Communication Design for Autonomous Meta-Analysis: A Study of Multi-Agent LLM Communication

Anonymous ACL submission

Abstract

Communication is a central but underexamined design choice in multi-agent language systems. While prior work relies on fixed interaction patterns, it remains unclear how communication structure itself shapes reliability, efficiency, and explainability. We present a communication-centric analysis that isolates communication design within a fixed document-grounded meta-analysis pipeline. By varying communication topology, interaction protocol, and message constraints only during quality-control stages, we enable controlled comparison across task-compatible designs. Using task-grounded metrics, we show that communication structure induces fundamental trade-offs. Highly connected interaction maximizes error correction but incurs high coordination cost and diffuses responsibility, whereas structured topologies such as committee-based and hierarchical verification achieve *competitive* reliability at significantly lower cost Our results demonstrate that communication structure is a primary determinant of multi-agent reasoning behavior and should be treated as a methodological choice rather than an implementation detail.

1 Introduction

Large language models are increasingly deployed as multi-agent systems to address complex reasoning tasks that exceed the reliability of a single model instance. By decomposing problems across multiple agents and introducing interaction, prior work has demonstrated improvements in robustness, error correction, and calibration across tasks such as question answering, planning, and document-grounded reasoning (Wang et al., 2022; Du et al., 2023; Choi et al., 2025) , as well as interactive agent-based systems that model sustained behavior and decision-making over time (Park et al., 2023).

Despite this progress, the *design of inter-agent communication* is rarely treated as a methodological object of study. Most existing systems adopt a small number of implicit communication patterns—such as fully connected debate, centralized aggregation, or ad hoc critique loops—without isolating communication structure as an experimental variable (Li et al., 2023; Shinn et al., 2023; Madaan et al., 2023; Subramaniam et al., 2024). As a result, it remains unclear which properties of interaction actually drive improvements in reliability or explainability.

This gap is particularly consequential for tasks that demand verification, accountability, and auditability. In collaborative reasoning, communication does not merely transmit information; it determines how errors propagate, how corrections are triggered, and how responsibility for decisions is distributed (Kroll, 2020; Horneber and Laumer, 2023), particularly in domains where correctness is externally evaluated rather than preference-based (Zheng et al., 2023).

Two multi-agent systems with identical models, prompts, and task pipelines may exhibit markedly different error dynamics and explanation quality solely due to differences in their communication structure.

In this work, we argue that **communication design should be treated as a first-class methodological choice** in multi-agent language systems. Rather than asking how agents should phrase messages or how many dialogue rounds are required, we focus on a more structural question: how does the organization of inter-agent communication shape reliability, coordination cost, and explanation traces?

To study this question, we adopt a protocol-centric perspective that formalizes communication as an explicit control structure. Each protocol is characterized by a communication topology specifying which agents may interact, an interaction schedule determining when communication occurs, and an update rule describing how received in-

