Minimal Signaling for Mediated Coordination in Multi-Agent LLM Systems

Medhansh Khattar
Rose-Hulman Institute of Technology

1 Introduction

Large Language Models (LLMs) are increasingly deployed in multi-agent settings where multiple models collaborate, critique, or coordinate to solve complex tasks. Recent work demonstrates that LLM-based multi-agent systems (MAAS) can exhibit structured role-play [3], collective reasoning through debate or self-consistency [4], and evaluative behavior when acting as judges for other models [5]. As model specialization increases and smaller domain-specific models become more common, such multi-agent configurations are expected to play an important role in the development of scalable AI systems.

Despite this progress, communication remains a central bottleneck. Nearly all existing LLM-to-LLM interaction relies on unrestricted natural language. Although expressive, natural language is verbose, ambiguous, and computationally expensive. Long messages consume limited context window capacity, increase inference cost, and introduce stylistic variation that leads to inconsistent interpretation across different models such as GPT, Claude, Gemini, and Llama. These limitations are amplified in resource-constrained environments, including edge devices, financial analysis pipelines, and cross-model interoperability scenarios.

Several techniques attempt to improve communication efficiency through summarization heuristics, message pruning, or hierarchical condensation. However, these methods are brittle, non-adaptive and may remove subtle but essential information. Long-context or attention-efficient architectures [1] improve input handling but do not address the underlying issue that free-form natural language is not an optimal medium for agent-to-agent communication. Classical multi-agent research suggests that compact, structured signaling supports more reliable and efficient coordination [2], yet LLM-based systems lack such structured communication substrates by default.

These observations motivate the central research question of our work:

What is the minimal amount of information that two LLM agents must exchange to remain aligned on a shared task?

Addressing this question requires mechanisms for explicitly constraining communication bandwidth and for evaluating how coordination quality changes as messages are compressed. Existing systems do not provide tools to enforce such bottlenecks, nor do they offer structured representations that can serve as low-bandwidth communication formats across different models.

To investigate this issue, we propose a mediated minimal-signaling architecture in which all inter-agent messages are processed by a two-stage mediator. The first stage applies learned natural-language compression using a lightweight summarization model like DistilBART, reducing verbosity while retaining essential intent. The second stage uses a compact LLM to extract discrete semantic keys that represent the compressed message in a concise, symbolic form. This mediator introduces an interpretable communication bottleneck that enables semantic compression before transferring the message further.

By combining learned compression, symbolic abstraction, and controllable bandwidth constraints, this approach provides a foundation for analyzing agent coordination, cross-model interoperability, and the emergence of low-bandwidth communication structures in LLM-based multi-agent environments.

2 Background

Multi-agent artificial intelligence systems involve multiple autonomous entities interacting within a shared environment to achieve individual or collective goals. Classical multi-agent research focuses on coordination, communication, negotiation, shared belief formation, and task allocation. In these systems, communication is typically implemented through structured, symbolic, and bandwidth-efficient messages that support reliable decision-making under any computational constraint.

LLM-based multi-agent systems inherit many of these principles but differ fundamentally in how communication is represented. Rather than relying on predefined symbolic vocabularies or compact messages, LLM agents communicate almost exclusively through natural language. This shift introduces both new opportunities and new limitations. Natural language offers high expressive capacity and generality, enabling agents to describe internal state, intentions, and plans in a detailed and open-ended form. However, it also introduces verbosity, redundancy, and potential ambiguity, particularly when models differ in architecture or training distribution.

The reliance on natural language as the default communication channel poses several challenges for coordination. First, message length directly competes with task-relevant in-

formation for space within the model's finite context window. Second, semantic variations across models can lead to inconsistent interpretation of same/similar instructions. Third, extended dialogue between agents may introduce drift in shared beliefs, especially when multiple rounds of generation and summarization accumulate noise over time. It is often the case that these systems lose track of the initial task due to accumulated noise.

These challenges are reflected in recent multi-agent LLM systems. Frameworks such as CAMEL demonstrate that structured role-based collaboration can emerge through natural language interaction [3], while self-consistency strategies highlight the benefits of aggregating multiple reasoning paths per response[4]. Other work shows that LLMs can evaluate one another's outputs through judgment systems such as MT-Bench and Chatbot Arena [5]. Although these systems showcase the potential of natural-language-mediated coordination, they also reveal the limitations of relying on free-form text for stable long-term communication.

In parallel, research on efficient representation of information within long text sequences has explored architectures such as Longformer, which introduces local and global attention patterns to handle extended inputs [1]. While such models fix some context limitations, they do not resolve the core issue that natural language itself is a suboptimal medium for machine-to-machine communication. Classical studies on emergent communication in multi-agent environments demonstrate that compact and structured signaling tends to arise when agents are optimized for coordination efficiency rather than explicit human interpretability [2].

