

Communication Alignment in Multi-Agent Systems: From MARL to LLM Agent Communication

Xinghai Wei¹, Jie Yuan¹, Hwei-Ming Chung², Tingting Yuan^{3*},

¹Beijing University of Posts and Telecommunications,

²Zhuhai Fudan Innovation Research Institute & Fudan University,

³IMC Krems University of Applied Sciences

junjuntvt@bupt.edu.cn, tingting.yuan@imc.ac.at

Abstract

Effective collaboration in multi-agent systems hinges on communication. Existing studies analyze agent communication indirectly through task performance, such as improvements in reward and success rate, without explicitly characterizing communication itself. As a result, the quality of communication protocols remains difficult to evaluate and interpret, limiting their reliability, scalability, and generalizability. To better characterize communication quality, this survey focuses on *communication alignment*, which emphasizes semantic fidelity and convention consistency across agents. We review communication alignment from emergent communication in multi-agent reinforcement learning (MARL) to latent communication powered by Large Language Models (LLMs), highlighting existing methods and the emerging challenges introduced by natural-language communication. Finally, we discuss future directions toward more effective, interpretable, and generalizable multi-agent communication.

1 Introduction

Multi-agent systems (MAS) have emerged as a foundational paradigm for designing intelligent systems, where multiple autonomous agents act within shared environments. These systems have found extensive applications in domains such as autonomous driving (Liu et al., 2024a, 2021), robotics (Nayak et al., 2024), network management (Naderalizadeh et al., 2021; Sana et al., 2020), and, more recently, Large Language Model (LLM)-driven collaborative software engineering (Hong et al., 2023; Tufano et al., 2024; Tang et al., 2024). In such contexts, agents typically operate under partial observability, limited local capabilities, and role-specific objectives, rendering communication critical for exchanging information and coordinat-

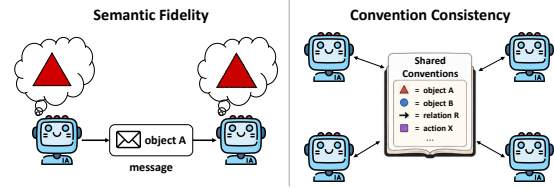


Figure 1: Semantic fidelity and convention consistency.

ing distributed decision-making (Wei et al., 2025; Yuan et al., 2023).

Agent communication can be viewed as a process of semantic compression and reconstruction, in which semantics encode an agent’s local observations, action intentions, or compressed summaries of interaction history (Ding et al., 2026). Existing studies and surveys have extensively examined agent communication in multi-agent reinforcement learning (MARL) (Peters et al., 2025; Brandizzi, 2023; Boldt and Mortensen, 2024; Zhu et al., 2024) and LLM MAS (Li et al., 2024b; Yuan et al., 2026; Kong et al., 2025; Tran et al., 2025; Chen et al., 2026; Yan et al., 2025; Guo et al., 2024), showing that communication can improve downstream task performance such as reward and success rate. However, these studies still lack a communication-centered understanding of whether and how agents can establish reliable communication. In particular, key properties of agent communication, such as semantic fidelity (Xie et al., 2022; Gündüz et al., 2022) and convention consistency (Rita et al., 2022b; Wang et al., 2025a; Li and Bowling, 2019) across agents, remain undercharacterized, as illustrated in Fig. 1. These properties are especially important in heterogeneous MAS, where agents may rely on different foundation models or interact under different contexts. In such settings, preserving intended semantics while maintaining shared communication conventions becomes challenging, forming the central problem of *communication alignment* (Lo and Sengupta, 2022; Lo et al., 2024; Du et al., 2026; Li et al., 2024a).

This survey reviews *communication alignment*,

*Corresponding author.

which focuses on semantic fidelity and convention consistency across agents, spanning emergent communication in MARL to latent communication in LLMs. To the best of our knowledge, this is the first work to provide a unified survey of multi-agent communication from this alignment-oriented perspective. A comparison with existing surveys is provided in Appendix A.1. In this survey, we formulate the communication alignment problem in Sec. 2. We then review relevant literature spanning emergent communication in MARL (Sec. 3), natural-language communication in LLM (Sec. 4), and LLM latent communication, where agents exchange compressed hidden representations instead of natural-language messages (Sec. 5). Across these paradigms, we highlight the recurring trade-off between semantic fidelity and convention consistency, as summarized in Appendix A.2. Overall, our contributions are threefold:

- We introduce *communication alignment* as a communication-centered perspective for analyzing multi-agent communication beyond task performance. This perspective emphasizes semantic fidelity and convention consistency, which are critical for effective and scalable multi-agent collaboration.
- We systematically organize and compare communication methods in MAS through the perspective of communication alignment. We further analyze the advantages, limitations, and trade-offs of emergent, natural-language, and latent communication across MARL and LLM-based MAS.
- We identify key challenges and future directions, including convention adaptation, hybrid communication between emergent languages and natural language, as well as safety and ethical risks in emerging multi-agent systems.

2 Definition and Formulation

Communication in MAS can be formulated as an encoding-decoding process between agents. Let $\mathcal{A} = \{a_1, \dots, a_N\}$ denote a set of agents. Each agent can act as both a speaker and a listener. Let $z_i \in \mathcal{Z}$ denote the task-relevant semantic representation of agent a_i , which may encode its local observation, action intent, and compressed interaction history, and $\mathcal{Z}$ denotes the space of task-relevant semantics¹. At a specific time, the speaker a_i maps its

semantics into a message through an encoder $E_i : \mathcal{Z} \rightarrow \mathcal{L}$, where $\mathcal{L}$ denotes the message space. The generated message is $m_i = E_i(z_i)$. The listener a_j receives the message and reconstructs the speaker’s intended semantics using its decoder $D_j : \mathcal{L} \rightarrow \mathcal{Z}$, denoted as $\hat{z}_{i \rightarrow j} = D_j(m_i) = D_j(E_i(z_i))$. The listener then uses $\hat{z}_{i \rightarrow j}$, together with its own local observation and policy, to make coordinated decisions. The representation reconstruction error from agent a_i to agent a_j is defined as:

$$\mathbf{R}_{ij} = \mathbb{E}_{z_i}[d(D_j(E_i(z_i)), z_i)], \quad i \neq j, \quad (1)$$

where $d(\cdot, \cdot)$ measures the discrepancy between the listener’s reconstructed semantics and the speaker’s intended task-relevant semantics. In contrast to autoencoder-style approaches (Shao et al., 2021; Zhang et al., 2025a), which focus on information reconstruction within a single encoder–decoder pair, communication alignment addresses the compatibility of semantics–message mappings across a population of agents. Formally, the communication alignment objective can be written as:

$$\min \mathbb{E}_{i,j}[\mathbf{R}_{ij}]. \quad (2)$$

This objective extends pairwise semantics reconstruction to interoperability. Rather than allowing agents to co-adapt with a small set of partners and develop locally valid but externally unintelligible communication conventions, it encourages the emergence of shared semantics–message mapping conventions among agents, thereby supporting generalizable and scalable cooperation.

In realistic settings, the true semantic state z is often not directly observable. Therefore, this objective is usually optimized through surrogate signals such as task success rate and collaborative return. The central problem addressed in this paper is how to preserve semantic recoverability when agents have been separately trained and have converged to heterogeneous communication conventions.

3 Aligning Emergent Communication in MARL

Emergent machine language refers to communication conventions that agents autonomously develop through interaction, rather than languages explicitly specified by humans (Peters et al., 2025). Since these conventions are optimized directly toward

¹Note that $\mathcal{Z}$ denotes the task-relevant semantic space rather than the agents’ internal representation space. There-

fore, as long as the agents operate within the same collaborative task setting, they can be assumed to share a common semantic space (Cao et al., 2018; Rita et al., 2022b).

collaborative rewards, they can be highly compact and task-optimal (Auersperger and Pecina, 2022; Lee, 2024; Jiang and Lu, 2018; Kim et al., 2019; Sukhbaatar et al., 2016; Das et al., 2019; Singh et al., 2019). Yet this setting often leads to extreme co-adaptation and overfitting within the training community², such that the learned conventions become locally effective but fail to generalize across agents or contexts, resulting in convention mismatch. This section reviews methods that address this bottleneck, as shown in Fig. 2.

3.1 Convention-Invariant Learning

Convention-invariant learning aims to make agents robust to specific communication conventions before deployment. Rather than aligning already-emerged conventions, this paradigm applies training-time pressures to improve the robustness of agents’ encoders and decoders, discouraging agents from overfitting to specific conventions.

Encoding-Level Invariance. One line of work targets the equivalence-class nature of communication (Lewis, 1969; Cao et al., 2018; Rita et al., 2022b), aiming to learn conventions that are robust to different communities. Other-Play (OP) (Hu et al., 2020) enforces such robustness by training agents under randomized permutations of action, observation, and state spaces. Bullard et al. (2021) and Yan et al. (2023) restrict the valid mapping space through external constraints, such as communication cost. However, enumerating all possible mappings is infeasible, since the mapping space is large and unobservable. This motivates simpler generalization mechanisms. For example, Michel et al. (2023) isolate speaker–listener pairs to reduce co-adaptation. Cope and Schoots (2024) introduce randomness into the communication channel to prevent agents from relying on a fixed encoding.

Population-Level Diversification. Population-based training draws on sociolinguistic (Raviv et al., 2020) and language-evolution evidence (Bromham et al., 2015; Bromham, 2025) that larger, more diverse populations promote more generalized languages (Mordatch and Abbeel, 2018; Michel et al., 2023; Chaabouni et al., 2020). Strouse et al. (2021) and Li and Bowling (2019) train an agent against a diverse set of partners to encourage it to produce messages that can be un-

derstood by agents from different training communities. MEP (Zhao et al., 2023) and TrajeDi (Lupu et al., 2021) introduce diversity objectives at the population and trajectory level, encouraging agents to experience a wider range of coordination patterns. Gupta et al. (2021) propose a dynamic population-based meta-learning approach, where new agents are continually introduced during training. Chaabouni et al. (2022); Rita et al. (2022a); Kim and Oh (2021) point out that additional mechanisms, such as imitation or aggregation, are needed to effectively leverage population pressure.

