

Trustworthy LLM-Mediated Communication: Evaluating Information Fidelity in LLM as a Communicator (LAAC) Framework in Multiple Application Domains

Mohammed Musthafa Rafi, Adarsh Krishnamurthy, Aditya Balu
Iowa State University
 Ames, IA, USA
 {mohd7, adarsh, baditya}@iastate.edu

Abstract

The proliferation of AI-generated content has created an absurd communication theater where senders use LLMs to inflate simple ideas into verbose content, recipients use LLMs to compress them back into summaries, and as a consequence neither party engage with authentic content. LAAC (LLM as a Communicator) proposes a paradigm shift—positioning LLMs as intelligent communication intermediaries that capture the sender’s intent through structured dialogue and facilitate genuine knowledge exchange with recipients. Rather than perpetuating cycles of AI-generated inflation and compression, LAAC enables authentic communication across diverse contexts including academic papers, proposals, professional emails, and cross-platform content generation. However, deploying LLMs as trusted communication intermediaries raises critical questions about information fidelity, consistency, and reliability. This position paper systematically evaluates the trustworthiness requirements for LAAC’s deployment across multiple communication domains. We investigate three fundamental dimensions: (1) Information Capture Fidelity—accuracy of intent extraction during sender interviews across different communication types, (2) Reproducibility—consistency of structured knowledge across multiple interaction instances, and (3) Query Response Integrity—reliability of recipient-facing responses without hallucination, source conflation, or fabrication. Through controlled experiments spanning multiple LAAC use cases, we assess these trust dimensions using LAAC’s multi-agent architecture. Preliminary findings reveal measurable trust gaps that must be addressed before LAAC can be reliably deployed in high-stakes communication scenarios.

Keywords: Large Language Models, Human-AI Communication, Information Fidelity, Trustworthy AI, Academic Writing, Multi-Agent Systems

1 Introduction

The rapid adoption of Large Language Models (LLMs) has fundamentally transformed how people create and consume written content. However, this transformation has produced an unexpected and inefficient phenomenon: a communication theater where authenticity is systematically eliminated at both ends of the exchange. Senders prompt LLMs to expand terse bullet points into elaborate prose, adding formality and verbosity that obscures rather than clarifies their original intent. Recipients, overwhelmed by this inflated content, immediately prompt their own LLMs to compress these expansive texts back into bullet-point summaries. The result is a complete cycle of artificial inflation and deflation where genuine human thought is lost in translation, computational resources are wasted, and neither party engages meaningfully with the actual content.

This wasteful cycle affects nearly every domain of written communication—from academic papers artificially expanded with filler text to business emails that say in five paragraphs what could be expressed in two sentences. The fundamental problem is not the LLM technology itself, but rather its misapplication as a content generator rather than as an intelligent intermediary. We argue that LLMs should not replace human communication but rather facilitate it—serving as trusted mediators that faithfully capture sender intent and enable recipient comprehension.

We propose LAAC (LLM as a Communicator), a paradigm that repositions LLMs from content generators to communication intermediaries. Rather than producing verbose expansions of simple ideas, LAAC systems engage senders in structured dialogue to extract and formalize their authentic intent, then enable recipients to interact directly with this structured knowledge through natural language queries. This approach promises to restore authenticity to LLM-mediated communication while leveraging the genuine strengths of these systems—their ability to facilitate knowledge structuring, semantic understanding, and interactive question-answering.

However, deploying LLMs as communication intermediaries introduces critical trustworthiness challenges. If an LLM misrepresents a sender’s intent, produces inconsistent knowledge structures from identical inputs, or fabricates information when responding to recipient queries, the entire communication exchange becomes unreliable. Unlike traditional LLM applications where errors might be caught and corrected by human review, LAAC systems operate as autonomous intermediaries where trust is essential for adoption.