Taken together, these observations underscore a gap between the communicative needs of multi-agent LLM systems and the capabilities provided by unconstrained natural language channels. The absence of low-bandwidth, interpretable, model-agnostic representations constrains the scalability of multi-agent coordination. This motivates the development of mechanisms that explicitly structure, compress, and regulate the information exchanged between LLM agents.

3 Literature Review

Research on communication within LLM-based multi-agent systems spans several areas, including natural-language coordination, structured summarization, long-context modeling, and emergent signaling in classical multi-agent environments. Together, these works outline the capabilities and limitations of existing approaches but do not provide mechanisms for regulating or minimizing the information passed between agents.

3.1 Natural-Language-Based Multi-Agent Coordination

A significant portion of recent MAAS research treats natural language as the default communication medium. The CAMEL framework demonstrates that structured role-playing can guide LLM agents toward cooperative problem solving through dialogue-based interaction [3]. Complementary work on self-consistency shows that generating and aggregating multiple reasoning paths can improve collaborative reasoning outcomes across agents [4]. Evaluation systems such as MT-Bench and Chatbot Arena extend this paradigm by showing that LLMs can serve as judges evaluating the reasoning quality of other models [5]. While these systems highlight the expressive power of natural-language coordination, they also reveal its limitations, including verbosity and susceptibility to drift in multi-turn interactions.

3.2 Summarization, Pruning, and Message Compression

A second line of work seeks to improve communication efficiency by reducing natural-language message length. Techniques include hierarchical summarization, iterative condensation, and heuristic pruning. These approaches can reduce context size and improve computational efficiency, but they often lack robustness. Summaries may omit subtle but contextually important details, and handcrafted templates generalize poorly across tasks. Moreover, these methods compress surface form rather than reasoning about what it chooses to compress, leaving open the question of what information is minimally necessary for coordination.

3.3 Long-Context and Efficient Attention Architectures

Architectural improvements such as Longformer provide mechanisms for handling extended sequences through sparse and global attention patterns [1]. These models significantly expand the feasible input length for transformer-based systems, enabling them to process long dialogues or documents more effectively. However, such models implicitly assume that large text volumes remain the medium of communication. They do not address whether natural language is inherently inefficient for machine-to-machine exchange or whether more compact representations could achieve similar/better coordination.

3.4 Emergent Signaling in Classical Multi-Agent Systems

Earlier work in classical multi-agent learning investigates how communication protocols emerge when agents are optimized for coordination under different pressures. Lazaridou et al. show that agents in referential games often converge to concise artificial communication

codes when required to coordinate efficiently [2]. These codes tend to be minimalistic, structured, and directly tied to the task. Although such agents are far simpler than modern LLMs, this line of research illustrates that when communication bandwidth is constrained, structured signaling can emerge naturally.

3.5 Gap in the Literature

Across these research areas, a consistent limitation appears. Existing multi-agent LLM systems either assume unconstrained natural-language communication or employ simple surface-level compression techniques that do not offer control over information content. Addressing this gap motivates the mediated minimal-signaling architecture introduced in the next section.

4 Proposed Approach

The objective of this work is to establish a framework for studying minimal information exchange between LLM agents. To do so, we introduce a mediated communication architecture that enforces an explicit bottleneck between agents through a two-stage transformation: learned natural-language compression and semantic-key extraction. This design makes it possible to vary communication bandwidth, evaluate the robustness of coordination under constrained signaling, and investigate whether stable low-bandwidth representations can support reliable multi-agent behavior.

The proposed approach builds on findings from prior work while addressing limitations identified in natural-language coordination and summarization-based compression. The architecture is model-agnostic and can be placed between any pair of LLM agents, enabling controlled experiments across different multi-agent systems.

4.1 Relation to Prior Work

Existing MAAS systems rely predominantly on natural-language dialogue for coordination, which provides flexibility but introduces verbosity and stylistic variability. Prior work demonstrates that structured role-play [3], multi-path reasoning [4], and evaluative judging [5] can improve the reliability of natural-language interaction. However, these systems do not regulate the amount of information transmitted for bandwidth-constrained settings.

Approaches based on summarization or heuristic pruning reduce message length but do not address the underlying problem of controlling semantic content. Long-context architectures such as Longformer [1] extend the feasible input size for transformers but implicitly assume that large volumes of natural-language text remain the communication pro-

tocol. Classical emergent communication research shows that compact signaling systems can arise when communication is optimized for efficiency [2], but these findings have not yet been operationalised for LLM-based agents.

The proposed mediated architecture differs from prior work in three key ways. First, it introduces an explicit, enforceable communication bottleneck between agents. Second, it separates semantic content from surface-level linguistic expression by using DistilBART for reasoning level compression. Third, it provides a mechanism for systematically measuring how coordination degrades under increasingly constrained communication, enabling approximation of a minimal signaling floor.

4.2 Mediated Minimal-Signaling Architecture

The proposed architecture introduces a mediator positioned between any pair of communicating LLM agents. All messages produced by one agent must pass through this mediator before reaching the next. The mediator performs two sequential transformations: learned natural-language compression and semantic-key extraction.