Despite their deployment-time appeal, these methods have two key limitations. First, the semantics–message mapping space is vast, making universal convention invariance difficult. Second, they rely on sufficient training-time diversity or structural pressure, which may not persist when agents or tasks shift after deployment.

3.2 Language Bridging

Language bridging aims to connect pre-existing communication systems without disrupting their learned conventions. It can be viewed as an intermediate mapping between E_i and D_j , expressed as $D_j(B(E_i(z_i)))$, where B denotes the bridging module. We categorize existing approaches to language bridging into the following two classes:

Common Language Bridging. It aligns agents from different communities by mapping their messages into a shared medium. Natural language is a common choice because it supports human-AI collaboration and provides a semantically rich interface (Lo and Sengupta, 2022; Toquebiau et al., 2025; Huh and Mohapatra, 2025; Levy et al., 2025; Li et al., 2024a; Sundar et al., 2025; Andreas et al., 2017). However, parallel corpora between emergent languages and common language are often scarce, motivating unsupervised neural machine translation (Levy et al., 2025), brief induced translation (Lee et al., 2018; Andreas et al., 2017; Zheng et al., 2026), and self-supervised representation learning (Lin et al., 2021).

Translator Bridging. It directly learns mappings between communication conventions. These methods differ from common-language bridging because they do not require all communities to share the same intermediate language; instead, they rely on pairwise or partner-conditioned translation modules. Lee et al. (2018); Cope and McBurney (2024) introduce trainable parameters that enable agents to learn a translation function for aligning conventions

²A community denotes a group of agents that co-adapt through joint training, forming shared conventions for mapping task-relevant semantics to messages and interpreting them.

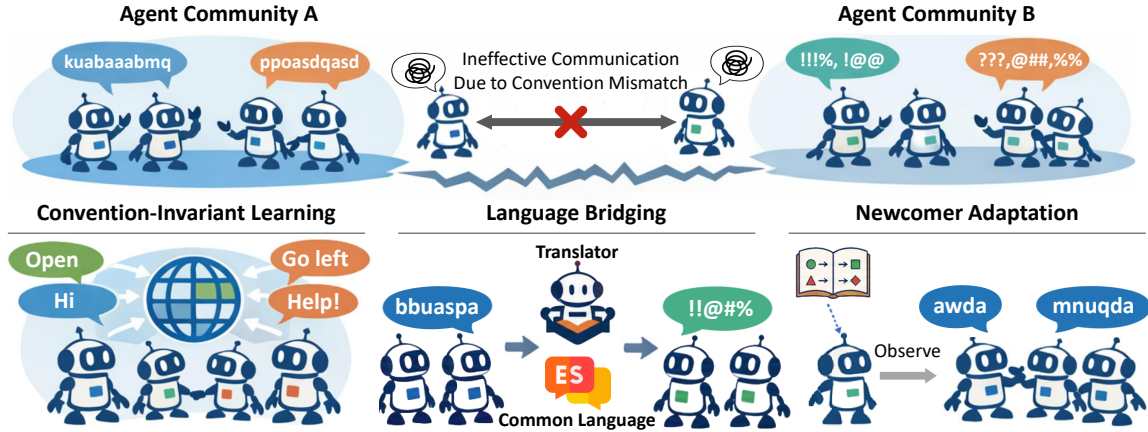


Figure 2: Taxonomy of communication alignment over emergent machine languages in MARL.

across communities. Thomas et al. (2021) equip each agent with multiple heads, each implementing a translation function specialized for communication with a distinct community. Kobrock et al. (2025); Graesser et al. (2019); Martinez-Lopez et al. (2025) designate a subset of agents as translators that interact simultaneously with agents from different communities, acting as intermediaries and transforming information.

Language bridging offers a lightweight alternative to enforcing a universal convention across communities. However, its flexibility comes at a cost. Common-language-based bridging introduces additional information loss, as agents must further compress and reconstruct their messages within the shared intermediate language space. Translator-based bridging faces a scalability bottleneck, as the number of translators grows rapidly with the number of interacting communities.

3.3 Newcomer Adaptation

Prior work has shown that, although agents in different communities often acquire distinct conventions, the implicit structural properties of the languages they learn (e.g., compositionality (Chaabouni et al., 2020) and Zipf regularities (Chaabouni et al., 2019)) exhibit a degree of universality and can support rapid adaptation to new languages. Motivated by this observation, several approaches enable newly introduced agents to integrate into established communication communities by adapting their encoders and decoders via few-shot learning.

Imitation. A straightforward approach is to expose the newly introduced agent directly to semantic-message pairs from the target community, allowing it to learn and imitate the communicative conven-

tions of the community. For example, ECTL (Cope and McBurney, 2024) provides a small set of observation-message pairs to the new agent and encourages it to generate messages consistent with those used by agents in the target community. At larger population scales, Chaabouni et al. (2022) incorporates a voting mechanism to improve language acquisition and cooperative performance.

Intention modeling. Direct imitation is simple and effective, yet it assumes access to semantics-message pairs, which are often difficult to obtain in practice. To overcome this limitation, another line of work draws on theory of mind (Jaques et al., 2019; Wellman, 2011; Astington and Jenkins, 1995), enabling an incoming agent to infer the intentions of target-community agents from a small number of online interactions. Zhu et al. (2021) construct a mental model of the newly introduced agent to predict its responses to different messages, and use meta-learning to adapt to new communication partners. Cope and McBurney (2022, 2025) propose to infer the speaker’s latent observations by leveraging causal relationships among messages, action trajectories, and the environment. Nevens et al. (2022) reconstruct the speaker’s intentions from messages, and then abstract these patterns through fluid construction grammar.

Despite its flexibility, newcomer adaptation also faces important limitations. First, imitation-based methods depend on access to aligned semantics-message examples from the target community, but such data are often difficult to collect in realistic settings. Second, intent-modeling requires the newcomer to recover conventions from limited interactions, which can be sample-inefficient and unstable.

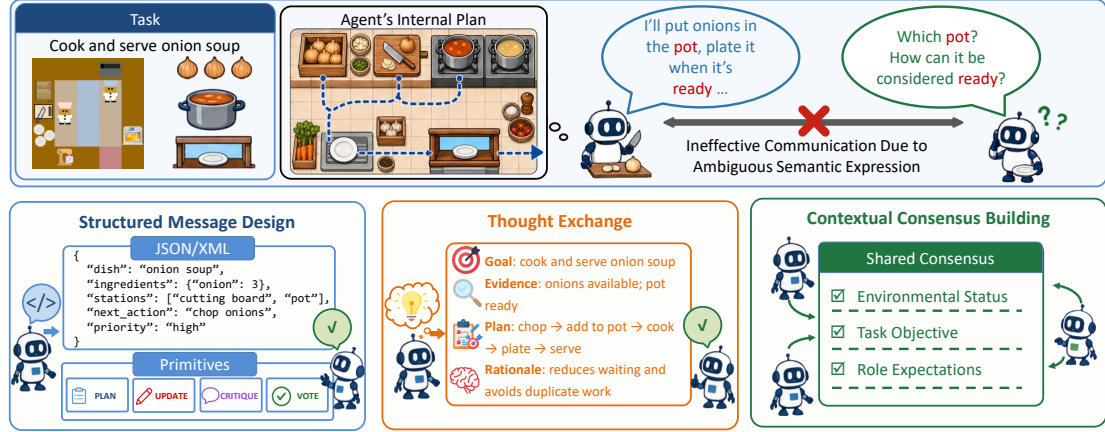


Figure 3: Taxonomy of communication alignment over natural language in LLM MAS.

3.4 Takeaway

The methods reviewed above improve interoperability across emergent communication conventions, but they do not eliminate the fundamental locality of machine-induced conventions. Emergent languages remain attractive because they are compact, task-adaptive, and efficient, yet these same properties make them brittle across independently trained communities. This motivates the turn to natural language in LLM-based agents. By leveraging natural language as a shared medium for information exchange, LLM agents can mitigate cross-community convention mismatches that emerge from independently trained MARL agents. However, natural language cannot fully eliminate the communication alignment problem.

4 Aligning Natural-Language Communication in LLM MAS

The shift from emergent languages to natural language changes, rather than eliminates, communication misalignment. Trained on massive corpora, LLMs can communicate directly through natural language, but natural-language messages remain lossy projections of agent semantics (Pham et al., 2024; Fu et al., 2025). Therefore, communication alignment in LLM agents focuses on structuring and enriching communication so that listeners can more reliably recover intended semantics. Existing studies mainly address this challenge through the three directions shown in Fig. 3.

4.1 Structured Message Design

Structured message design mitigates semantic misalignment by imposing constraints on the expressive form of natural-language messages, thereby

reducing the admissible message space $\mathcal{L}$. Instead of allowing agents to exchange unconstrained free-form text, these methods define schemas, fields, or typed communicative acts that make the semantic role of each message explicit.

Schema-Grounded Message Structuring. Schema-grounded communication decomposes free-form utterances into predefined or task-adaptive semantic fields, making the speaker’s intended meaning easier to recover. For example, TalkHier (Wang et al., 2025b) factorizes each message into structured fields such as background, conversation history, and intermediate output. Similar design principles are adopted by Sun et al. (2025); Shi et al. (2023); Hong et al. (2023); Fleming et al. (2026). AutoForm (Chen et al., 2024b) allows agents to choose or devise alternative schema formats for specific tasks.

Communication Primitives. These methods treat communication as a workflow composed of typed operations, so that the semantics of each message is defined not only by its text but also by its role in the interaction. Fleming et al. (2025) frame communication primitives as programmable interfaces at the network-protocol layer. Anemoi (Ren et al., 2025) incorporates communication primitives into the A2A protocol. LACP (Li et al., 2025) proposes a three-layer communication framework with minimal message types, each carrying explicit semantic intent and supporting transactional guarantees.

Overall, structured message design improves alignment by making natural-language communication more machine-checkable and semantically typed. However, structure is not universally beneficial. Overly restrictive formats can limit expressiveness and hinder collaboration (Geng et al.,

2025; Tam et al., 2024). In addition, message structures are often manually designed and task-specific, which limits their generality and scalability.