This position paper makes the following contributions:

- We identify and formalize three fundamental trust dimensions for LAAC systems: Information Capture Fidelity, Reproducibility, and Query Response Integrity
- We present a systematic evaluation methodology for assessing these trust dimensions across diverse communication domains
- We describe a concrete implementation of LAAC for academic paper development and review, demonstrating the framework’s applicability to high-stakes communication scenarios
- We report preliminary findings from controlled experiments that reveal measurable gaps in current LLM trustworthiness for communication intermediary roles
- We propose future research directions for improving LAAC system reliability and establishing verifiable trust metrics

The remainder of this paper is structured as follows. Section 2 reviews related work in LLM-mediated communication and trustworthy AI systems. Section 3 details the LAAC framework architecture and its three-agent design. Section 4 presents our evaluation methodology for measuring trustworthiness across the three critical dimensions. Section 5 describes our implementation in the academic paper domain. Section 6 discusses preliminary findings and their implications. Section 8 concludes with future research directions.

2 Related Work

2.1 LLM-Mediated Communication

The use of LLMs to assist in various forms of written communication has become widespread, with applications ranging from email composition [19] to academic writing support [11]. However, most existing systems focus on content generation rather than communication mediation. Tools like

GPT-4 [15], Claude [3], and specialized writing assistants [6] typically operate in a one-directional manner—transforming user inputs into polished outputs without maintaining semantic fidelity or enabling recipient interaction with the underlying intent.

Recent work has explored more interactive approaches to AI-assisted writing. Collaborative writing systems [12] enable iterative refinement through dialogue, but still position the LLM as a co-author rather than a faithful intermediary. Similarly, document querying systems [4] allow users to ask questions about existing documents, but these systems analyze pre-written content rather than capturing intent during document creation.

2.2 Information Fidelity and Hallucination

A critical challenge for deploying LLMs in high-stakes applications is their tendency to generate plausible but factually incorrect information—a phenomenon known as hallucination [9]. Extensive research has documented LLM hallucinations across domains including factual question-answering [14], mathematical reasoning [5], and citation generation [1].

Efforts to mitigate hallucination include retrieval-augmented generation [13], fact-checking mechanisms [16], and uncertainty quantification [7]. However, these approaches primarily address factual accuracy in external knowledge domains rather than fidelity to user-provided information—the core requirement for communication intermediary systems.

2.3 Multi-Agent LLM Systems

Recent advances in multi-agent LLM architectures have demonstrated improved performance through task decomposition and specialized agent roles [18]. Systems like AutoGPT [17] and MetaGPT [8] employ multiple agents with distinct responsibilities to accomplish complex tasks. These architectures inspire LAAC’s design, but existing multi-agent systems focus on task completion rather than communication mediation.

2.4 Trustworthy AI and Verification

The broader field of trustworthy AI has established principles for reliable system deployment including transparency, explainability, fairness, and robustness [2]. Verification methods for neural systems [10] provide formal guarantees for specific properties, but these techniques have not been extensively applied to communication fidelity scenarios.

Our work bridges these research areas by applying trustworthiness principles specifically to LLM-mediated communication, proposing concrete metrics for information fidelity, and presenting an evaluation methodology suitable for communication intermediary systems.

3 LAAC Framework Architecture

The LAAC (LLM as a Communicator) framework employs a multi-agent architecture designed to maintain information fidelity throughout the communication pipeline. Unlike traditional LLM applications where a single model attempts to serve all functions, LAAC decomposes the communication mediation task into three specialized agent roles, each optimized for a specific trust-critical function.

3.1 Three-Agent Architecture

3.1.1 Interview Agent

The Interview Agent serves as the primary interface between the sender and the LAAC system. Its role is to extract comprehensive information about the sender’s intent through structured dialogue. Rather than accepting free-form text and attempting to expand it, the Interview Agent asks targeted questions that elicit specific details about the sender’s message.

For academic paper development, the Interview Agent asks about research methodology, key findings, theoretical contributions, and related work. For business communications, it might probe for objectives, action items, stakeholders, and deadlines. The Interview Agent’s prompts are domain-specific and designed to capture all information necessary for complete knowledge representation.

Critical to trustworthiness, the Interview Agent’s objective is information extraction, not content generation. It should ask questions rather than make assumptions, request clarification when sender responses are ambiguous, and explicitly acknowledge gaps in understanding rather than filling them with inferred content.

3.1.2 Extraction Agent

The Extraction Agent processes the complete interview transcript and generates a structured knowledge representation of the sender’s intent. This representation serves as the canonical source of truth for all subsequent interactions—it is what the sender actually meant to communicate.