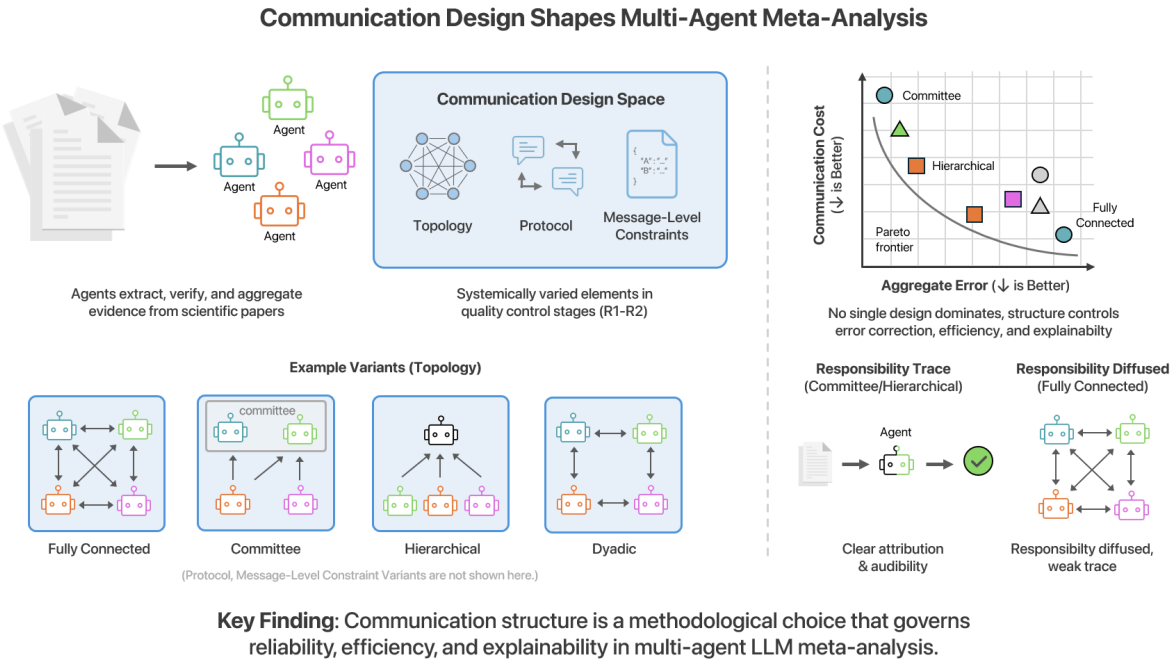

Figure 1: Communication design shapes multi-agent meta-analysis. Agents perform independent extraction and centralized aggregation in a fixed document-grounded pipeline, while only the communication structure used for verification and revision—including topology, protocol, and message-level constraints—is systematically varied, revealing trade-offs among reliability, communication cost, and responsibility attribution.

formation influences subsequent reasoning. Crucially, all other factors—including agent architecture, prompts, tools, and task definition—are held constant.

We instantiate this methodology in a document-grounded meta-analysis task, where agents must extract, verify, and aggregate structured evidence from scientific papers. Meta-analysis provides a particularly stringent testbed: it requires independent extraction, explicit verification, and centralized aggregation, and therefore sharply constrains which forms of interaction are task-compatible. Within this setting, we define a canonical multi-agent baseline that externalizes implicit human verification as explicit quality-control steps, and systematically vary communication topology, interaction protocol, and message constraints during verification.

Our results reveal that communication structure induces qualitatively distinct reasoning regimes. Some designs favor rapid convergence but amplify early extraction errors, while others preserve epistemic diversity at higher coordination cost. Notably, we find that explanation quality is closely tied to *responsibility structure* rather than message verbosity: protocols that preserve identifiable decision ownership produce explanations that are easier to audit, even when final accuracy is comparable.

Overall, this work makes three contributions. First, we formalize inter-agent communication as an explicit, executable design space rather than an implementation detail. Second, we provide a controlled empirical study isolating the effects of communication design on reliability, efficiency, and explainability in document-grounded reasoning. Third, we show that explanation traces emerge from communication structure itself, rather than from linguistic expressiveness or debate length.

By reframing communication as a methodological choice, this work provides a principled foundation for analyzing and designing multi-agent language systems beyond task-specific heuristics. Our findings further demonstrate that communication structure shapes both performance trade-offs and explanation traces, as visualized in Figures 3 and 4.

2 Related Work

2.1 Multi-Agent Reasoning with Language Models

A growing body of work explores multi-agent formulations of language model reasoning, including debate, self-consistency, critique–revision, and ensemble-style aggregation (Wang et al., 2022;

Du et al., 2023; Liang et al., 2024; Choi et al., 2025; Subramaniam et al., 2024; Li et al., 2023; Shinn et al., 2023; Madaan et al., 2023; Bai et al., 2022). These approaches demonstrate that interaction among agents can mitigate individual model biases and surface alternative reasoning paths.

However, in most prior work, communication is implemented using a fixed interaction pattern—typically fully connected dialogue or centralized aggregation. Comparisons are often limited to the presence or absence of interaction, rather than differences in communication structure (Yan et al., 2025; Du et al., 2025). As a result, the causal role of communication design remains underspecified.

In contrast, our work isolates communication topology, protocol, and message constraints as independent design dimensions, while holding agent architectures and task pipelines constant. This enables systematic analysis of how interaction structure shapes collective reasoning behavior.

2.2 Communication Topology in Multi-Agent Systems

The effects of communication topology have been extensively studied in distributed systems and classical multi-agent research. Centralized, decentralized, and hierarchical interaction structures are known to influence convergence speed, robustness, and fault tolerance (Lynch, 1996; Wooldridge, 2009; Olfati-Saber et al., 2007).