4.2 Thought Exchange

Natural-language messages can be viewed as compressed projections of an agent’s internal reasoning state, often omitting critical premises, obscuring uncertainty, or blurring underlying intentions. To mitigate these limitations, a line of work augments messages m with intermediate reasoning r , i.e., $m \leftarrow m + r$, to improve inter-agent alignment. Comm-CoT (Qi et al., 2025) and Exchange-of-Thoughts (Yin et al., 2023) incorporate Chain-of-Thought (CoT) reasoning directly into communication messages. OPTAgent (Bi et al., 2025) adapts the communication topology among agents to preserve informative reasoning content. Mind-Map (Kim et al., 2025) allows agents to expose their dialogue summary to prevent redundancy.

Thought exchange improves semantic recoverability by making latent semantics more explicit, but it can not guarantee stable alignment across long interactions. Exposed reasoning may be incomplete, strategically filtered, or inconsistent (Turpin et al., 2023; Madsen et al., 2024; Parcalabescu and Frank, 2024), motivating methods that build and maintain common ground over time.

4.3 Contextual Consensus Building

Contextual consensus building shifts communication alignment from single-message interpretation to the iterative construction of consensus³. In this context, the semantic reconstruction process can be reformulated as $\hat{z}_{i \rightarrow j} = D_j(m_i, G)$, where G denotes the consensus. Note that, unlike thought exchange, which explicitly augments m with additional reasoning text, G implicitly emerges from memory, intermediate artifacts, and other interaction traces.

Memory Accumulation. These methods maintain common ground primarily through agents’ own memories, allowing them to adapt their interpretations, conventions, and social assumptions. Chen et al. (2023); Kouwenhoven et al. (2025b,a); Ashery et al. (2025); Ruan et al. (2025) show that shared communicative conventions can emerge

from memory. Qian et al. (2024); Zhang et al. (2024a); Park et al. (2023) enable agents to infer partners’ intentions from memory and received messages to improve collaboration. Deng et al. (2026) introduces consensus judgment to enable agents to form consensus more stably under limited interaction. From a social-psychological perspective, AlKhamissi et al. (2024); Ki et al. (2025); Zhang et al. (2024b) reveal that multi-turn interaction helps agents surface and revise cultural or social priors that shape intent interpretation.

Artifact-Mediated Coordination. Whereas memory accumulation methods rely on agents to form consensus through their individual memories, another line of work (Huang et al., 2025; Hong et al., 2023; Liu et al., 2025; Ma et al., 2026; Kim and Oh, 2021; Zhang et al., 2025b; Chopra et al., 2025) adopts a more centralized coordination paradigm by externalizing shared semantics into persistent and jointly accessible objects. For example, MetaGPT (Hong et al., 2023) and KGMAF (Huang et al., 2025) employ a shared message pool to store role-specific reasoning traces, allowing subsequent agents to operate directly on shared semantic states. RTADev (Liu et al., 2025) introduces an intent alignment mechanism to prevent misinterpretations from propagating and compounding throughout the interaction process. AutoData (Ma et al., 2026) introduces a hypergraph-based mechanism that exposes each agent only to the artifacts relevant to its current context, thereby reducing context contamination. NexusSum (Kim and Kim, 2025) restructures unordered distributed communication into a three-stage pipeline and employs interactive compression to shorten artifacts.

Contextual consensus building offers a promising solution for long-horizon cooperation. However, such approaches typically come at the cost of substantial communication overhead. Moreover, because consensus emerges implicitly through multiple rounds of interaction, it is difficult for humans to inspect, verify, or correct. As a result, harmful biases introduced early in the interaction process may persist and remain uncorrected.

4.4 Takeaway

The methods reviewed above show that natural language provides a more general substrate for communication alignment than emergent machine languages. We summarize the three categories of methods mentioned above in Table 1. We further conduct a controlled comparison of represen-

³In this sense, consensus does not mean agreement on the answer to a particular question. Rather, it refers to an emergent shared understanding of task objectives, environmental states, role expectations, and conventions that are gradually formed through repeated interaction.

Table 1: Summary of different communication alignment methods in LLM MAS.

Category	Focused Bottleneck	Working Mechanism	Overhead	Failure Modes	Representative Methods
Structured message design	Ambiguous free-form messages	Constrain message space $\mathcal{L}$	Low	Improper schema/primitive design, coverage gaps	Wang et al. (2025b); Sun et al. (2025); Shi et al. (2023); Li et al. (2024a); Ren et al. (2025)
Thought exchange	Hidden beliefs, plans, and intentions	Augment message m with exposed reasoning r	Medium	Unfaithful thought exposure, rationale inconsistency	Qi et al. (2025); Yin et al. (2023); Bi et al. (2025); Kim et al. (2025)
Contextual consensus building	Lack of persistent common ground	Accumulate shared consensus G	High	Error propagation, persuasion overriding truth, harmful bias persistence	Chen et al. (2023); AlKhamissi et al. (2024); Hong et al. (2023); Liu et al. (2025)

tative methods from the three categories under matched task settings, as shown in Tables 2 and Table 3. Specifically, we evaluate AutoForm (Chen et al., 2024b), EoT (Yin et al., 2023), and BCCS (Deng et al., 2026) on Collab-Overcooked (Sun et al., 2025) tasks at difficulty levels 3–4 and the Minecraft tasks in MultiAgentBench (Zhu et al., 2025), using two different backbone models. These results empirically highlight a clear trade-off between communication effectiveness and efficiency in natural-language communication in MAS. The shared bottleneck of these methods lies in the textual channel itself. Natural-language messages must compress high-dimensional semantic states into token sequences, which may omit implicit premises and introduce redundancy. These limitations suggest that further progress may require moving beyond natural language as the sole communication substrate. The next section therefore turns to latent communication, an emerging paradigm that directly transmits model-internal representations to preserve higher-fidelity semantic exchange between LLM agents.

5 Aligning Latent Communication for Efficient LLM-Agent Collaboration

Latent communication emerges from the limitations of natural-language communication discussed above. It transmits model-internal representations that are closer to the agent’s latent semantic state. This paradigm can be viewed as replacing the discrete natural-language space $\mathcal{L}^{nl}$ with a compact continuous vector space $\mathcal{L}^{cv}$, which is more suitable for machine processing. This design reduces the lossy projection from hidden states to discrete tokens and can preserve fine-grained information that is difficult to verbalize.

5.1 Dense Embedding Sharing

These methods remove the token sampling step. CIPHER (Pham et al., 2024) is the first to use raw transformer output embeddings as the communication medium between agents, showing that embedding-based communication outperforms natural language in most settings. Wu et al. (2025) extend this paradigm by introducing a trainable seq2seq module, enabling communication via embeddings across different foundation LLMs. SDE (Tang et al., 2025) transmits token-wise state transition trajectories alongside natural language tokens. ThoughtComm (Zheng et al., 2025) allows agents to produce latent thought representations and route them to relevant agents.

5.2 KV Cache Sharing

Beyond embeddings, several works have investigated KV-cache-based communication (Liu et al., 2024b; Fu et al., 2025; Shi et al., 2025; Bian et al., 2025; Shkolnikov, 2026). Compared with embedding-based approaches, KV cache can preserve substantially richer attention information. This advantage, however, comes at the cost of significantly higher communication overhead. To improve efficiency, DroidSpeak (Liu et al., 2024b) proposes selective KV-cache reuse and recomputation to reduce the amount of cache that must be loaded. KVComm (Shi et al., 2025) selectively transmits only the most important KV cache entries according to attention-based importance. C2C (Fu et al., 2025) introduces a projection module to align cache representations across models, enabling KV-cache-based communication between different foundation LLMs.

Table 2: Results with Deepseek-v4-flash.

Method	Category	Collab-Overcooked		MultiAgentBench	
		Task Score	Token Cost	Task Score	Token Cost
Autoform (Chen et al., 2024b)	Structured message design	55.9	1.0×	32.4	1.0×
EoT (Yin et al., 2023)	Thought exchange	83.0	11.2×	46.9	15.3×
BCCS (Deng et al., 2026)	Contextual consensus building	86.2	48.5×	44.6	66.1×

Table 3: Results with Qwen3.7-max.

Method	Category	Collab-Overcooked		MultiAgentBench	
		Task Score	Token Cost	Task Score	Token Cost
Autoform (Chen et al., 2024b)	Structured message design	48.7	1.0×	29.6	1.0×
EoT (Yin et al., 2023)	Thought exchange	72.9	10.8×	43.2	18.9×
BCCS (Deng et al., 2026)	Contextual consensus building	76.1	51.7×	43.8	91.1×

5.3 Takeaway

Latent communication resembles emergent machine communication in its efficiency, but with an important difference: the latent spaces of pretrained LLMs are shaped by natural language and world knowledge, so they may already contain reusable semantic structure. Investigating what semantic structures exist in the latent spaces of pretrained LLMs and how these structures can be leveraged for inter-agent communication is a promising research direction. Although latent communication can offer higher information density and fidelity than natural language, it renders communication among LLM agents largely opaque. Moreover, such methods typically require additional training to align the semantics of embeddings or KV caches across different foundation models, limiting their generalizability and scalability.

6 Future Directions

In this section, we review challenges and discuss future directions for communication alignment.

6.1 Semantic-Fidelity Evaluation

In practice, the true semantic representation z is often not directly observable, making it difficult to directly evaluate communication alignment using Eq. (2). Existing studies typically rely on accessible signals, such as agents’ observations and rewards, to indirectly assess the degree of communication alignment. We summarize several evaluation metrics adopted in prior work in Appendix A.4.2. Nevertheless, these metrics suffer from limited reliability and fail to distinguish genuine semantic recovery from accidental coordination. Future work should develop diagnostic metrics that explicitly evaluate semantic fidelity, including goal recovery, belief consistency, and plan reconstruction.

6.2 Adaptive Message Structure Design

As a simple yet effective mechanism for improving communication alignment, structured message design has been widely adopted in existing multi-agent frameworks, yet its effectiveness often depends on manually designed structures and task-specific assumptions. This creates a gap between fixed structure design and open-ended multi-agent environments, where agents may differ in model architecture, capability, role, tool access, and communication conventions (Chen et al., 2024b). Future research should move toward adaptive message design that can be negotiated, revised, and personalized during interaction.

