chance = 33%). Role identity crystallises in the upper half of the network: classifier accuracy rises from approximately 0.96 at early transformer layers to approximately 0.99 at late layers, indicating that role-specific representations emerge through depth.

Within- vs. cross-agent similarity. Within-agent cosine similarity is approximately 1.0000 across all rounds and all tasks, meaning that each agent's hidden state is essentially identical from round to round. This extreme within-agent stability implies that agents are not updating their representations in response to the latent information they receive. Cross-agent similarity is 0.85 to 0.87 across tasks (Table 2), giving a within/cross ratio of 1.15 to 1.18. While this ratio confirms that agents are geometrically distinguishable, the absolute cross-agent similarity is high: agents are far more similar to each other than one would expect from a compositionally specialised system.

Anchoring dynamics. The Planner's influence on downstream agents varies by task. On GSM8K, anchoring decays sharply: $\cos(\text{A0}, \text{A2}) = 0.527$, $\cos(\text{A0}, \text{A3}) = 0.256$, indicating that the Planner's initial representation is largely overwritten by later agents. On ARC-Challenge, anchoring persists strongly: $\cos(\text{A0}, \text{A3}) = 0.543$, suggesting the Planner's framing of the problem survives through the full pipeline. MBPP+ shows intermediate behaviour. No significant mirror-neuron effects are observed ($p > 0.05$ in all conditions), ruling out simple copying between agents.

5.5. Most of the Representation Space Is Inactive

PCA on all collected latent hidden states (n=16,536 vectors, D=2560) places the elbow at k=69 components, which together explain 86.9% of total variance. The top component alone accounts for 26.9% of variance, with a rapid drop-off: the top 10 components explain 68.4% combined.

The **dead dimension fraction** is 93.6%: 2,398 of 2,560 dimensions have near-zero variance ($< 10^{-4}$ times the maximum). Active communication occurs in roughly 162 dimensions. Despite this concentration, the predictive subspace analysis (Method B) identifies zero dimensions that are above threshold for predicting collaboration outcomes,

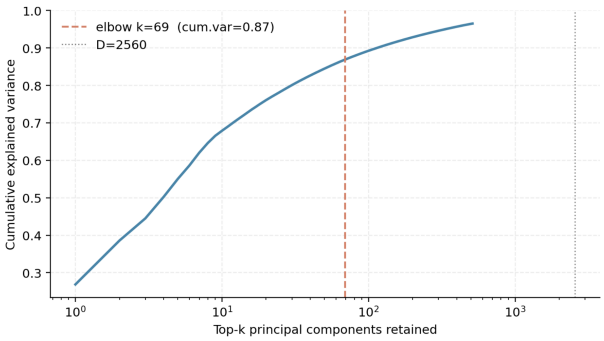

Figure 3. Latent communication subspace. Left: Cumulative explained variance vs. PCA components – 87% captured by 69 dimensions out of 2560. Right: Dimension activity distribution – 93.6% of dimensions have near-zero variance.

Table 3. Round-to-round mutual information (nats) by task and correctness subset. MBPP+ shows substantially higher and more persistent MI than math/science tasks.

Task	Subset	k=1	k=2	k=3
GSM8K	All	0.150	0.060	0.000
	Correct	0.137	0.057	0.000
	Incorrect	0.002	0.000	0.000
ARC	All	0.096	0.027	0.001
MBPP+	All	0.322	0.225	0.001
	Incorrect	0.167	0.119	0.026

suggesting that the active subspace reflects general model geometry rather than task-specific or coordination-specific structure.

Figure 3 shows the PCA variance spectrum and dead-dimension distribution. The contrast between the 69-component elbow and the full 2560-dimensional nominal capacity of the hidden state is stark: only 2.7% of the representational capacity is meaningfully used.

5.6. Information Decays Rapidly Across Rounds

We estimate mutual information between hidden states at different communication rounds using the Kraskov kNN estimator on 16-dimensional PCA projections. Results are summarised in Table 3.