While these insights provide important theoretical foundations, they have largely been developed in settings without natural language reasoning or explicit explanation traces. Our work builds on this literature by translating communication topology into the context of language-based, document-grounded reasoning, where messages are semantic objects and errors manifest as structured extraction or verification failures.

2.3 Verification, Deliberation, and Collective Decision-Making

Research in verification, argumentation, and deliberative decision-making has examined how structured interaction supports error correction and accountability. Dyadic verification, committee-based review, and hierarchical adjudication are widely used in domains such as systematic reviews and scientific evaluation, and have also motivated semi-automated evidence synthesis pipelines (Marshall et al., 2017, 2020; Li et al., 2025).

We connect these ideas to multi-agent language systems by operationalizing verification and deliberation as executable communication protocols. Rather than modeling open-ended negotiation or persuasion, we focus on task-compatible quality-control interactions that preserve independent judgment and centralized aggregation.

2.4 Explainability in Collaborative Reasoning

Explainability in language models is often framed in terms of model internals or the linguistic quality of generated rationales, and more broadly in terms of interpretability and trust in learned systems (Ribeiro et al., 2016; Doshi-Velez and Kim, 2017). In multi-agent systems, explanations are frequently treated as post-hoc summaries of dialogue or aggregated reasoning steps.

We adopt a different perspective. Instead of viewing explanation as a property of individual messages, we treat it as an emergent property of communication structure. By analyzing how responsibility and revision rights are distributed across protocols, we show that explanation quality depends less on how much agents communicate and more on how decision authority is organized, aligning with broader notions of accountability in computer systems (Kroll, 2020; Horneber and Laumer, 2023). This perspective is also consistent with classical accounts of collective intelligence, which emphasize structured coordination and role differentiation over unstructured deliberation (Malone and Bernstein, 2015; Woolley et al., 2010).

2.5 Communication Without Explicit Correction

Several multi-agent frameworks emphasize information sharing, deliberation, or opinion aggregation without enforcing revision. Such designs are effective for exploration, brainstorming, or consensus-building, but do not support explicit error correction or responsibility attribution under external correctness constraints.

Our work intentionally excludes these settings. We focus on communication designs that are compatible with document-grounded meta-analysis, meaning they preserve independent extraction, permit explicit verification and revision, and produce structured outputs suitable for centralized aggregation. This restriction defines the feasible design space analyzed in this study, rather than an exhaustive taxonomy of multi-agent communication.

3 Methodology

Our study examines how communication design choices shape the behavior of multi-agent systems for automated meta-analysis, with a focus on reliability, efficiency, and explainability. We adopt AutoMETA (Ryu and Lee, 2025; Li et al., 2025) as a canonical reference framework and analyze the effects of communication design by systematically varying how agents interact during verification and revision, while holding all other system components fixed.

Figure 2 presents a protocol-centric decomposition of the AutoMETA pipeline. The system is organized into four stages: independent extraction (R0), verification (R1), revision (R2), and centralized aggregation (R3). Independent extraction and aggregation are fixed across all experimental conditions, while communication design choices are instantiated only during the intermediate stages (R1–R2). This decomposition allows observed differences in system behavior to be attributed directly to communication topology, interaction protocol, or message constraints, rather than to agent capability or task formulation.

3.1 Task Definition

We consider the task of automated meta-analysis, in which a system extracts structured information from multiple research papers and synthesizes these records into aggregate statistical summaries. The task imposes strict constraints: remain traceable throughout aggregation, reflecting broader concerns about interpretability and accountability in automated decision systems (Kroll, 2020; Doshi-Velez and Kim, 2017).

Following standard meta-analysis practice, we decompose the task into four stages. In R0, agents independently extract structured records from documents without interaction. In R1, extracted records may be inspected by peers for potential errors, inconsistencies, or missing evidence. In R2, original extractors revise their records in response to identified issues, producing finalized entries. In R3, these finalized records are pooled by a centralized aggregator to compute the final meta-analytic estimates.

Interaction is permitted only during R1–R2. No new information beyond the source documents is introduced, and all revisions must be explicitly justified.