6.3 Convention Acquisition by Newcomers

LLM-agent systems face a scalability demand similar to the newcomer adaptation problem in MARL: a new agent may need to join an already established agent community and contribute to collaboration. In MARL settings, this problem has been studied through few-shot adaptation. However, in LLM-agent systems, conventions may contain coordination norms that implicitly emerge from context and memory. Future work should study how to abstract these conventions into transferable representations that can be consumed by newcomer agents.

6.4 Hybrid Communication Between Natural Language and Latent Representations

Natural language supports training-free coordination but is ambiguous and lossy, while emergent languages and latent representations can preserve richer semantic information but are harder to generalize. Although several works (Elberg et al., 2026; Du et al., 2026; Tucker et al., 2022) have attempted to balance these properties, they often struggle to satisfy all of them simultaneously. Fu-

ture systems should explore hybrid communication architectures, where agents may use latent representations for high-fidelity machine-to-machine coordination while generating natural-language summaries for verification and human oversight.

6.5 Interpretability and Oversight

As LLM-based MAS become more autonomous, communication alignment should be treated not only as an agent-to-agent coordination problem, but also as a human-agent governance problem. Several works (Sarkar et al., 2025; Mathew et al., 2025; Motwani et al., 2024) have found that efficient agent communication may gradually move beyond forms that humans can easily inspect. This creates risks of hidden misunderstanding, harmful collective bias, or coordination strategies that are effective for agents but misaligned with human intent. Besides, latent communication among LLM agents raises unresolved challenges regarding how to prevent latent messages from encoding irrelevant, private, or unsafe information. Future work should develop oversight mechanisms that allow humans to monitor, interpret, and intervene in inter-agent communication, ensuring that agent communication remains aligned with human expectations for privacy, ethics, and safety.

7 Conclusion

In this paper, we present a comprehensive survey of communication alignment in multi-agent cooperation. We review communication alignment methods across emergent communication, natural-language communication, and latent communication. By connecting MARL and LLM-based MAS under a unified perspective, we highlight the fundamental trade-offs among semantic fidelity and communication consistency. We hope this survey provides useful insights for future research on scalable, adaptive, and human-compatible multi-agent communication.

Limitations

This paper presents a comprehensive survey and synthesis of existing methodologies and analyses on communication alignment in MAS. Nevertheless, given the extensive body of related work, particularly the rapidly expanding literature on LLM-agent communication, we may have overlooked some valuable contributions. Moreover, several methods, such as MetaGPT (Hong et al., 2023) and

AutoDev (Tufano et al., 2024), integrate multiple alignment strategies discussed in this paper. Therefore, assigning them to a single specific category may be imprecise.

References

- Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422.
- Jacob Andreas, Anca Dragan, and Dan Klein. 2017. Translating neuralese. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 232–242.
- Ariel Flint Ashery, Luca Maria Aiello, and Andrea Baronchelli. 2025. Emergent social conventions and collective bias in llm populations. *Science Advances*, 11(20):eadu9368.
- Janet Wilde Astington and Jennifer M Jenkins. 1995. Theory of mind development and social understanding. *Cognition & Emotion*, 9(2-3):151–165.
- Michal Auersperger and Pavel Pecina. 2022. Defending compositionality in emergent languages. In *Proceedings of the 2022 Conference of the North American chapter of the association for computational linguistics: Human language technologies: Student research workshop*, pages 285–291.
- Zhenyu Bi, Meng Lu, Yang Li, Swastik Roy, Weijie Guan, Morteza Ziyadi, and Xuan Wang. 2025. Optagent: Optimizing multi-agent llm interactions through verbal reinforcement learning for enhanced reasoning. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 1713–1728.
- Zhuohang Bian, Feiyang Wu, Teng Ma, and Youwei Zhuo. 2025. Tokencake: A kv-cache-centric serving framework for llm-based multi-agent applications. *arXiv preprint arXiv:2510.18586*.
- Brendon Boldt and David Mortensen. 2024. A review of the applications of deep learning-based emergent communication. *arXiv preprint arXiv:2407.03302*.
- Nicolo’ Brandizzi. 2023. Toward more human-like ai communication: A review of emergent communication research. *IEEE Access*, 11:142317–142340.
- Lindell Bromham. 2025. Population size and language change: an evolutionary perspective. *Annual Review of Linguistics*, 11(1):183–208.
- Lindell Bromham, Xia Hua, Thomas G Fitzpatrick, and Simon J Greenhill. 2015. Rate of language evolution

- is affected by population size. *Proceedings of the National Academy of Sciences*, 112(7):2097–2102.
- Kalesha Bullard, Douwe Kiela, Franziska Meier, Joelle Pineau, and Jakob Foerster. 2021. Quasi-equivalence discovery for zero-shot emergent communication. *arXiv preprint arXiv:2103.08067*.
- Kris Cao, Angeliki Lazaridou, Marc Lanctot, Joel Z Leibo, Karl Tuyls, and Stephen Clark. 2018. Emergent communication through negotiation. In *Proceedings of the International Conference on Learning Representations*, pages 3128–3142.
- Rahma Chaabouni, Eugene Kharitonov, Diane Bouchacourt, Emmanuel Dupoux, and Marco Baroni. 2020. Compositionality and generalization in emergent languages. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4427–4442.
- Rahma Chaabouni, Eugene Kharitonov, Emmanuel Dupoux, and Marco Baroni. 2019. Anti-efficient encoding in emergent communication. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 6293–6303.
- Rahma Chaabouni, Florian Strub, Florent Alth  , Eugene Tarassov, Corentin Tallec, Elnaz Davoodi, Kory Wallace Mathewson, Olivier Tieleman, Angeliki Lazaridou, and Bilal Piot. 2022. Emergent communication at scale. In *International conference on learning representations*, pages 3184–3213.
- Huaben Chen, Wenkang Ji, Lufeng Xu, and Shiyu Zhao. 2023. Multi-agent consensus seeking via large language models. *arXiv preprint arXiv:2310.20151*.
- Jingdi Chen, Hanqing Yang, Zongjun Liu, and Carlee Joe-Wong. 2026. The five ws of multi-agent communication: Who talks to whom, when, what, and why—a survey from marl to emergent language and llms. *arXiv preprint arXiv:2602.11583*.
- Junzhe Chen, Xuming Hu, Shuodi Liu, Shiyu Huang, Wei-Wei Tu, Zhaofeng He, and Lijie Wen. 2024a. Llmarena: Assessing capabilities of large language models in dynamic multi-agent environments. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13055–13077.
- Weize Chen, Chenfei Yuan, Jiarui Yuan, Yusheng Su, Chen Qian, Cheng Yang, Ruobing Xie, Zhiyuan Liu, and Maosong Sun. 2024b. Beyond natural language: Llms leveraging alternative formats for enhanced reasoning and communication. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10626–10641.
- Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and 1 others. 2024c. Tombench: Benchmarking theory of mind in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15959–15983.
- Amit K Chopra, Matteo Baldoni, Samuel H Christie, and Munindar P Singh. 2025. Azorus: Commitments over protocols for bdi agents. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*, pages 490–499.
- Dylan Cope and Peter McBurney. 2022. Joining the conversation: Towards language acquisition for ad hoc team play. In *International Conference on Learning Representations 2022: Workshop on Emergent Communication*.
- Dylan Cope and Peter McBurney. 2024. Learning translations: emergent communication pretraining for cooperative language acquisition. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 40–48.
- Dylan Cope and Peter McBurney. 2025. Decoding communications with partial information. In *Artificial Life Conference Proceedings 37*, volume 2025, page 58.
- Dylan Cope and Nandi Schoots. 2024. Channel randomisation methods for zero-shot communication. In *Proceedings of the 27th European Conference on Artificial Intelligence*, pages 3620–3627.
- Abhishek Das, Th  ophile Gerv  , Joshua Romoff, Dhruv Batra, Devi Parikh, Mike Rabbat, and Joelle Pineau. 2019. Tarmac: Targeted multi-agent communication. In *International Conference on machine learning*, pages 1538–1546. PMLR.
- Wentao Deng, Jiahuan Pei, Zhiwei Xu, Zhaochun Ren, Zhumin Chen, and Pengjie Ren. 2026. Belief-calibrated multi-agent consensus seeking for complex nlp tasks. In *Proceedings of the International Conference on Neural Information Processing Systems*, volume 38, pages 116412–116445.
- Ling Ding, Tianbai Lyu, Zhiliang Bi, Hao Wang, Shanshan Feng, and Wei Yu. 2026. Communication-efficient multi-agent reinforcement learning with spatiotemporal information hub. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 20799–20807.
- Wei Du, Benyu Wu, Yuqing Sun, Wei Guo, Yuntao Du, Zhongmin Yan, Guoxian Yu, and Lizhen Cui. 2026. Learning efficient and interpretable multi-agent communication. In *The Fourteenth International Conference on Learning Representations*.
- Rafael Elberg, Felipe del Rio, Mircea Petrache, and Denis Parra. 2026. A compressive-expressive communication framework for compositional representations. In *Proceedings of International Conference on Neural Information Processing Systems*, volume 38, pages 35562–35603.