For GSM8K and ARC, MI is near zero at lag k=2, indicating that information from two rounds earlier is essentially absent. MBPP+ shows substantially higher and more persistent MI (0.225 nats at k=2), consistent with code generation's higher collaborative benefit. Correct examples maintain consistently higher inter-round MI than incorrect examples across all tasks: when the system is heading toward a correct answer, there is more persistent structure across the communication rounds. Incorrect trajectories show near-zero MI even at lag k=1 for GSM8K and ARC.

Table 4. Cross-agent R^2 matrices. Entry (i, j) is the R^2 of a linear regression predicting agent j's hidden state from agent i's. High values indicate high redundancy across all pairs.

GSM8K – Mean $R^2 = 0.752$			
	A0	A1	A2
A0	1.000	0.809	0.775
A1	0.662	1.000	0.765
A2	0.722	0.778	1.000
ARC-Challenge – Mean $R^2 = 0.795$			
	A0	A1	A2
A0	1.000	0.797	0.762
A1	0.746	1.000	0.819
A2	0.799	0.848	1.000

5.7. Agents Are Highly Informationally Redundant

For each ordered pair of agents (i, j), we train a ridge regression predicting $h^{(j)}$ from $h^{(i)}$ (using a 64-dimensional PCA projection) and compute the coefficient of determination R^2. Results are shown in Table 4.

Mean redundancy is 0.752 (GSM8K), 0.795 (ARC), and 0.749 (MBPP+), leaving only 20 to 25% unique information per agent. Critically, redundancy does not predict correctness: Spearman correlations between per-example cross-agent R^2 and final correctness are $r < 0.013$ with $p > 0.15$ on all tasks. High redundancy is a structural property of the system independent of whether it gets the answer right.

6. Discussion

Base-model dominance. Across all experiments, the most consistent finding is that properties attributed to multi-agent collaboration are equally present in single-agent baselines. Task separability (near-perfect in both conditions), correctness encoding (AUC 0.67 to 0.71 in both), and the layer-wise emergence of correctness signal are all properties of the base model's geometry. The multi-agent protocol inherits rather than learns these properties.

Structure without function. Agents are role-identifiable with high accuracy and maintain distinct representational lanes (within-agent cosine ≈ 1.0), yet 75 to 80% of their information overlaps. This creates an important distinction between *structural differentiation* and *semantic complementarity*. The system has learned the former without the latter: agents look different but largely say the same thing. This distinction is central to the compositional learning literature: a truly compositional system requires semantic complementarity, not merely structural diversity.

Implications for the LatentMAS design. Two specific design choices deserve reconsideration in light of our results. First, the alignment matrix $\mathbf{W}_a$ appears to learn a near-trivial mapping (see companion paper for ablation details), which combined with the high cross-agent redundancy suggests that the communication channel is not carrying task-useful information. Second, the protocol communicates final-layer hidden states, but our layer-wise analysis shows correctness information peaks at mid-network layers. Alternative designs that extract and transmit mid-layer representations may achieve better information transfer.

Implications for compositional learning. Genuine compositionality requires sub-components that contribute distinct, reusable representations. Our results – high redundancy, near-perfect within-agent consistency, and inactive representation dimensions – suggest that current LatentMAS architectures do not achieve this. Explicit interventions such as diversity regularisation, bottleneck communication channels, or role-specific pre-training may be necessary to produce the semantic complementarity that compositional multi-agent systems require.

Limitations. This study examines a single model family (Qwen3-4B) and a four-agent configuration. Logit-lens data was unavailable due to a data collection bug (the `_decode_latent_topk` function threw silent exceptions), precluding faithfulness gap analysis . Activation patching data was not fully collected. Future work should examine larger model families, more agent configurations, and alternative communication protocols.

7. Conclusion

We have presented a systematic mechanistic study of a latent multi-agent system spanning more than one thousand examples and multiple analysis methods. The consistent picture is one of base-model dominance: task domains are linearly separable by the model itself, correctness information peaks in mid-network layers identically in single-agent settings, and the vast majority of representation dimensions are inactive. Despite high role identifiability, agents are highly informationally redundant, with cross-agent R^2 of 0.75 to 0.80. Round-to-round mutual information decays rapidly except for code generation tasks, where collaboration provides more genuine benefit.