Stage 1: Learned Natural-Language Compression: The first stage applies a learned compression model to reduce the verbosity of the agent's original message while preserving essential intent. In this work, we employ DistilBART, a lightweight sequence-to-sequence summarization model capable of producing concise abstractions of longer text. DistilBART is chosen due to its balance of efficiency and semantic range, making it suitable for real-time multi-agent settings.

This compression stage may be applied recursively. If the compressed message exceeds a predefined communication budget, measured in tokens or bytes, the mediator invokes DistilBART again on its own output. Recursion continues until the message is below the specified threshold or until compression quality deteriorates, as determined by pre-defined verification heuristics. This mechanism provides control over communication bandwidth and allows experiments with different message sizes to evaluate how task performance varies with reduced information exchange.

Stage 2: Semantic-Key Extraction: After compression, the mediator transforms the condensed message into a discrete set of semantic keys. A compact and small parameter LLM is used for this abstraction step, mapping the compressed text to symbolic units that represent the core meaning of the message. These semantic keys are intentionally low-dimensional and follow a consistent schema, such as `INSTRUCTION(verify)`, `STATE(confidence: high)`, or `GOAL(what-to-do)`, etc.

The purpose of semantic keys is to provide a model-agnostic communication protocol that minimizes surface-level linguistic variability. Unlike natural-language messages, which differ substantially across LLM families, semantic keys offer a stable interface that can be interpreted reliably by agents. This stage therefore serves as a bridge between the subjective expressiveness of LLM-generated text and the efficiency required for objective machine-oriented communication.

Optional Verification and Extensions: The mediator may optionally include a verification layer—for example, a lightweight judge model that ensures the semantic keys accurately reflect the intent of the compressed message. Additional components such as confidence evaluators or rule-based filters, can be added as needed.

Communication Flow and Purpose. The complete communication flow is:

$$\text{Agent A} \rightarrow \text{DistilBART Compression} \rightarrow \text{Semantic-Key Extraction} \rightarrow \text{Agent B}.$$

By enforcing this structured bottleneck, the architecture enables controlled experiments on how multi-agent coordination degrades or adapts under increasingly constrained signaling.

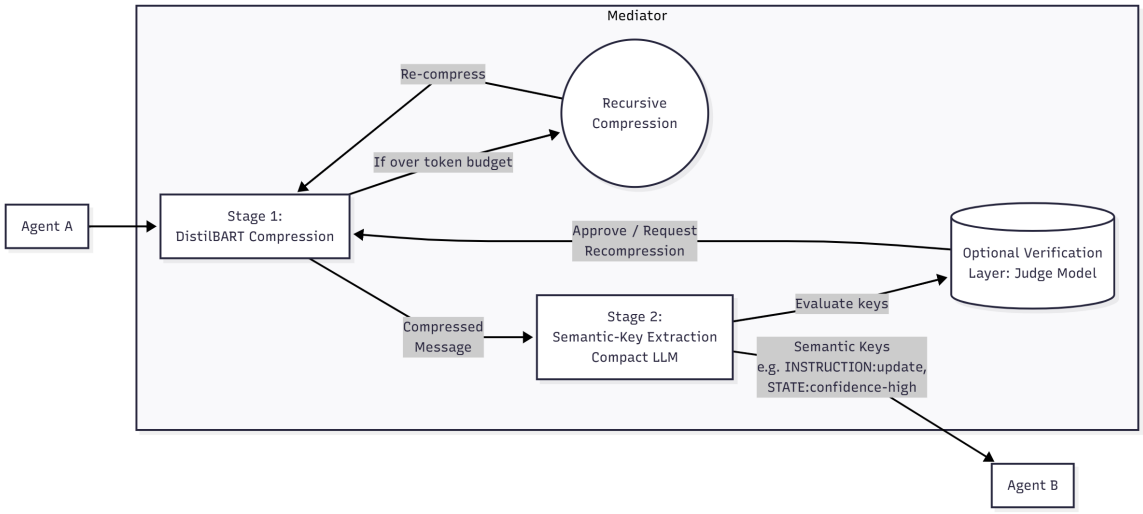

Figure 1: Mediated minimal-signaling architecture.

References

[1] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020. doi: 10.48550/arXiv.2004.05150.

[2] Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. Multi-agent cooperation and the emergence of (natural) language. *arXiv preprint arXiv:1612.07182*, 2017. doi: 10.48550/arXiv.1612.07182.

[3] Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for "mind" exploration of large language model society. *arXiv preprint arXiv:2303.17760*, 2023. doi: 10.48550/arXiv.2303.17760.

[4] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022. doi: 10.48550/arXiv.2203.11171.

[5] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Hao Zhang, Eric P Xing, Joseph E Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*, 2023. doi: 10.48550/arXiv.2306.05685.