- Charles Fleming, Ramana Kompella, Peter Bosch, and Vijoy Pandey. 2026. Scaling multi-agent systems: A smart middleware for improving agent interactions. *arXiv preprint arXiv:2604.03430*.
- Charles Fleming, Luca Muscariello, Vijoy Pandey, and Ramana Kompella. 2025. A layered protocol architecture for the internet of agents. *arXiv preprint arXiv:2511.19699*.
- Tianyu Fu, Zihan Min, Hanling Zhang, Jichao Yan, Guohao Dai, Wanli Ouyang, and Yu Wang. 2025. Cache-to-cache: Direct semantic communication between large language models. *arXiv preprint arXiv:2510.03215*.
- Zihong Gao, Hongjian Liang, Lei Hao, and Liangjun Ke. 2026. Decomposing communication gain and delay cost under cross-timestep delays in cooperative multi-agent reinforcement learning. *arXiv preprint arXiv:2604.03785*.
- Saibo Geng, Hudson Cooper, Michał Moskal, Samuel Jenkins, Julian Berman, Nathan Ranchin, Robert West, Eric Horvitz, and Harsha Nori. 2025. Generating structured outputs from language models: Benchmark and studies. *arXiv e-prints*, pages arXiv–2501.
- Laura Harding Graesser, Kyunghyun Cho, and Douwe Kiela. 2019. Emergent linguistic phenomena in multi-agent communication games. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3700–3710.
- Deniz Gündüz, Zhijin Qin, Inaki Estella Aguerri, Harpreet S Dhillon, Zhaohui Yang, Aylin Yener, Kai Kit Wong, and Chan-Byoung Chae. 2022. Beyond transmitting bits: Context, semantics, and task-oriented communications. *IEEE Journal on Selected Areas in Communications*, 41(1):5–41.
- Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xi-angliang Zhang. 2024. Large language model based multi-agents: a survey of progress and challenges. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 8048–8057.
- Abhinav Gupta, Marc Lanctot, and Angeliki Lazaridou. 2021. Dynamic population-based meta-learning for multi-agent communication with natural language. In *Proceedings of the International Conference on Neural Information Processing Systems*, volume 34, pages 16899–16912.
- Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, and 1 others. 2023. Metagpt: Meta programming for a multi-agent collaborative framework. In *Proceedings of the twelfth international conference on learning representations*, pages 35980–36008.
- Hengyuan Hu, Adam Lerer, Alex Peysakhovich, and Jakob Foerster. 2020. "other-play" for zero-shot coordination. In *Proceedings of the 37th International Conference on Machine Learning*, pages 4399–4410.
- Jiangping Huang, Dongmin Jin, Weisong Sun, Yang Liu, and Zhi Jin. 2025. Envisioning intelligent requirements engineering via knowledge-guided multi-agent collaboration. In *Proceedings of the 40th International Conference on Automated Software Engineering-New Ideas and Emerging Results Track*, pages 1–6. IEEE Computer Society.
- Dom Huh and Prasant Mohapatra. 2025. Grounding natural language for multi-agent decision-making with multi-agentic llms. *arXiv preprint arXiv:2508.07466*.
- Natasha Jaques, Angeliki Lazaridou, Edward Hughes, Caglar Gulcehre, Pedro Ortega, DJ Strouse, Joel Z Leibo, and Nando De Freitas. 2019. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *International conference on machine learning*, pages 3040–3049. PMLR.
- Jiechuan Jiang and Zongqing Lu. 2018. Learning attentional communication for multi-agent cooperation. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 7265–7275.
- Dayeon Ki, Rachel Rudinger, Tianyi Zhou, and Marine Carpuat. 2025. Multiple llm agents debate for equitable cultural alignment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24841–24877.
- Daewoo Kim, Sangwoo Moon, David Hostallero, Wan Ju Kang, Taeyoung Lee, Kyunghwan Son, and Yung Yi. 2019. Learning to schedule communication in multi-agent reinforcement learning. In *International Conference on Learning Representations*, pages 6458–6474.
- Hyungmin Kim, Minsu Jang, Jaehong Kim, and 1 others. 2025. Mind-map agent: Enhancing cooperative task planning through communication alignment with large language models. In *NeurIPS 2025 Workshop on Bridging Language, Agent, and World Models for Reasoning and Planning*.
- Hyuntak Kim and Byung-Hak Kim. 2025. Nexussum: Hierarchical llm agents for long-form narrative summarization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10120–10157.
- Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Le Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2023. Fantom: A benchmark for stress-testing machine theory of mind in interactions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP 2023*, pages 14397–14413.

- Jooyeon Kim and Alice Haeyun Oh. 2021. Emergent communication under varying group sizes and connectivities. In *Proceedings of the 35th Conference on Neural Information Processing Systems, NeurIPS 2021*, pages 17579–17591.
- Kristina Kobrock, Xenia Ohmer, Elia Bruni, and Nicole Gotzner. 2025. Agents generalize to novel levels of abstraction by using adaptive linguistic strategies. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8685–8699.
- Dezhang Kong, Shi Lin, Zhenhua Xu, Zhebo Wang, Minghao Li, Yufeng Li, Yilun Zhang, Hujin Peng, Xiang Chen, Zeyang Sha, and 1 others. 2025. A survey of llm-driven ai agent communication: Protocols, security risks, and defense countermeasures. *arXiv preprint arXiv:2506.19676*.
- T Kouwenhoven, MCA Peepkorn, RE de Kleijn, T Verhoef, and J Kwok. 2025a. Shaping shared languages: human and large language models’ inductive biases in emergent communication. In *International Joint Conference on Artificial Intelligence*, pages 10298–10306.
- Tom Kouwenhoven, Max Peepkorn, and Tessa Verhoef. 2025b. Searching for structure: Investigating emergent communication with large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9977–9991.
- Heeyoung Lee. 2024. One-to-many communication and compositionality in emergent communication. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20794–20811.
- Jason Lee, Kyunghyun Cho, Jason Weston, and Douwe Kiela. 2018. Emergent translation in multi-agent communication. In *International Conference on Learning Representations*, pages 5135–5152.
- Ido Levy, Orr Paradise, Boaz Carmeli, Ron Meir, Shafi Goldwasser, and Yonatan Belinkov. 2025. Unsupervised translation of emergent communication. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 23231–23239.
- David Lewis. 1969. *Convention: A Philosophical Study*. Harvard University Press, Cambridge, MA.
- Fushan Li and Michael Bowling. 2019. Ease-of-teaching and language structure from emergent communication. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, pages 15851–15861.
- Hua Li, Hossein Nourkhiz Mahjoub, Behdad Chalaki, Vaishnav Tadiparthi, Kwongjoon Lee, Ehsan Moradi-Pari, Michael Lewis, and Katia Sycara. 2024a. Language grounded multi-agent reinforcement learning with human-interpretable communication. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pages 87908–87933.
- Xin Li, Mengbing Liu, and Chau Yuen. 2025. Llm agent communication protocol (lap) requires urgent standardization: A telecom-inspired protocol is necessary. In *NeurIPS 2025 Workshop: AI and ML for Next-Generation Wireless Communications and Networking*.
- Xinyi Li, Sai Wang, Siqi Zeng, Yu Wu, and Yi Yang. 2024b. A survey on llm-based multi-agent systems: workflow, infrastructure, and challenges. *Vicinity*, 1(1):9.
- Toru Lin, Minyoung Huh, Chris Stauffer, Ser-Nam Lim, and Phillip Isola. 2021. Learning to ground multi-agent communication with autoencoders. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, pages 15230–15242.
- Bingyi Liu, Weizhen Han, Enshu Wang, Xin Ma, Shengwu Xiong, Chunming Qiao, and Jianping Wang. 2021. An efficient message dissemination scheme for cooperative drivings via multi-agent hierarchical attention reinforcement learning. In *Proceedings of the 41st International Conference on Distributed Computing Systems (ICDCS)*, pages 326–336.
- Haochen Liu, Li Chen, Yu Qiao, Chen Lv, and Hongyang Li. 2024a. Reasoning multi-agent behavioral topology for interactive autonomous driving. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pages 92605–92637.
- Jie Liu, Guohua Wang, Ronghui Yang, Jiajie Zeng, Mengchen Zhao, and Yi Cai. 2025. Rtadev: Intention aligned multi-agent framework for software development. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 1548–1581.
- Yuhan Liu, Yuyang Huang, Jiayi Yao, Shaoting Feng, Zhuohan Gu, Kuntai Du, Hanchen Li, Yihua Cheng, Junchen Jiang, Shan Lu, and 1 others. 2024b. Droid-speak: Kv cache sharing for cross-llm communication and multi-llm serving. *arXiv preprint arXiv:2411.02820*.
- Yat Long Lo and Biswa Sengupta. 2022. Learning to ground decentralized multi-agent communication with contrastive learning. In *Emergent Communication Workshop at ICLR 2022*.
- Yat Long Richie Lo, Biswa Sengupta, Jakob Foerster, and Michael Noukhovitch. 2024. Learning multi-agent communication with contrastive learning. In *International Conference on Learning Representations*, volume 2024, pages 51029–51046.
- Andrei Lupu, Brandon Cui, Hengyuan Hu, and Jakob Foerster. 2021. Trajectory diversity for zero-shot coordination. In *International conference on machine learning*, pages 7204–7213.
- Tianyi Ma, Yiyue Qian, Zheyuan Zhang, Zehong Wang, Xiaoye Qian, Feifan Bai, Yifan Ding, Xuwei Luo,