3.2 Datasets

We use the same benchmark datasets introduced in AutoMETA. These datasets consist of collections of peer-reviewed studies commonly used in empirical meta-analyses, together with predefined extraction schemas and ground-truth aggregate statistics.

The datasets are well suited to our study because they require precise numerical extraction, exhibit heterogeneous reporting styles across papers, and demand transparent verification and correction. As a result, system performance is highly sensitive to how verification and revision are organized.

3.3 Canonical Baseline

We adopt AutoMETA as our canonical baseline configuration. In this baseline, agents independently extract structured records in R0, perform peer-based verification and revision in R1–R2, and submit finalized entries for centralized aggregation in R3.

Crucially, the baseline pipeline already includes two fixed components. First, verification and correction are implemented through a critique–revision mechanism, a pattern widely used in iterative self-correction and agent-based refinement (Shinn et al., 2023; Madaan et al., 2023). Second, all extractions and critiques are represented using a structured JSON schema with mandatory evidence pointers.

These components are treated as integral parts of the AutoMETA pipeline and are not considered experimental variants. Unless explicitly stated otherwise, all experimental conditions inherit this baseline verification mechanism and message schema.

3.4 Communication Design Space

Starting from the canonical baseline, we define a controlled communication design space along three orthogonal axes: communication topology, interaction protocol, and message constraints. Each experimental condition deviates from the baseline along exactly one axis, while all other components are held fixed. This one-axis-at-a-time design ensures interpretability and enables controlled comparison.

Topology ablations modify only the communication graph among agents during R1–R2, drawing on classical distinctions between centralized, decentralized, and hierarchical interaction structures in multi-agent systems (Lynch, 1996; Wooldridge, 2009). Protocol ablations alter the rules governing when and how verification or revision occurs, ex-

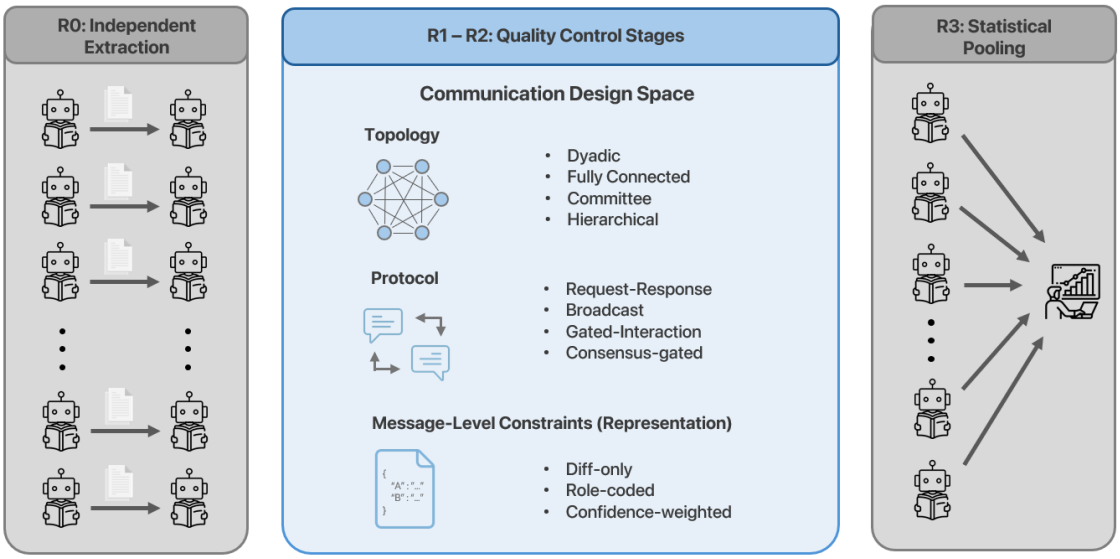

Figure 2: Protocol-centric decomposition of AutoMETA. Independent extraction (R0) and centralized aggregation (R3) are fixed, while communication design choices are applied only during the verification and revision stages (R1–R2).

tending prior work on structured agent interaction and decision protocols (Choi et al., 2025; Yan et al., 2025). Constraint ablations modify message-level constraints—i.e., what information agents are permitted or required to transmit during verification and revision—including the granularity of updates, semantic annotation of critiques, and conditional triggering of verification.