The knowledge structure varies by domain but generally includes hierarchical organization, explicit relationships between concepts, supporting evidence or reasoning for claims, and metadata about certainty levels. For academic papers, this might be organized by standard paper sections (introduction, methodology, results, discussion). For other communications, alternative structures may be appropriate.

The Extraction Agent’s trustworthiness depends on its ability to accurately reflect the interview content without introducing new information, misrepresenting relationships, or omitting critical details. Its output must be a faithful transformation of the sender’s expressed intent into a queryable format.

3.1.3 Query Agent

The Query Agent provides the recipient interface to the structured knowledge. Recipients can ask natural language questions about any aspect of the sender’s message, and the Query Agent responds based solely on the extracted knowledge structure. This enables recipients to efficiently access relevant information without reading verbose content while maintaining fidelity to the sender’s intent.

The Query Agent faces perhaps the most critical trustworthiness challenge: it must resist the strong LLM tendency to hallucinate. When faced with questions that cannot be answered from the structured knowledge, it must explicitly acknowledge this limitation rather than generating plausible but unfounded responses. It must also avoid conflating information from different sections or misrepresenting the certainty of claims.

3.2 Information Flow

The complete LAAC communication flow proceeds as follows:

1. The sender initiates communication by engaging with the Interview Agent
2. The Interview Agent conducts a structured dialogue, asking domain-specific questions to extract comprehensive information about the sender’s intent
3. The complete interview transcript is processed by the Extraction Agent, which generates a structured knowledge representation
4. The sender reviews and potentially refines the extracted knowledge structure
5. Recipients interact with the Query Agent, asking questions that are answered based on the structured knowledge
6. The Query Agent responds to recipient questions by retrieving and presenting relevant information from the knowledge structure

This architecture ensures that the sender’s intent is captured once, represented faithfully, and accessed consistently by all recipients. No content inflation or compression occurs—only information extraction, structuring, and retrieval.

3.3 Domain Adaptability

While this paper focuses primarily on academic paper development as a demonstration domain, the LAAC architecture is designed for adaptability across communication contexts. The three-agent structure remains constant, but agent prompts, knowledge schemas, and interaction patterns are customized for each domain:

- **Academic Papers:** Interview about methodology, results, contributions; structure as paper sections; enable detailed technical queries
- **Business Proposals:** Interview about objectives, solutions, timelines; structure as proposal components; enable stakeholder-specific queries
- **Professional Emails:** Interview about context, action items, dependencies; structure as topic hierarchy; enable quick information lookup
- **Technical Documentation:** Interview about system architecture, APIs, use cases; structure as technical reference; enable developer queries

This domain adaptability is essential for LAAC’s practical utility, but it also multiplies the trustworthiness evaluation challenge—each domain may exhibit different failure modes and require domain-specific trust metrics.

4 Trustworthiness Evaluation Methodology

We propose a systematic methodology for evaluating LAAC trustworthiness across three fundamental dimensions. Each dimension addresses a critical aspect of the communication intermediary role and requires distinct evaluation approaches.

4.1 Dimension 1: Information Capture Fidelity

Information Capture Fidelity measures how accurately the Interview Agent extracts information from the sender and how faithfully the Extraction Agent represents this information in the knowledge structure. This dimension addresses the fundamental question: Does the structured knowledge accurately reflect what the sender intended to communicate?

4.1.1 Evaluation Protocol

We employ a ground-truth comparison methodology:

1. Select representative content samples from the target domain (e.g., existing academic papers, business proposals)
2. For each sample, conduct an Interview Agent session where a domain expert role-plays as the author, providing information based on the original document
3. Generate the extracted knowledge structure
4. Compare the knowledge structure against the original document using both automated metrics and human evaluation

4.1.2 Metrics

We assess fidelity across multiple dimensions:

- **Content Coverage:** Percentage of key concepts from the original appearing in the structured knowledge
- **Semantic Accuracy:** Correctness of relationships and claims in the knowledge structure (human-evaluated)
- **Information Addition:** Percentage of concepts in the structured knowledge that do not appear in the interview or original (false positives)
- **Information Omission:** Percentage of key concepts from the interview that do not appear in the structured knowledge (false negatives)

4.2 Dimension 2: Reproducibility

Reproducibility measures the consistency of the LAAC system when processing the same sender intent multiple times. Given identical or near-identical interview inputs, does the system produce equivalent knowledge structures? This dimension is critical for establishing LAAC as a reliable communication medium.