- Shinan Zhang, Keerthiram Murugesan, and 1 others. 2026. Autodata: A multi-agent system for open web data collection. In *Proceedings of the International Conference on Neural Information Processing Systems*, volume 38, pages 173416–173448.
- Andreas Madsen, Sarath Chandar, and Siva Reddy. 2024. Are self-explanations from large language models faithful? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 295–337.
- Fernando Martinez-Lopez, Tao Li, Yingdong Lu, and Juntao Chen. 2025. In-context reinforcement learning via communicative world models. *arXiv preprint arXiv:2508.06659*.
- Yohan Mathew, Ollie Matthews, Robert McCarthy, Joan Velja, Christian Schroeder de Witt, Dylan Cope, and Nandi Schoots. 2025. Hidden in plain text: Emergence & mitigation of steganographic collusion in llms. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 585–624.
- Paul Michel, Mathieu Rita, Kory Wallace Mathewson, Olivier Tieleman, and Angeliki Lazaridou. 2023. Re-visiting populations in multi-agent communication. In *Proceedings of the Eleventh International Conference on Learning Representations*, pages 27800–27815.
- Igor Mordatch and Pieter Abbeel. 2018. Emergence of grounded compositional language in multi-agent populations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, pages 1495–1502.
- Sumeet Ramesh Motwani, Mikhail Baranchuk, Martin Strohmeier, Vijay Bolina, Philip HS Torr, Lewis Hammond, and Christian Schroeder de Witt. 2024. Secret collusion among ai agents: multi-agent deception via steganography. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pages 73439–73486.
- Navid Naderializadeh, Jaroslaw J Sydir, Meryem Simsek, and Hosein Nikopour. 2021. Resource management in wireless networks via multi-agent deep reinforcement learning. *IEEE Transactions on Wireless Communications*, 20(6):3507–3523.
- Parinaz Naghizadeh, Maria Gorlatova, Andrew S Lan, and Mung Chiang. 2019. Hurts to be too early: Benefits and drawbacks of communication in multi-agent learning. In *Proceedings of the IEEE conference on computer communications*, pages 622–630.
- Siddharth Nayak, Adelmo Morrison Orozco, Marina Ten Have, Vittal Thirumalai, Jackson Zhang, Darren Chen, Aditya Kapoor, Eric Robinson, Karthik Gopalakrishnan, James Harrison, and 1 others. 2024. Long-horizon planning for multi-agent robots in partially observable environments. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, pages 67929–67967.
- Jens Nevens, Jonas Doumen, Paul Van Eecke, and Katrien Beuls. 2022. Language acquisition through intention reading and pattern finding. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 15–25.
- Letitia Parcalabescu and Anette Frank. 2024. On measuring faithfulness or self-consistency of natural language explanations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6048–6089.
- Joon Sung Park, Joseph O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22.
- Jannik Peters, Constantin Waubert de Puiseau, Hasan Tercan, Arya Gopikrishnan, Gustavo Adolpho Lucas de Carvalho, Christian Bitter, and Tobias Meisen. 2025. Emergent language: a survey and taxonomy. *Autonomous Agents and Multi-Agent Systems*, 39(1):18.
- Chau Pham, Boyi Liu, Yingxiang Yang, Zhengyu Chen, Tianyi Liu, Jianbo Yuan, Bryan A Plummer, Zhaoran Wang, and Hongxia Yang. 2024. Let models speak ciphers: Multiagent debate through embeddings. In *Proceedings of the 12th International Conference on Learning Representations*, pages 41704–41733.
- Runnan Qi, Yuming Quan, Yanan Ni, Zongyuan Li, Xiaojie Xu, Kuihua Huang, and Xian Guo. 2025. Comm-cot: Standardized chain-of-thought communication framework for efficient llm based multi-agent decision-making in real-time strategy games. In *2025 IEEE 2nd International Conference on Electronics, Communications and Intelligent Science (ECIS)*, pages 1–8.
- Chen Qian, Yufan Dang, Jiahao Li, Wei Liu, Zihao Xie, Yifei Wang, Weize Chen, Cheng Yang, Xin Cong, Xiaoyin Che, and 1 others. 2024. Experiential co-learning of software-developing agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5628–5640.
- Limor Raviv, Antje Meyer, and Shiri Lev-Ari. 2020. The role of social network structure in the emergence of linguistic structure. *Cognitive Science*, 44(8):e12876.
- Xinxing Ren, Caelum Forder, Qianbo Zang, Ahsen Tahir, Roman J Georgio, Suman Deb, Peter Carroll, Önder GÜRCAN, and Zekun Guo. 2025. Anemol: A semi-centralized multi-agent system based on agent-to-agent communication mcp server from coral protocol. In *NeurIPS 2025 Workshop on Bridging Language, Agent, and World Models for Reasoning and Planning*.

- Mathieu Rita, Florian Strub, Jean-Bastien Grill, Olivier Pietquin, and Emmanuel Dupoux. 2022a. On the role of population heterogeneity in emergent communication. In *Proceedings of the International Conference on Learning Representations*, pages 4698–5716.
- Mathieu Rita, Corentin Tallec, Paul Michel, Jean-Bastien Grill, Olivier Pietquin, Emmanuel Dupoux, and Florian Strub. 2022b. Emergent communication: generalization and overfitting in lewis games. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pages 1389–1404.
- Kai Ruan, Mowen Huang, Ji-Rong Wen, and Hao Sun. 2025. Benchmarking llms’ swarm intelligence. *arXiv preprint arXiv:2505.04364*.
- Mohamed Sana, Antonio De Domenico, Emilio Calvanese Strinati, and Antonio Clemente. 2020. Multi-agent deep reinforcement learning for distributed handover management in dense mmwave networks. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8976–8980.
- Bidipta Sarkar, Warren Xia, C Karen Liu, and Dorsa Sadigh. 2025. Training language models for social deduction with multi-agent reinforcement learning. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*, pages 1830–1839.
- Jiawei Shao, Yuyi Mao, and Jun Zhang. 2021. Learning task-oriented communication for edge inference: An information bottleneck approach. *IEEE Journal on Selected Areas in Communications*, 40(1):197–211.
- Xiangyu Shi, Marco Chiesa, Gerald Q Maguire Jr, and Dejan Kostic. 2025. Kvcomm: Enabling efficient llm communication through selective kv sharing. *arXiv preprint arXiv:2510.03346*.
- Zijing Shi, Meng Fang, Shunfeng Zheng, Shilong Deng, Ling Chen, and Yali Du. 2023. Cooperation on the fly: Exploring language agents for ad hoc teamwork in the avalon game. *arXiv preprint arXiv:2312.17515*.
- Yakov Pyotr Shkolnikov. 2026. Agent memory below the prompt: Persistent q4 kv cache for multi-agent llm inference on edge devices. *arXiv preprint arXiv:2603.04428*.
- Amanpreet Singh, Tushar Jain, and Sainbayar Sukhbaatar. 2019. Learning when to communicate at scale in multiagent cooperative and competitive tasks. In *International Conference on Learning Representations*, pages 3911–3926.
- DJ Strouse, Kevin R McKee, Matt Botvinick, Edward Hughes, and Richard Everett. 2021. Collaborating with humans without human data. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, pages 14502–14515.
- Sainbayar Sukhbaatar, Arthur Szlam, and Rob Fergus. 2016. Learning multiagent communication with backpropagation. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 2252–2260.
- Haochen Sun, Shuwen Zhang, Lujie Niu, Lei Ren, Hao Xu, Hao Fu, Fangkun Zhao, Caixia Yuan, and Xiaojie Wang. 2025. Collab-overcooked: Benchmarking and evaluating large language models as collaborative agents. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4922–4951.
- Anirudh Sundar, Sinead Williamson, Katherine Metcalf, Barry-John Theobald, Skyler Seto, and Masha Fedzechkina. 2025. Steering into new embedding spaces: Analyzing cross-lingual alignment induced by model interventions in multilingual language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2375–2401.
- Zhi Rui Tam, Cheng-Kuang Wu, Yi-Lin Tsai, Chieh-Yen Lin, Hung-yi Lee, and Yun-Nung Chen. 2024. Let me speak freely? a study on the impact of format restrictions on large language model performance. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1218–1236.
- Xunzhu Tang, Kisub Kim, Yewei Song, Cedric Lothritz, Bei Li, Saad Ezzini, Haoye Tian, Jacques Klein, and Tegawendé F Bissyandé. 2024. Codeagent: Autonomous communicative agents for code review. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11279–11313.
- Yichen Tang, Weihang Su, Yujia Zhou, Yiqun Liu, Min Zhang, Shaoping Ma, and Qingyao Ai. 2025. Augmenting multi-agent communication with state delta trajectory. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10230–10251.
- Jonathan D Thomas, Mihai Anca, Robert Piechocki, and 1 others. 2021. Multi-lingual agents through multi-headed neural networks. *arXiv preprint arXiv:2111.11129*.
- Maxime Toquebiau, Jae-Yun Jun, Faiz Benamar, and Nicolas Bredeche. 2025. Towards language-augmented multi-agent deep reinforcement learning. *arXiv preprint arXiv:2506.05236*.
- Khanh-Tung Tran, Dung Dao, Minh-Duong Nguyen, Quoc-Viet Pham, Barry O’Sullivan, and Hoang D Nguyen. 2025. Multi-agent collaboration mechanisms: A survey of llms. *arXiv preprint arXiv:2501.06322*.
- Mycal Tucker, Roger Levy, Julie Shah, and Noga Zaslavsky. 2022. Trading off utility, informativeness, and complexity in emergent communication. In *Proceedings of the 36th International Conference*