We restrict attention to communication designs that are compatible with automated meta-analysis. Specifically, all designs must preserve independent extraction, support explicit verification and revision, and produce structured outputs suitable for centralized aggregation. Interaction patterns that merely share information or aggregate opinions without enabling correction fall outside the scope of this study.

Table 1 summarizes all task-compatible design variants considered in our experiments. All variants preserve independent extraction (R0) and centralized aggregation (R3), and differ only in the design of verification and revision during R1–R2.

3.5 Experimental Control

All agents operate under identical language models, prompts, and document access. Decoding parameters and random seeds are fixed across all conditions. As a result, observed differences in performance arise solely from variations in communication topology, interaction protocol, or message constraints.

3.6 Evaluation Metrics

We evaluate system behavior using task-grounded metrics computed from structured outputs and interaction logs.

Aggregate Error. Aggregate Error measures the absolute deviation between the final meta-analytic estimates produced at R3 and the corresponding ground-truth values. This metric captures end-to-end correctness and reflects whether residual extraction or verification errors affect the final scientific conclusions.

Correction Rate. Correction Rate measures the proportion of extraction errors present after R0 that are corrected by the end of R2. This metric isolates the effect of interaction by focusing on changes induced during verification and revision.

Communication Cost. Communication Cost is measured as the total number of tokens exchanged during R1–R2, including critiques, responses, and revision messages. It captures the coordination overhead induced by a given communication design.

Attribution Completeness. Attribution Completeness measures the proportion of finalized fields and aggregate contributions that can be traced

Table 1: Task-compatible communication design variants. All variants share the same AutoMETA baseline (independent extraction, critique–revision, structured JSON, centralized aggregation) and differ only in how verification and revision are organized during R1–R2.

Variant	Description
Topology Variants	
Dyadic	Pairwise verification: each agent checks exactly one peer.
Committee	Restricted verification: a subset of agents acts as reviewers for all others.
Hierarchical	Multi-level verification: subgroup checks followed by higher-level review.
Protocol Variants	
Request–Response	Pull-based verification for uncertain fields only.
Broadcast	Push-based signaling of detected issues to all peers.
Gated Interaction	Verification triggered only under low confidence or detected conflicts.
Consensus-gated	Revised entries admitted to aggregation only after sufficient peer approval.
Message-Constraint Variants	
Diff-only	Only modified fields are transmitted during revision, with justification and evidence.
Role-coded	Critiques are annotated with semantic role labels (e.g., STAT, EVIDENCE, SCHEMA).
Confidence-weighted	Each value includes a confidence score used to trigger conditional verification.

to identifiable agents, evidence spans, and revision steps. This metric treats explainability as a structural property of the system rather than a linguistic one, aligning with system-level notions of accountability and responsibility assignment in automated decision-making (Kroll, 2020; Horneber and Laumer, 2023).

4 Results

We compare communication design variants against a canonical AutoMETA baseline. Results are grouped by design axis (communication topology, interaction protocol, and message-level constraints), with all other components held fixed. All values are averaged over three runs with identical decoding settings.

Aggregate Error denotes the mean relative error of final R3 meta-analytic estimates with respect to a canonical ground-truth meta-analysis, computed as $|R3 - GT|/|GT|$ and averaged across effect size, confidence bounds, and heterogeneity statistics. Lower Aggregate Error and Communication Cost indicate better performance, while higher Correction Rate and Attribution Completeness indicate more effective and interpretable revision.

4.1 Effects of Communication Topology

We first examine the effect of communication topology while holding the critique–revision protocol and message constraints fixed. Table 2 reports results for topology variants.

Changing communication topology does not uniformly improve Aggregate Error relative to the fully connected baseline. Instead, different topologies induce distinct trade-offs among calibration accuracy, coordination cost, and responsibility attribution.

Fully connected interaction enables broad cross-checking and sustained revision activity, but it does not yield the lowest Aggregate Error among topology variants. This suggests that dense mutual verification can increase revision activity without guaranteeing calibrated correction.

In contrast, committee-based and hierarchical topologies substantially reduce communication cost and improve attribution completeness, while trading off calibration accuracy. Overall, topology primarily determines how verification effort and revision responsibility are distributed, rather than serving as a direct lever for improving end-to-end calibration. Figure 3 visualizes this accuracy–efficiency trade-off.