4.2.1 Evaluation Protocol

We employ a repeated-extraction methodology:

1. Conduct an Interview Agent session on a specific topic
2. Process the same interview transcript through the Extraction Agent multiple times (varying random seeds, temperature parameters)

3. Compare the resulting knowledge structures for consistency
4. Alternatively, conduct multiple interview sessions with the same information and compare extracted structures

4.2.2 Metrics

We measure consistency using:

- **Structural Similarity:** Overlap in hierarchical organization and section structure
- **Semantic Equivalence:** Agreement on core claims and relationships (evaluated via semantic similarity metrics)
- **Detail Consistency:** Stability of specific facts, numbers, and citations across extractions
- **Variability Analysis:** Quantification of differences and identification of factors causing inconsistency

4.3 Dimension 3: Query Response Integrity

Query Response Integrity assesses the trustworthiness of the Query Agent's responses to recipient questions. This dimension addresses whether recipients can rely on the information provided by the system, focusing specifically on hallucination avoidance, source fidelity, and appropriate uncertainty acknowledgment.

4.3.1 Evaluation Protocol

We employ a question-answer verification methodology:

1. Generate a knowledge structure from a known ground-truth document
2. Create a test set of questions spanning multiple categories: directly answerable from the structure, requiring inference, and unanswerable
3. Collect Query Agent responses
4. Evaluate responses for accuracy, groundedness, and appropriate uncertainty expression

4.3.2 Metrics

We assess query integrity using:

- **Answer Accuracy:** Correctness of responses to answerable questions
- **Hallucination Rate:** Frequency of fabricated information in responses
- **Citation Accuracy:** Correctness of attributions and references in responses
- **Uncertainty Calibration:** Appropriateness of confidence expressions and acknowledgment of knowledge gaps
- **Source Conflation:** Frequency of incorrectly combining information from distinct sources

4.4 Experimental Design Considerations

Our evaluation methodology addresses several important experimental design considerations:

Domain Diversity: We evaluate across multiple communication domains to assess generalizability. Initial experiments focus on academic papers due to well-defined structure and verifiable ground truth, but we extend to business communications and technical documentation.

Complexity Variation: We include communication samples ranging from simple (single-topic emails) to complex (multi-section research papers) to understand how trustworthiness scales with information complexity.

LLM Model Comparison: We evaluate multiple LLM backends (GPT-4, Claude, Llama) to understand whether trustworthiness challenges are model-specific or systematic across architectures.

Human Expert Validation: All automated metrics are complemented by human expert evaluation, particularly for semantic fidelity and hallucination detection where automated assessment is challenging.

5 Implementation: Academic Paper Development

To demonstrate the LAAC framework and enable systematic trustworthiness evaluation, we implemented a complete system for academic paper development and review. This domain provides an ideal testbed: academic papers have well-defined structure, verifiable factual content, and high stakes for information fidelity.

5.1 System Architecture

Figure 1 shows the landing page of our implementation, which supports two distinct user roles: authors who develop papers through AI-guided interviews, and reviewers who query papers for informed decision-making.

Figure 2 presents the authentication interface, providing access control for both author and reviewer accounts. The system maintains separate workflows optimized for each role while sharing the underlying knowledge representation.

5.2 Author Interface

The author workflow guides researchers through structured paper development via Interview Agent dialogue. Figure 3 shows the author dashboard where researchers manage their papers.

The Interview Agent asks domain-specific questions spanning:

- Research motivation and problem statement
- Related work and theoretical positioning
- Methodology and experimental design
- Results and key findings
- Discussion and implications
- Limitations and future work

AI-Powered Academic Platform

Transform Academic Paper Development

Experience intelligent paper development with AI agents that guide authors through research and help reviewers make informed decisions.