- on *Neural Information Processing Systems*, pages 22214–22228.
- Michele Tufano, Anisha Agarwal, Jinu Jang, Roshanak Zilouchian Moghaddam, and Neel Sundaresan. 2024. Autodev: Automated ai-driven development. *arXiv preprint arXiv:2403.08299*.
- Miles Turpin, Julian Michael, Ethan Perez, and Samuel R Bowman. 2023. Language models don’t always say what they think: unfaithful explanations in chain-of-thought prompting. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 74952–74965.
- Yanhu Wang, Ruidong Li, Changhong Wang, Jia Ye, Chenyuan Feng, and Shuaishuai Guo. 2025a. Collaborative learning for task-oriented semantic communications: Overcoming data mismatch between transceivers. *IEEE Open Journal of the Communications Society*.
- Zhao Wang, Sota Moriyama, Wei-Yao Wang, Briti Gangopadhyay, and Shingo Takamatsu. 2025b. Talk structurally, act hierarchically: A collaborative framework for llm multi-agent systems. *arXiv preprint arXiv:2502.11098*.
- Xinghai Wei, Tingting Yuan, Jie Yuan, Dongxiao Liu, and Xiaoming Fu. 2025. Rescom: Reward-shaped curriculum for efficient multi-agent communication learning. In *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems*, pages 2153–2161.
- Henry M Wellman. 2011. Developing a theory of mind. *The Wiley-Blackwell handbook of childhood cognitive development*, 2:258–284.
- Shiguang Wu, Yaqing Wang, and Quanming Yao. 2025. Dense communication between language models. *arXiv preprint arXiv:2505.12741*.
- Huiqiang Xie, Zhijin Qin, Xiaoming Tao, and Khaled B Letaief. 2022. Task-oriented multi-user semantic communications. *IEEE Journal on Selected Areas in Communications*, 40(9):2584–2597.
- Frank F Xu, Yufan Song, Boxuan Li, Yuxuan Tang, Kri-tanjali Jain, Mengxue Bao, Zora Z Wang, Xuhui Zhou, Zhitong Guo, Murong Cao, and 1 others. 2024. Theagentcompany: Benchmarking llm agents on consequential real world tasks. *arXiv preprint arXiv:2412.14161*.
- Bingyu Yan, Zhibo Zhou, Litian Zhang, Lian Zhang, Ziyi Zhou, Dezhuang Miao, Zhoujun Li, Chaozhao Li, and Xiaoming Zhang. 2025. Beyond self-talk: A communication-centric survey of llm-based multi-agent systems. *arXiv preprint arXiv:2502.14321*.
- Xue Yan, Jiaxian Guo, Xingzhou Lou, Jun Wang, Haifeng Zhang, and Yali Du. 2023. An efficient end-to-end training approach for zero-shot human-ai coordination. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 2636–2658.
- Zhangyue Yin, Qiushi Sun, Cheng Chang, Qipeng Guo, Junqi Dai, Xuan-Jing Huang, and Xipeng Qiu. 2023. Exchange-of-thought: Enhancing large language model capabilities through cross-model communication. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15135–15153.
- Dun Yuan, Fuyuan Lyu, Ye Yuan, Weixu Zhang, Bowei He, Jiayi Geng, Linfeng Du, Zipeng Sun, Yankai Chen, Changjiang Han, and 1 others. 2026. Beyond message passing: A semantic view of agent communication protocols. *arXiv e-prints*, pages arXiv–2604.
- Tingting Yuan, Hwei-Ming Chung, Jie Yuan, and Xiaoming Fu. 2023. Dacom: Learning delay-aware communication for multi-agent reinforcement learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 11763–11771.
- Ceyao Zhang, Kaijie Yang, Siyi Hu, Zihao Wang, Guanghe Li, Yihang Sun, Cheng Zhang, Zhaowei Zhang, Anji Liu, Song-Chun Zhu, and 1 others. 2024a. Proagent: building proactive cooperative agents with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17591–17599.
- Guangyi Zhang, Hanlei Li, Yunlong Cai, Qiyu Hu, Guanding Yu, and Runmin Zhang. 2025a. Learned image transmission with hierarchical variational autoencoder. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 13215–13223.
- Jintian Zhang, Xin Xu, Ningyu Zhang, Ruibo Liu, Bryan Hooi, and Shumin Deng. 2024b. Exploring collaboration mechanisms for llm agents: A social psychology view. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14544–14607.
- Yiran Zhang, Ruiyin Li, Peng Liang, Weisong Sun, and Yang Liu. 2025b. Knowledge-based multi-agent framework for automated software architecture design. In *Proceedings of the 33rd ACM International Conference on the Foundations of Software Engineering*, pages 530–534.
- Rui Zhao, Jinming Song, Yufeng Yuan, Haifeng Hu, Yang Gao, Yi Wu, Zhongqian Sun, and Wei Yang. 2023. Maximum entropy population-based training for zero-shot human-ai coordination. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6145–6153.
- Weihua Zheng, Chang Liu, Zhengyuan Liu, Xin Huang, Kui Wu, Muhammad Huzaifah Md Shahrin, Aiti Aw, and Roy Ka-Wei Lee. 2026. Bridging linguistic gaps: Cross-lingual mapping in pre-training and dataset for enhanced multilingual llm performance. *arXiv preprint arXiv:2604.10590*.
- Yujia Zheng, Zhuokai Zhao, Zijian Li, Yaqi Xie, Mingze Gao, Lizhu Zhang, and Kun Zhang. 2025. Thought

communication in multiagent collaboration. *arXiv preprint arXiv:2510.20733*.

Changxi Zhu, Mehdi Dastani, and Shihan Wang. 2024. A survey of multi-agent deep reinforcement learning with communication. *Autonomous Agents and Multi-Agent Systems*, 38(1):4.

Hao Zhu, Graham Neubig, and Yonatan Bisk. 2021. Few-shot language coordination by modeling theory of mind. In *Proceedings of the International Conference on Machine Learning*, pages 12901–12911.

Kunlun Zhu, Hongyi Du, Zhaochen Hong, Xiaocheng Yang, Shuyi Guo, Daisy Zhe Wang, Zhenhailong Wang, Cheng Qian, Robert Tang, Heng Ji, and 1 others. 2025. Multiagentbench: Evaluating the collaboration and competition of llm agents. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8580–8622.

A Appendix

A.1 Overview of Related Surveys

We summarize existing surveys related to multi-agent communication in Table 4. This paper offers an in-depth treatment of convention consistency, a critical property for scalable multi-agent communication that has been largely overlooked in prior surveys, thereby addressing an important gap in the literature. Compared with existing surveys, our review further broadens the scope by covering not only emergent communication in MARL and natural-language communication in LLM-based MAS, but also latent communication, an emerging communication paradigm.

A.2 Comparison of Reviewed Communication Paradigms

Table 5 presents a comparative summary of the three communication paradigms reviewed in this paper. The comparison suggests that communication in multi-agent systems has continually evolved under a fundamental trade-off between semantic fidelity and convention consistency.

A.3 Summary Tables of Reviewed Work

Table 6 and Table 7 summarize the methods reviewed in this paper, covering their core ideas, strengths, limitations, and key features.

A.4 Benchmarks and Evaluation Metrics

A.4.1 Benchmarks

We summarize a set of benchmarks that support the evaluation of communication alignment in LLM MAS. Unlike static Q/A benchmarks such as MMLU and HotpotQA, these benchmarks provide interactive environments in which agents engage in multi-turn interactions. This enables researchers to observe how communication conventions emerge, stabilize, and evolve over extended interaction horizons.

Collab-Overcooked. Collab-Overcooked (Sun et al., 2025) is an LLM-based multi-agent benchmark built upon Overcooked-AI, where agents collaborate in interactive cooking environments through natural-language communication. It is particularly suitable for evaluating communication alignment, as it features diverse open-ended tasks and process-oriented metrics that capture goal interpretation, active collaboration, and continuous adaptation.

MultiAgentBench. MultiAgentBench (Zhu et al., 2025) is a comprehensive benchmark for evaluating LLM MAS across diverse interactive scenarios involving both collaboration and competition. Beyond final task success, it introduces metrics such as communication score and coordination score, making it useful for studying how communication quality influences multi-agent coordination.

TheAgentCompany. TheAgentCompany (Xu et al., 2024) constructs a realistic software-company environment in which agents complete tasks by browsing the web, writing code, and communicating with coworkers. This long-horizon collaborative setting makes it a strong benchmark for evaluating persistent common ground and the robustness of agent collaboration in realistic organizational contexts.

LLMArena. LLMArena (Chen et al., 2024a) provides a dynamic multi-agent environment comprising seven interactive game scenarios. It is well suited for assessing long-horizon agent collaboration, as its environments require models to reason under evolving states, adapt to other agents or opponents, and sustain effective communication across multi-turn interactions.

SwarmBench. SwarmBench (Ruan et al., 2025) evaluates the swarm intelligence of LLM-based agents under decentralized constraints. It provides a 2D grid-world environment where agents must collaborate through local communication with nearby peers. Notably, the authors find that agents gradually converge toward shared communication conventions over repeated interactions, a phenomenon closely related to the communication alignment discussed in this survey.

A.4.2 Evaluation Metrics

We summarize the main evaluation metrics adopted in prior work as follows.

Reward. Reward, or task success rate, is the most direct metric and has been widely used in both MARL settings and LLM-based multi-agent collaboration. However, raw reward signals may conflate the effect of communication with other task-related factors. To address this issue, some studies introduce a variant known as communication gain (Wei et al., 2025; Gao et al., 2026; Naghizadeh et al., 2019), which is computed as the difference in reward between agents operating with and without communication.

Semantic Reconstruction (Kim et al., 2023;

Table 4: Overview of existing surveys. A check mark (✓) indicates direct focus, a triangle (△) indicates partial but limited relevance, and a cross (×) indicates absence.

Survey Paper	MARL Emergent Communication	LLM Natural-Language Communication	LLM Latent Communication	Semantic Fidelity	Convention Consistency
Peters et al. (2025)	✓ It focuses on emergent language in MAS.	× Absent.	× Absent.	△ It discusses grounding and semantic evaluation.	△ It discusses generalization.
Brandizzi (2023)	✓ It reviews emergent communication research.	△ It mentions language models as motivation.	× Absent.	△ It discusses Theory of Mind modeling.	△ It discusses language drift.
Boldt and Mortensen (2024)	✓ It surveys deep emergent communication.	× Absent.	× Absent.	△ It discusses communication evaluation metrics.	△ It discusses language change and evolution.
Zhu et al. (2024)	✓ It surveys communication in MADRL.	× Absent.	× Absent.	△ It discusses useful semantics behind messages.	△ It discusses learnable communication.
Li et al. (2024b)	× Absent.	✓ It surveys LLM multi-agent interaction.	× Absent.	△ It discusses message delivery.	△ It discusses shared message pools for coordination.
Yuan et al. (2026)	× Absent.	✓ It surveys LLM-agent communication protocols.	× Absent.	✓ It discusses semantic repair and verification.	△ It discusses semantic alignment within protocols rather than alignment between agents.
Kong et al. (2025)	× Absent.	✓ It surveys LLM-driven agent communication.	× Absent.	△ It discusses ambiguity and semantic risks.	△ It discusses semantic alignment within protocols rather than alignment between agents.
Tran et al. (2025)	× Absent.	✓ It surveys LLM multi-agent collaboration.	× Absent.	△ It discusses semantic communication in 5G/6G networks.	△ It discusses common ground, not convention consistency.
Chen et al. (2026)	✓ It covers MARL and emergent language.	✓ It covers LLM-powered communication.	× Absent.	△ It discusses message interpretation through interaction.	△ It discusses common ground between agents.
Yan et al. (2025)	△ It gives traditional MAS as background.	✓ It surveys LLM-MAS communication.	× Absent.	△ It discusses how structured message improve information clarity.	△ It discusses the stability of protocols across agents.

Chen et al., 2024c). In LLMs, the speaker can be prompted to explicitly express its current intention in textual form, which can serve as an approximate substitute for its internal semantics. Similarly, the listener can be prompted to output its inferred intention of the speaker from the received message. Therefore, the reconstruction error defined in Eq. (1) can be approximated by measuring the textual discrepancy (usually measured by edit distance) between these two outputs, expressed as:

$$\mathbf{R}_{ij} \approx \mathbb{E}_{m_i} [d(a_i(\text{EXP}), a_j(\text{INF}(m_i)))], \quad (3)$$

where EXP and INF denote the prompts provided to the speaker for expressing its true semantics and to the listener for inferring the semantics from m_i , respectively.