These results challenge the assumption that latent communication enables richer compositional coordination than text-based alternatives, and motivate architectural interventions that actively enforce representational complementarity between agents. The gap between structural role differentiation and semantic complementarity identified here provides a concrete target for future work on compositional multi-agent architectures.

Impact Statement

This paper advances understanding of multi-agent AI systems and interpretability methods. Improved transparency of latent communication mechanisms supports safer deployment of multi-agent systems. There are no immediate negative societal impacts we feel must be highlighted.

References

Jacob Austin et al. Program synthesis with large language models. arXiv preprint arXiv:2108.07732, 2021.

Peter Clark et al. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. arXiv preprint arXiv:1803.05457, 2018.

Karl Cobbe et al. Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168, 2021.

Yilun Du et al. Improving factuality and reasoning in language models through multiagent debate. arXiv preprint arXiv:2305.14325, 2023.

Sirui Hong et al. MetaGPT: Meta programming for a multi-agent collaborative framework. arXiv preprint arXiv:2308.00352, 2023.

Tian Liang et al. Encouraging divergent thinking in large language models through multi-agent debate. arXiv preprint arXiv:2305.19118, 2023.

Qwen Team. Qwen3 technical report. arXiv preprint arXiv:2505.09388, 2025.

Michael Wooldridge and Nicholas R. Jennings. Intelligent agents: Theory and practice. *The Knowledge Engineering Review*, 10(2):115–152, 1995.

Qingyun Wu et al. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. arXiv preprint arXiv:2308.08155, 2023.

J. Zou et al. Latent collaboration: Training-free latent-space communication for multi-agent systems. arXiv preprint arXiv:2511.20639, 2025.

A. Full Accuracy Table

Table 5. Headline accuracy across all conditions and tasks. Entries marked with a sample size used subsets; all others used the full evaluation set.

Condition	GSM8K	ARC-Challenge	MBPP+
latent_mas	92.0% (n=500)	92.6% (n=500)	73.3% (n=378)
single_agent_latent_sampled	90.8%	92.8%	66.9%
single_agent_latent_greedy	90.5% (n=200)	92.5% (n=200)	65.5% (n=200)
text_mas	91.0% (n=200)	90.5% (n=200)	89.0% (n=200)
kv_blocked	89.5% (n=200)	89.0% (n=200)	69.0% (n=200)
no_transfer	90.8%	93.2%	67.5%
latent_mas_random_wa_spectrum	91.5% (n=200)	92.0% (n=200)	77.5% (n=200)
exp_m_identity_wa	91.0% (n=100)	91.0% (n=100)	74.0% (n=100)

B. Full Per-Layer AUC

Table 6. Per-layer AUC of logistic regression predicting correctness from individual transformer layer hidden states (GSM8K, $n = 200$, $L = 37$).

Layer	AUC	p-value
0	0.664	0.017
3	0.690	0.004
4	0.719	<0.001
5	0.728	<0.001
6	0.750	<0.001
7	0.778	<0.001
8	0.813	<0.001
9	0.833	<0.001
10	0.821	<0.001
11	0.810	<0.001

C. Full Mutual Information Results

D. Experimental Conditions

We evaluate thirteen distinct conditions: `latent_mas`, `single_agent_latent_sampled`, `single_agent_latent_greedy`, `text_mas`, `kv_blocked`, `no_transfer`, `latent_mas_random_wa_spectrum`, and `exp_m_identity_wa`, with additional ablation variants for activation patching across 25 examples per task. All conditions use Qwen/Qwen3-4B with temperature 0.7 and top-p 0.9 (temperature 0 for greedy decoding).

Table 7. Round-to-round mutual information (nats), Kraskov kNN estimator, 16-dim PCA projection. Split by correctness.

Task	Subset	$k = 1$	$k = 2$	$k = 3$
GSM8K	All	0.150	0.060	0.000
	Correct	0.137	0.057	0.000
	Incorrect	0.002	0.000	0.000
ARC	All	0.096	0.027	0.001
	Correct	0.106	0.014	0.001
	Incorrect	0.003	0.000	0.000
MBPP+	All	0.322	0.225	0.001
	Correct	0.271	0.131	0.000
	Incorrect	0.167	0.119	0.026