4.2 Effects of Interaction Protocol

We next analyze interaction protocols while holding the canonical topology and message constraints fixed. Results are reported in Table 3.

Unlike topology, interaction protocol directly determines when and under what conditions revisions are permitted. Protocols that expose information without enforcing revision authority, such as broadcast and request–response, reduce communication cost but often yield higher Aggregate Error, indicating insufficient correction of biased intermediate estimates.

In contrast, gated and consensus-gated protocols explicitly regulate revision authority. Consensus-gated interaction achieves the lowest Aggregate Error across all settings, outperforming the canonical baseline. This result demonstrates that protocol-

Table 2: Effects of communication topology on calibration accuracy, communication cost, and attribution.

Topology	Aggregate Error ↓	Correction Rate ↑	Communication Cost ↓	Attribution Completeness ↑
Baseline (Fully Connected)	**0.315**	0.41	12.2k	0.071
Dyadic	0.491	0.53	**4.1k**	**0.160**
Committee	0.494	**0.61**	4.9k	0.105
Hierarchical	0.420	0.58	5.5k	0.108

Table 3: Effects of interaction protocol on calibration accuracy and attribution.

Protocol	Aggregate Error ↓	Correction Rate ↑	Communication Cost ↓	Attribution Completeness ↑
Baseline (Critique–Revision)	0.315	0.41	12.2k	0.071
Request–Response	0.545	0.05	**1.3k**	**0.400**
Broadcast	0.281	0.39	11.5k	0.158
Gated Interaction	0.345	0.49	11.1k	0.100
Consensus-gated	**0.226**	**0.55**	10.9k	0.211

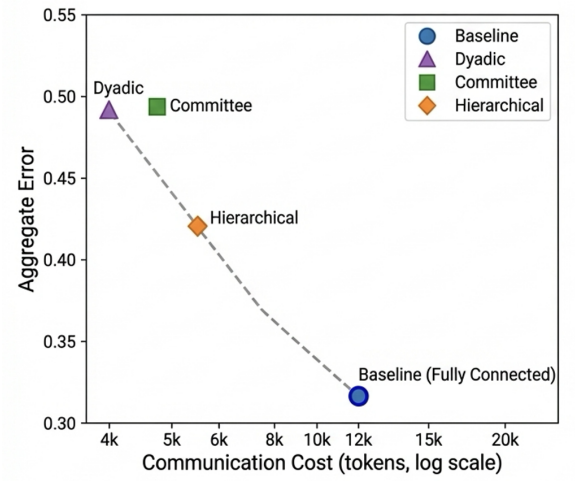

Figure 3: Trade-off between Aggregate Error (mean relative error w.r.t. ground truth) and communication cost across topology variants.

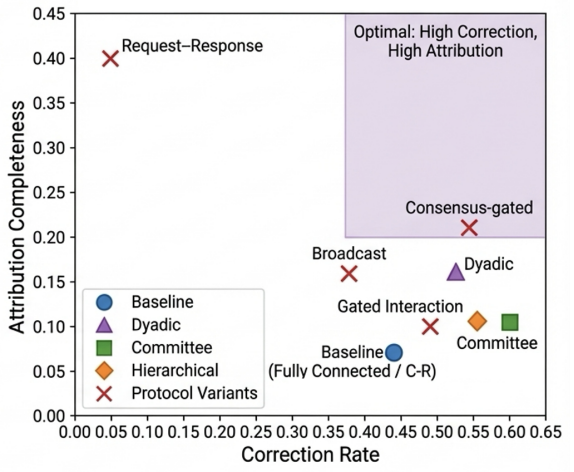

Figure 4: Relationship between correction rate and attribution completeness across communication designs.

level control over agreement and revision is critical for improving calibration, even when topology and agent capabilities are held constant.

Correction rate alone does not explain these outcomes. For example, request–response interaction exhibits high attribution completeness but the highest Aggregate Error, reflecting the absence of iterative correction.

4.3 Effects of Message-Level Constraints

Finally, we evaluate message-level constraints while holding topology and protocol fixed. Across all constraint variants, message-level constraints do not reduce Aggregate Error relative to the baseline.

While constraints affect communication cost and Attribution Completeness, these changes are not accompanied by consistent improvements in calibration accuracy. This indicates that syntactic restrictions on message content alone are insufficient to improve meta-analytic performance.