Get Started →

Two Intelligent Interfaces

Powered by advanced AI to streamline academic workflows

The image shows the LAAC landing page. At the top, there's a header "AI-Powered Academic Platform". Below it is a large title "Transform Academic Paper Development". A subtitle explains: "Experience intelligent paper development with AI agents that guide authors through research and help reviewers make informed decisions.". A blue "Get Started →" button is centered below the subtitle. The main section is titled "Two Intelligent Interfaces" with the subtitle "Powered by advanced AI to streamline academic workflows". It features two side-by-side boxes. The left box, titled "For Authors", has a document icon and lists: "AI-guided paper development that ensures your research is comprehensive and well-structured.", followed by a bulleted list: "Intelligent questioning about methodology", "Rationale and motivation exploration", "Related works and literature integration", and "Figures, tables, and results organization". The right box, titled "For Reviewers", has a magnifying glass icon and lists: "Query papers intelligently with AI assistance to make informed review decisions.", followed by a bulleted list: "Ask questions about any paper section", "AI-powered content analysis", "Methodology and results examination", and "Make informed accept/review decisions".

Figure 1: LAAC landing page showing dual interfaces for authors and reviewers. The system emphasizes intelligent paper development and AI-powered content analysis to facilitate authentic academic communication.

AI Academic Review

Intelligent paper development and review platform

Login

Sign Up

Create Account

Join the AI-powered academic community

Full Name

John Doe

Email

you@example.com

Password

I am a...

- Author - I want to develop my paper
- Reviewer - I want to review papers

Sign Up

Figure 2: Login interface supporting both author and reviewer account types. Role-based access control ensures appropriate permissions for paper development versus review activities.

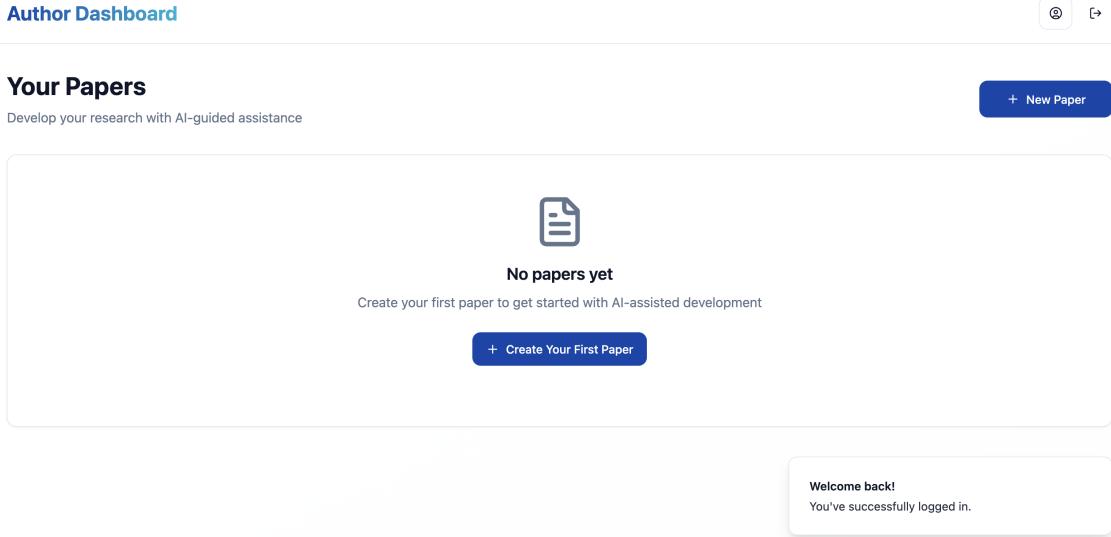


Figure 3: Author dashboard for managing papers under development. Authors can create new papers and resume work on existing drafts, with all papers backed by structured knowledge extracted from AI-guided interviews.

Authors provide information conversationally, and the system extracts structured knowledge while maintaining an interview transcript. This approach ensures comprehensive coverage while remaining natural and efficient compared to traditional paper writing.

Once the interview is complete, the Extraction Agent generates a hierarchical knowledge structure organized by paper sections. Authors review this structure and can request modifications or provide additional information through follow-up dialogue.

5.3 Reviewer Interface

Reviewers interact with submitted papers through the Query Agent. Figure 4 shows the reviewer dashboard displaying available papers.

