
Compression and Interpretability in Multi-Agent Latent Communication

August Andersen
Yale University
august.andersen@yale.edu

Abstract

Recent work has demonstrated that multi-agent LLM systems achieve $4\text{--}24\times$ inference speedups by communicating via compressed latent representations rather than natural language [Zou et al., 2025, Du et al., 2025]. Prior interpretability research has applied logit lens analysis to single-agent latent reasoning [Dilgren and Wiegrefe, 2026] and probing methods to multi-agent natural-language communication [Rose et al., 2026]. However, to our knowledge, no published work examines interpretability of inter-agent latent communication, particularly under variable compression. We address this gap through an empirical measurement of the compression-interpretability tradeoff¹ in multi-agent latent communication. Using LatentMAS-style multi-agent latent communication on Qwen3-8B on the GSM8K-Platinum benchmark, we apply logit lens decoding to the inter-agent latent messages. We sweep compression levels from 1 to 80 latent steps and measure how natural-language recoverability of message content changes with compression. *Empirical results forthcoming*. These findings characterize the relationship between compression level and interpretability in multi-agent systems using latent communication and raise broader questions about efficiency and interpretability in agentic AI.

References

- Dilgren, A. and Wiegrefe, S. (2026). Are latent reasoning models easily interpretable? arXiv preprint arXiv:2604.04902. <https://arxiv.org/abs/2604.04902>
- Du, Z., Wang, R., Bai, H., Cao, Z., Zhu, X., Cheng, Y., Zheng, B., Chen, W., and Ying, H. (2025). Enabling agents to communicate entirely in latent space. arXiv preprint arXiv:2511.09149. <https://arxiv.org/abs/2511.09149>
- Rose, A., Cullen, C., Kaplowitz, B., and Schroeder de Witt, C. (2026). Detecting multi-agent collusion through multi-agent interpretability. arXiv preprint arXiv:2604.01151. <https://arxiv.org/abs/2604.01151>
- Zou, J., Yang, X., Qiu, R., Li, G., Tieu, K., Lu, P., Shen, K., Tong, H., Choi, Y., He, J., et al. (2025). Latent collaboration in multi-agent systems. arXiv preprint arXiv:2511.20639. <https://arxiv.org/abs/2511.20639>

¹We use the term “tradeoff” for semantic clarity; whether such a relationship exists, and what shape it takes, has not been proven empirically and is the question this paper aims to answer.