4.4 Explainability and Responsibility Structure

Figure 4 plots Correction Rate against Attribution Completeness across topology and protocol variants.

High correction rates do not necessarily correspond to high explainability. Fully connected interaction achieves frequent revision but exhibits low Attribution Completeness, indicating diffused responsibility.

In contrast, committee-based, hierarchical, and gated designs preserve identifiable revision pathways, achieving higher Attribution Completeness at comparable levels of Aggregate Error. These results show that explainability emerges from how responsibility is structurally allocated, rather than

from interaction volume.

Overall, the results indicate that explainability in multi-agent systems is not a byproduct of interaction volume or correction frequency. Instead, it emerges from how revision authority and responsibility are structurally allocated within the communication design.

5 Discussion

Our results demonstrate that communication design is a primary determinant of multi-agent system behavior, even when agent models, prompts, and task pipelines are held constant. Across all experiments, differences in reliability, efficiency, and explainability can be attributed to how interaction is structured, rather than to the presence or absence of communication itself.

Communication as Structured Error Control. Figure 3 shows that communication induces a fundamental trade-off between reliability and coordination cost. Importantly, the fully connected interaction maximizes error correction but incurs substantial overhead and dilutes responsibility, while structured topologies such as committee-based and hierarchical verification substantially reduce coordination cost and yield clearer responsibility attribution, at the expense of higher Aggregate Error. This suggests that communication primarily functions as a form of structured error control, whose effectiveness depends on how correction authority is distributed (Malone and Bernstein, 2015; Woolley et al., 2010).

Explainability Emerges from Responsibility Structure. Our analysis further reveals that explainability is not a byproduct of interaction volume or linguistic richness. As shown in Figure 4, protocols that preserve identifiable decision ownership yield clearer attribution traces, even when their final accuracy is similar to that of more interactive designs. This perspective aligns with system-level accounts of interpretability and accountability, which emphasize traceability and responsibility assignment over post-hoc rationalization (Kroll, 2020; Horneber and Laumer, 2023).

When Communication Fails. Notably, some protocols incur non-trivial communication cost without yielding consistent improvements in correction rate or aggregate error. These cases illustrate that communication alone is insufficient: error correction requires protocols that explicitly permit and structure revision. This finding challenges the common assumption that increased interaction necessarily improves multi-agent reasoning, echoing recent survey-level observations that communication structure, rather than volume, is the key driver of performance gains.

Implications for Multi-Agent System Design. Taken together, our findings suggest that effective multi-agent systems should be designed around explicit communication structures that balance error correction, coordination cost, and responsibility assignment. Rather than defaulting to unrestricted debate or fully connected interaction, designers should treat communication topology and protocol as first-class design choices aligned with task requirements.

6 Conclusion

We presented a protocol-centric analysis of communication design in multi-agent language systems. By isolating communication topology, interaction protocol, and message representation within a fixed automated meta-analysis pipeline, we demonstrated that communication structure alone induces distinct and systematic reasoning regimes, even when agent models, prompts, and task formulations are held constant.

Our results show that no single communication design dominates across reliability, efficiency, and explainability. Highly connected interaction, which serves as a common baseline in prior multi-agent systems, enables sustained revision activity but incurs substantial coordination cost and diffuses responsibility. In contrast, structured designs such as committee-based and hierarchical verification substantially reduce communication overhead and yield clearer attribution, while trading off calibration accuracy. These findings highlight that communication structure primarily governs how corrections are distributed and audited, rather than uniformly improving end-to-end calibration.

More broadly, this work reframes communication as a methodological design choice rather than an implementation detail. By articulating a controlled communication design space and introducing task-grounded metrics that jointly capture accuracy, cost, and responsibility, we provide a principled foundation for analyzing and engineering multi-agent language systems beyond ad hoc or fully connected interaction patterns.

7 Limitations

This work has several limitations. First, our analysis is restricted to communication protocols that explicitly support verification and revision. Protocols that only share information or aggregate opinions without enabling correction are outside the scope of this study.

Second, our experiments focus on a document-grounded meta-analysis task. While this task provides a stringent and well-defined testbed for studying communication design, the observed trade-offs may differ in tasks that prioritize open-ended generation, negotiation, or creative collaboration.