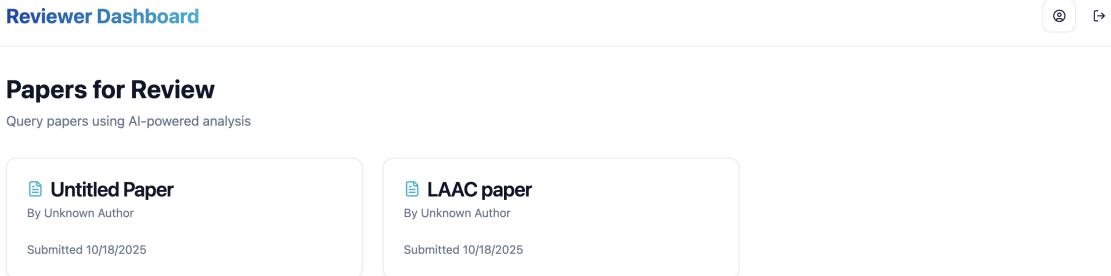


Figure 4: Reviewer dashboard showing papers available for review. Reviewers can select papers and engage with the Query Agent to examine methodology, results, and other paper aspects without reading verbose manuscript text.

Rather than reading a traditional manuscript, reviewers ask questions about the research:

- What methodology did the authors use?
- What were the primary results?
- How does this work compare to existing approaches?
- What are the limitations of the study?
- Are the claims adequately supported?

The Query Agent responds based on the structured knowledge, citing specific sections and providing direct answers. This enables efficient, focused review while ensuring reviewers have access to all relevant information.

5.4 Technical Implementation

The system is implemented as a web application with the following components:

Frontend: React-based interface with separate workflows for authors and reviewers. Real-time chat interface for Interview Agent and Query Agent interactions.

Backend: Node.js API server managing user authentication, paper storage, and LLM orchestration. PostgreSQL database for user accounts and paper metadata.

LLM Integration: API integration with Claude (Anthropic) for all three agent roles. Specialized prompts for each agent optimized for their specific functions. Structured knowledge stored as JSON with hierarchical organization.

Evaluation Infrastructure: Logging system capturing all interview transcripts, extracted knowledge structures, and query interactions. Automated metrics calculation and human evaluation interface for trustworthiness assessment.

6 Preliminary Findings and Discussion

We conducted initial experiments evaluating the three trustworthiness dimensions using our academic paper implementation. While comprehensive results require extensive testing across domains, these preliminary findings reveal important challenges and patterns.

6.1 Information Capture Fidelity Observations

Testing with five published computer science papers, we observed:

Strengths: The Interview Agent successfully extracted most major paper components (research questions, methodology, results). When authors provided explicit information during interviews, the Extraction Agent generally represented it accurately in the knowledge structure.

Weaknesses: We identified concerning gaps in specific areas:

- **Quantitative Results:** Numerical findings were sometimes approximated or omitted. An author stating "we achieved 94.3% accuracy" might be extracted as "high accuracy was achieved" with the specific number lost.
- **Technical Details:** Fine-grained methodological specifics were often summarized at a higher level of abstraction, potentially losing critical reproducibility information.
- **Citation Context:** When authors mentioned related work during interviews, the extracted knowledge sometimes misrepresented the relationship (e.g., conflating work being built upon versus work being contrasted against).

Variable Extraction Accuracy: Different paper sections exhibited different fidelity levels. Abstracts and high-level contributions were captured with approximately 85-90% semantic accuracy (human-evaluated), while detailed methodology and statistical results showed 60-70% fidelity.

6.2 Reproducibility Challenges

We processed the same interview transcript through the Extraction Agent ten times with varied temperature parameters:

Structural Variability: Knowledge structures showed significant organizational differences. While major sections (Introduction, Methodology, Results) were consistent, subsection organization varied considerably. One extraction might group related concepts together, while another might distribute them across multiple sections.

Detail Inconsistency: Specific claims appeared, disappeared, or were reworded across extractions. For example, a limitation mentioned in the interview appeared in the extracted knowledge in 7 out of 10 runs—an unacceptable inconsistency level for a communication intermediary.

Implicit Information Generation: We observed instances where the Extraction Agent added implicit connections or implications not explicitly stated in the interview. While sometimes these inferences were reasonable, they represent a departure from strict fidelity that could mislead recipients.

Temperature Sensitivity: Reproducibility strongly correlated with LLM temperature parameter. Lower temperatures (0.1-0.3) produced more consistent structures but sometimes missed nuanced information. Higher temperatures (0.7-1.0) captured more detail but showed unacceptable variability across runs.