Convention Symmetry (Lo et al., 2024; Lo and Sengupta, 2022). Convention symmetry estimates the degree of communication alignment by comparing the similarity between agents’ encoding strategies. Given the same observation o for two agents, let the messages generated by them be $m_i = E_i(o)$ and $m_j = E_j(o)$, respectively. The convention symmetry between the two agents can then be com-

puted as the expected similarity between these messages, i.e., $CS_{i,j} = \mathbb{E}_o[\text{sim}(m_i, m_j)]$. A higher convention symmetry indicates greater similarity between the agents’ encoders, suggesting that the agents are more likely to achieve alignment in communication.

Overall, metrics for evaluating communication alignment remain scarce, and existing ones suffer from substantial limitations. Specifically, reward cannot distinguish genuine semantic recovery from accidental coordination. Furthermore, semantic reconstruction compares free-form textual outputs directly, making the measurement potentially unreliable; convention symmetry only evaluates encoder-side and overlooks decoder-side alignment.

Table 5: Comparison of emergent communication, natural-language communication, and latent communication.

Communication paradigm	Communication substrate	Convention consistency	Semantic fidelity	Human interpretability	Typical failure mode
Emergent communication	Discrete or continuous protocols learned through interaction	Low. Learned conventions are often local to a training community and may not transfer well across independently trained agents	High. Communication can be compact and highly task-adaptive	Low. Messages are usually opaque to humans	Overfitting to local conventions; failure under new partners, tasks, or communities
Natural-language communication	natural-language text	High. Natural language provides a general interface for heterogeneous agents	Medium. It supports broad coordination but may introduce redundancy and communication overhead	High. Humans can inspect, debug, and intervene in communication	Ambiguous interpretation, hidden premises, accumulated bias, or consensus drift
Latent communication	Dense embeddings, KV caches, or model-internal representations	Low. Generalization is constrained by model architecture, representation compatibility, and cross-model alignment	High. It can preserve dense information and reduce lossy token-level communication	Low. Messages are usually opaque to humans	Cross-model incompatibility, privacy leakage, opaque coordination, extra alignment cost

Table 6: Summary of communication alignment methods in MARL.

Category	Core Idea	Pros & Cons	Method	Key Features
Convention-Invariant Learning	Introducing additional training pressure to make the agents' encoders and decoders more robust.	Pros: No additional training is required after deployment. Cons: Universal convention invariance is hard to achieve.	Cao et al. (2018)	Studies communication through grounded negotiation games.
			Rita et al. (2022b)	Analyzes co-adaptation and overfitting in Lewis games.
			Hu et al. (2020)	Uses symmetry randomization for zero-shot coordination.
			Bullard et al. (2021)	Discovers quasi-equivalent conventions under constraints.
			Yan et al. (2023)	Restricts mappings through external coordination constraints.
			Michel et al. (2023)	Reduces co-adaptation via isolated pair training.
			Cope and Schoots (2024)	Randomizes channels to avoid fixed encodings.
			Lo and Sengupta (2022)	Grounds decentralized messages using contrastive learning.
			Raviv et al. (2020)	Links social structure with language emergence.
			Bromham (2025)	Relates population size to language evolution.
			Mordatch and Abbeel (2018)	Studies grounded compositional language.
			Chaabouni et al. (2020)	Studies compositionality and language generalization.
			Strouse et al. (2021)	Trains with diverse partners for collaboration.
			Li and Bowling (2019)	Shows teaching pressure shapes language structure.
			Zhao et al. (2023)	Encourages diversity via maximum-entropy training.
			Lupu et al. (2021)	Promotes trajectory diversity for coordination.
			Gupta et al. (2021)	Uses dynamic population-based meta-learning.
			Chaabouni et al. (2022)	Examines communication at larger population scales.
			Rita et al. (2022b)	Studies heterogeneity in emergent communication.
			Kim and Oh (2021)	Examines group-size and connectivity effects.
Language Bridging	Introducing an intermediate mapping B between E_i and D_j , i.e., $D_j(B(E_i(z_i)))$.	Pros: Enables interoperability while preserving existing protocols. Cons: Introduces additional information loss and scaling overhead.	Toquebiau et al. (2025)	Adds natural language to MARL communication.
			Huh and Mohapatra (2025)	Grounds language for multi-agent decision-making.
			Levy et al. (2025)	Translates emergent languages without parallel data.
			Li et al. (2024a)	Grounds communication in interpretable language.
			Sundar et al. (2025)	Studies cross-lingual embedding-space alignment.
			Andreas et al. (2017)	Translates neural messages through shared beliefs.
			Lee et al. (2018)	Learns translation through visual grounding.
			Zheng et al. (2026)	Improves alignment via cross-lingual mapping.
			Lin et al. (2021)	Grounds communication with autoencoder representations.
			Cope and McBurney (2024)	Learns translations for language acquisition.
			Thomas et al. (2021)	Uses multi-headed multilingual agent networks.
			Kobrock et al. (2025)	Uses adaptive strategies across abstractions.
			Graesser et al. (2019)	Studies language contact in communication games.
			Martinez-Lopez et al. (2025)	Uses communicative world models for in-context RL.
Newcomer Adaptation	Updating the encoders and decoders of newly introduced agents via few-shot learning.	Pros: Supports onboarding with limited target-community interaction. Cons: Sample inefficient and unstable.	Cope and McBurney (2024)	Uses few-shot imitation and translation.
			Chaabouni et al. (2020)	Studies scalable language acquisition in populations.
			Jaques et al. (2019)	Uses social influence to motivate communication.
			Zhu et al. (2021)	Models partners' minds for few-shot coordination.
			Cope and McBurney (2022)	Frames acquisition as ad hoc teamwork.
			Cope and McBurney (2025)	Decodes communication under partial information.
			Nevens et al. (2022)	Combines intention reading with grammar abstraction.

Table 7: Summary of communication alignment methods in LLM MAS.

Category	Core Idea	Pros & Cons	Method	Key Features
Structured Message Design	Constraining the message space $\mathcal{L}$ to reduce ambiguity.	Pros: Simple to implement. Cons: Rigid formats may reduce expressiveness and scalability.	Wang et al. (2025b)	Studies structured messages in hierarchical multi-agent collaboration.
			Sun et al. (2025)	Divides agents' output into dedicated fields such as <i>plan</i> , <i>analysis</i> , and <i>say</i> .
Thought Exchange	Augment messages m with explicit intermediate reasoning r to reveal hidden beliefs, plans, and intentions.	Pros: Simple to implement. Cons: Reasoning may be incomplete or inconsistent.	Shi et al. (2023)	Studies ad hoc teamwork in Avalon-style communication.
			Hong et al. (2023)	Encodes SOPs into structured multi-agent workflows.
			Chen et al. (2024b)	Lets agents select task-adaptive communication formats.
			Fleming et al. (2026)	Middleware structures interaction through semantic grounding.
			Fleming et al. (2025)	Proposes layered protocols for agent communication.
			Ren et al. (2025)	Uses A2A primitives for semi-centralized coordination.
			Li et al. (2025)	Defines minimal message types with transactional guarantees.
			Geng et al. (2025)	Evaluates limitations of structured LLM outputs.
			Tam et al. (2024)	Shows format restrictions can harm LLM reasoning.
Thought Exchange			Qi et al. (2025)	Standardizes CoT communication for real-time decision-making.
			Yin et al. (2023)	Enables cross-model reasoning through thought exchange.
			Bi et al. (2025)	Adapts agent interaction topology to preserve informative reasoning content.
			Kim et al. (2025)	Allows agents to expose dialogue summary.
Contextual Consensus Building	Accumulate shared consensus G through multi-round interaction to improve agents' understanding of messages.	Pros: Supports long-horizon coordination and convention formation. Cons: High overhead, errors or biases may persist.	Chen et al. (2023)	Studies consensus seeking among LLM agents.
			Kouwenhoven et al. (2025b)	Studies structure emergence in LLM communication.
			Kouwenhoven et al. (2025a)	Compares human and LLM inductive biases.
			Ashery et al. (2025)	Models emergent conventions and collective bias.
			Ruan et al. (2025)	Shows that agents gradually converge toward a target messaging style through multi-turn interactions.
			Qian et al. (2024)	Enables agents to learn from past experiences.
			Zhang et al. (2024a)	Infers partner intentions for proactive cooperation.
			Park et al. (2023)	Uses memory, reflection, and planning for agents.
			Deng et al. (2026)	Calibrates beliefs to stabilize agent consensus.
			AlKhamissi et al. (2024)	Measures and improves cultural alignment in LLMs.
			Ki et al. (2025)	Uses multi-agent debate for equitable cultural alignment.
			Zhang et al. (2024b)	Analyzes collaboration from social psychology perspectives.
			Hong et al. (2023)	Stores shared artifacts for role-based coordination.
			Liu et al. (2025)	Verifies shared messages to prevent intent drift.
			Ma et al. (2026)	Uses hypergraphs to manage task-relevant artifacts.
			Kim and Kim (2025)	Compresses distributed communication into structured summaries.
Dense Embeddings Sharing	Share hidden embeddings instead of sampled natural-language tokens.	Pros: Preserves dense semantics beyond natural language. Cons: Requires cross-model alignment	Pham et al. (2024)	Uses raw transformer embeddings for agent communication.
			Wu et al. (2025)	Trains seq2seq modules for communication across different foundation LLMs.
			Tang et al. (2025)	Shares token-level state transition trajectories with text.
			Zheng et al. (2025)	Routes latent thought representations to relevant agents.
KV Cache Sharing	Share KV caches containing attention-level context.	Pros: Preserves richer context information. Cons: High overhead, introduces privacy leakage risks, and requires cross-model alignment.	Liu et al. (2024b)	Reuses selected KV caches to reduce loading cost.
			Shi et al. (2025)	Sends important KV entries via attention-based selection.
			Fu et al. (2025)	Projects KV caches across different foundation LLMs.
			Bian et al. (2025)	Provides KV-centric serving for multi-agent LLMs.
			Shkolnikov (2026)	Stores persistent quantized KV caches below prompts.