Third, we fix agent architectures, prompts, and decoding parameters in order to isolate the effects of communication design. Future work could explore interactions between communication structure and agent heterogeneity, learning dynamics, or adaptive protocol selection.

Our design space is also intentionally restricted to task-compatible variants under explicit verification and revision constraints. This enables controlled comparison, but it does not constitute an exhaustive taxonomy of multi-agent communication patterns, nor does it guarantee that the same trade-offs hold under looser interaction objectives.

In addition, we evaluate a fixed system scale and interaction budget. Topology and protocol effects may change with the number of agents, the amount of available communication, and deployment constraints such as synchronization and parallelism. A more complete characterization would sweep agent counts and budgets to study scaling behavior.

Our reported values are averaged over three runs, and we do not provide confidence intervals or formal significance tests. Given the stochasticity of language model outputs, future work should quantify variance more systematically, for example through additional random seeds or bootstrap estimates over documents.

Finally, our operational measures have limits. Communication cost is measured in exchanged tokens, which is an informative proxy but does not capture deployment-relevant costs such as latency, model-call counts, or monetary cost under specific inference APIs. Similarly, attribution completeness captures structural auditability and responsibility assignment, but it does not directly measure how explanations are perceived or used by human decision-makers in realistic review workflows.

References

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, and 1 others. 2022. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*.

Hyeong Kyu Choi, Xiaojin Zhu, and Sharon Li. 2025. Debate or vote: Which yields better decisions in multi-agent large language models? *arXiv preprint arXiv:2508.17536*.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Shangheng Du, Jiabao Zhao, Jinxin Shi, Zhentao Xie, Xin Jiang, Yanhong Bai, and Liang He. 2025. A survey on the optimization of large language model-based agents. *arXiv preprint arXiv:2503.12434*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first International Conference on Machine Learning*.

David Horneber and Sven Laumer. 2023. Algorithmic accountability. *Business & Information Systems Engineering*, 65(6):723–730.

Joshua A Kroll. 2020. Accountability in computer systems. *The Oxford handbook of ethics of AI*, pages 179–196.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008.

Lingbo Li, Anuradha Mathrani, and Teo Susnjak. 2025. Transforming evidence synthesis: A systematic review of the evolution of automated meta-analysis in the age of ai. *arXiv preprint arXiv:2504.20113*.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 17889–17904.

Nancy A Lynch. 1996. *Distributed algorithms*. Elsevier.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback, 2023. *URL https://arxiv.org/abs/2303.17651*.

Thomas W Malone and Michael Bernstein. 2015. *Handbook of collective intelligence*. MIT press.

Iain Marshall, Joël Kuiper, Edward Banner, and Byron C Wallace. 2017. Automating biomedical evidence synthesis: Robotreviewer. In *Proceedings of ACL 2017, system demonstrations*, pages 7–12.

Iain J Marshall, Blair T Johnson, Zigeng Wang, Sanguthevar Rajasekaran, and Byron C Wallace. 2020. Semi-automated evidence synthesis in health psychology: current methods and future prospects. *Health psychology review*, 14(1):145–158.

Reza Olfati-Saber, J Alex Fax, and Richard M Murray. 2007. Consensus and cooperation in networked multi-agent systems. *Proceedings of the IEEE*, 95(1):215–233.

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.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Kunhee Ryu and Keeheon Lee. 2025. AutoMETA: A multi-agent LLM system for autonomous meta-analysis. In *The 25th International Conference on Autonomous Agents and Multi-Agent Systems*.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652.

Vighnesh Subramaniam, Antonio Torralba, and Shuang Li. 2024. Debategpt: Fine-tuning large language models with multi-agent debate supervision.

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

Michael Wooldridge. 2009. *An introduction to multiagent systems*. John wiley & sons.

Anita Williams Woolley, Christopher F Chabris, Alex Pentland, Nada Hashmi, and Thomas W Malone. 2010. Evidence for a collective intelligence factor in the performance of human groups. *science*, 330(6004):686–688.

Bingyu Yan, Zhibo Zhou, Litian Zhang, Lian Zhang, Ziyi Zhou, Dezhuang Miao, Zhoujun Li, Chaozhuo Li, and Xiaoming Zhang. 2025. Beyond self-talk: A communication-centric survey of llm-based multi-agent systems. *arXiv preprint arXiv:2502.14321*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.