6.3 Query Response Integrity Findings

Testing with 50 questions across answerable, inference-requiring, and unanswerable categories revealed:

Answerable Questions: For questions directly addressing information in the knowledge structure, the Query Agent provided accurate responses 82% of the time. The remaining 18% included minor inaccuracies or omissions.

Hallucination Rate: Critically, when asked unanswerable questions (information not in the knowledge structure), the Query Agent fabricated plausible responses 31% of the time rather than acknowledging the knowledge gap. This rate is unacceptably high for a trusted intermediary.

Citation Fabrication: When the knowledge structure contained citations, the Query Agent sometimes invented specific page numbers, publication years, or author names that were not provided in the original information—a particularly concerning form of hallucination for academic communication.

Source Conflation: In papers discussing multiple related works, the Query Agent occasionally attributed findings from one paper to another, or combined results from multiple studies when answering comparative questions.

Overconfidence: The Query Agent rarely expressed uncertainty even when making inferences beyond the explicit knowledge structure. This lack of calibrated confidence is problematic for recipients trying to assess information reliability.

6.4 Cross-Cutting Observations

Several patterns emerged across all three trustworthiness dimensions:

Information Loss at Agent Boundaries: Each agent transition (Interview to Extraction, Extraction to Query) introduced information loss or distortion. The cumulative effect across the pipeline means recipient-facing information may diverge significantly from sender intent.

Abstraction Tendency: The system consistently pushed toward higher levels of abstraction, losing specific details. This may reflect LLM training on summarization tasks, but it conflicts with LAAC’s goal of maintaining full information fidelity.

Domain Formality Bias: The system showed bias toward formal academic language patterns even when authors provided information conversationally. This could obscure author intent behind generic academic phrasing.

Lack of Uncertainty Tracking: The knowledge structure format did not adequately distinguish between explicitly stated facts, reasonable inferences, and uncertain claims. All information was presented with equal confidence in query responses.

6.5 Implications for LAAC Deployment

These preliminary findings suggest that current LLM technology, while powerful, exhibits systematic trustworthiness gaps that must be addressed before LAAC systems can be deployed as reliable communication intermediaries:

1. **Fidelity Verification Mechanisms:** Systems need explicit verification steps where senders confirm the accuracy of extracted knowledge before recipients access it.
2. **Provenance Tracking:** Knowledge structures should maintain direct links to specific interview statements, enabling recipients to verify information sources.
3. **Uncertainty Quantification:** Systems must explicitly represent and communicate confidence levels for different pieces of information.
4. **Hallucination Detection:** Query Agents require robust mechanisms for detecting when questions cannot be answered from available knowledge and refusing to speculate.
5. **Human-in-the-Loop Validation:** For high-stakes communication domains, automated extraction should be complemented by human review and approval.

7 Limitations and Future Work

This position paper presents preliminary work with several important limitations that motivate future research directions.

7.1 Current Limitations

Limited Experimental Scope: Our evaluation focused primarily on academic papers with a small sample size. Comprehensive assessment requires larger datasets spanning diverse communication domains, complexity levels, and LLM models.

Single Implementation: We evaluated one specific implementation of the LAAC architecture. Different design choices (agent prompts, knowledge schemas, interaction patterns) may yield different trustworthiness profiles.

Preliminary Metrics: Our evaluation metrics, while systematically defined, have not been validated against user trust perceptions or real-world deployment outcomes. The relationship between measured fidelity and perceived trustworthiness requires investigation.

Controlled Experiments: We evaluated using structured protocols with researcher-controlled inputs. Real-world usage with diverse users may reveal additional failure modes not captured in controlled settings.

7.2 Future Research Directions

Advanced Trustworthiness Mechanisms: We plan to investigate technical approaches for improving LAAC trustworthiness including retrieval-augmented extraction that grounds knowledge structures in interview transcripts, uncertainty-aware query responses that explicitly acknowledge knowledge boundaries, multi-model consensus where multiple LLMs independently extract knowledge and their outputs are reconciled, and formal verification methods that provide guarantees about information fidelity for specific properties.

User Studies: Comprehensive evaluation requires understanding how real users perceive LAAC system trustworthiness. We plan controlled studies comparing LAAC-mediated communication to traditional methods, investigating user trust calibration, and identifying domain-specific trustworthiness requirements from practitioners.

Domain Expansion: We will extend LAAC implementation and evaluation to additional communication domains including business proposals, technical documentation, grant applications, policy briefs, and cross-cultural communication scenarios.

Standardized Benchmarks: The broader research community would benefit from standardized benchmarks for evaluating communication intermediary systems, enabling rigorous comparison across approaches and tracking progress over time.

Regulatory and Ethical Considerations: As LAAC-style systems become more prevalent, important questions arise about liability when AI intermediaries misrepresent information, disclosure requirements for AI-mediated communication, and standards for acceptable trustworthiness levels in different domains.

8 Conclusion

The proliferation of LLMs has created a paradoxical situation where these powerful tools are being used to inflate and then compress information, wasting computational resources while eliminating authentic human communication. The LAAC framework offers an alternative paradigm: positioning LLMs as trusted intermediaries that faithfully capture sender intent and facilitate recipient understanding through structured knowledge and interactive querying.

However, our systematic evaluation reveals that deploying LLMs in this communication intermediary role faces critical trustworthiness challenges. Current systems exhibit measurable gaps in information capture fidelity, reproducibility, and query response integrity. Interview Agents achieve variable extraction accuracy, Extraction Agents produce inconsistent knowledge structures from identical inputs, and Query Agents show concerning tendencies toward hallucination and citation fabrication.

These findings should not be interpreted as fundamental limitations of the LAAC approach, but rather as empirical validation that trustworthiness cannot be assumed—it must be systematically evaluated and engineered. The path forward requires technical innovations in fidelity verification, uncertainty quantification, and hallucination prevention, combined with appropriate human-in-the-loop validation for high-stakes communication domains.

As LLMs continue to evolve, their potential role as communication intermediaries represents a more authentic and efficient application than content generation. However, realizing this potential requires the AI research community to prioritize trustworthiness alongside capability, developing

systems that users can rely on to faithfully represent their intent. This paper provides a framework for understanding and measuring that trustworthiness, establishes preliminary baselines, and charts a path toward reliable LLM-mediated communication.

References

- [1] A. Alkaissi and S. I. McFarlane. Artificial hallucinations in chatgpt: Implications in scientific writing. *Cureus*, 15(2), 2023.
- [2] D. Amodei et al. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.
- [3] Anthropic. Claude: Constitutional ai and harmlessness. <https://www.anthropic.com>, 2023.
- [4] Y. Chen et al. Document question answering with contextualized language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5621–5634. Association for Computational Linguistics, 2022.
- [5] H. Frieder et al. Mathematical capabilities of chatgpt. *arXiv preprint arXiv:2301.13867*, 2023.
- [6] Grammarly. Ai-powered writing assistant. <https://www.grammarly.com>, 2024.
- [7] J. He et al. Quantifying uncertainty in large language models. *arXiv preprint arXiv:2302.09664*, 2023.
- [8] S. Hong et al. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*, 2023.
- [9] Z. Ji et al. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023.
- [10] M. Katz et al. Verifying neural networks: Current methods and future challenges. *Formal Methods in System Design*, 57:261–296, 2021.
- [11] S. Kumar and J. Rodriguez. Large language models for academic writing: Opportunities and challenges. *Nature Human Behaviour*, 7(8):1234–1245, 2023.
- [12] R. Lee et al. Collaborative writing with ai: Understanding the human-ai partnership. *ACM Transactions on Computer-Human Interaction*, 30(4):1–28, 2023.
- [13] P. Lewis et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 9459–9474, 2020.
- [14] S. Maynez et al. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1906–1919. Association for Computational Linguistics, 2020.
- [15] OpenAI. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.
- [16] N. Rashkin et al. Fact-checking meets fauxtography: Verifying claims about images. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2374–2390. Association for Computational Linguistics, 2021.
- [17] T. Richards. Autogpt: An autonomous gpt-4 experiment. GitHub repository, 2023.

- [18] Q. Wu et al. Autogen: Enabling next-gen llm applications via multi-agent conversation. *arXiv preprint arXiv:2308.08155*, 2023.
- [19] M. Zhang et al. Ai-assisted email composition: Usage patterns and user satisfaction. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–14. ACM, 2023.