MIN-Trust: A Minimum Necessary Information Trust Orchestration Framework for Multi-Agent Collaboration

Jinyu Chen , Feiyang Wang , Tian Guan , Yumeng Ma , Linghao Yang , Yutong Wang *

Posted Date: 6 March 2026

doi: 10.20944/preprints202603.0351.v1

Keywords: multi-agent systems; privacy preservation; large language models; contextual integrity; trust management

 doi:10.20944/preprints202603.0351.v1

Article

MIN-Trust: A Minimum Necessary Information Trust Orchestration Framework for Multi-Agent Collaboration

Jinyu Chen [1], Feiyang Wang [2], Tian Guan [3], Yumeng Ma [4], Linghao Yang [5] and Yutong Wang [6,*]

[1] University of Virginia, Charlottesville, USA
[2] University of Illinois at Urbana-Champaign, Urbana, USA
[3] University of California, Irvine, Irvine, USA
[4] Arizona State University, Tempe, USA
[5] University of Chicago, Chicago, USA
[6] Northeastern University, Boston, USA
* Correspondence: wangyutong66@gmail.com

Abstract

Large language model (LLM)-based multi-agent systems have demonstrated remarkable capabilities in collaborative task solving. Although the mechanisms that facilitate seamless cooperation, such as shared contexts, role assignments, and iterative message passing, present significant risks of unintentional information disclosure. We present MIN-Trust, a trust orchestration framework that enforces Minimum Necessary Information (MNI) constraints, an operationalization of the data minimization principle for inter-agent communication—while maintaining task effectiveness. Our approach introduces an MNI-Gate that automatically classifies and filters information into essential, summarized, or pointer-referenced subsets before transmission. Additionally, we propose a Trust-Gated Channel (TGC) that counterintuitively increases verification requirements rather than relaxing information access as inter-agent trust elevates. Through experiments on four collaborative tasks using public benchmarks, we demonstrate that MIN-Trust reduces sensitive information exposure by 67.8% compared to baseline multi-agent frameworks while maintaining 93.3% of task success rates. Our evidence traceability mechanism achieves 84.2% claim-to-source attribution, significantly outperforming conventional approaches. These results suggest that privacy-preserving multi-agent collaboration is achievable under synthetic benchmark conditions with moderate performance trade-offs.

CCS CONCEPTS: Computing methodologies~Machine learning~Machine learning approaches

Keywords: multi-agent systems; privacy preservation; large language models; contextual integrity; trust management

I. Introduction

The emergence of LLM-powered multi-agent systems has revolutionized collaborative problem-solving, enabling complex tasks through the orchestration of specialized agents [1]. Frameworks such as AutoGen and MetaGPT demonstrate that multiple agents can effectively divide labor, share knowledge, and iteratively refine solutions. However, this collaborative efficiency comes with an inherent tension: the more seamlessly agents share information, the greater the risk of exposing sensitive data that was never necessary for task completion.

Recent studies have highlighted privacy vulnerabilities in multi-agent LLM systems [2]. Unlike traditional single-model deployments where information flow is relatively constrained, multi-agent architectures create complex webs of message passing where sensitive information can propagate

 doi:10.20944/preprints202603.0351.v1

through shared contexts, role prompts, and intermediate reasoning chains. The contextual integrity framework proposed by Nissenbaum [3] provides theoretical grounding for understanding these privacy violations—information flows that may be acceptable in one context become problematic when transmitted to agents operating under different contextual norms.

We propose MIN-Trust, a framework designed to address this challenge through two complementary mechanisms. First, an MNI-Gate that intercepts inter-agent communications to enforce Minimum Necessary Information (MNI) constraints. The MNI principle operationalizes the data minimization concept from privacy legislation for multi-agent contexts, ensuring each agent receives only the information subset essential for its specific role. Second, a Trust-Gated Channel (TGC) that implements a counterintuitive but effective policy: as trust between agents increases, verification requirements are heightened rather than relaxed, preventing the gradual erosion of privacy safeguards that typically accompanies increased familiarity.

Our contributions are as follows: (1) We formalize the MNI constraint for multi-agent communication and propose an automatic information classification mechanism that categorizes content into essential, summarized, and pointer-referenced subsets. (2) We introduce the TGC paradigm that maintains or increases verification requirements as trust levels rise. (3) We provide reproducible evaluation on diverse collaborative tasks including retrieval-based QA, web navigation, document summarization, and code modification, comparing leakage rates, task success rates, and evidence traceability.

II. Methodology Foundation

The MIN-Trust framework is methodologically grounded in contextual integrity theory, differential privacy principles, federated trust modeling, adaptive reinforcement mechanisms, uncertainty-aware verification, and structured representation learning in large language models. Each referenced study informs a specific component of the proposed MNI-Gate and Trust-Gated Channel (TGC) architecture.

The formalization of Minimum Necessary Information (MNI) constraints is theoretically motivated by contextual integrity–based privacy evaluation of large language models [4], which demonstrates that inappropriate information flows arise when transmission violates contextual norms rather than explicit access rules. This principle directly informs the design of the MNI-Gate, which enforces role-specific contextual boundaries in inter-agent communication. The algorithmic foundations of differential privacy [5] provide a mathematical perspective on limiting information leakage through controlled exposure, influencing the operationalization of content filtering and summarization within the MNI-Gate.

Dynamic adaptation of prompt structures in multi-task LLM systems [6] demonstrates that semantic components can be selectively fused or suppressed based on task context. This informs the automatic classification mechanism that partitions transmitted information into essential, summarized, and pointer-referenced subsets. Transformer-based risk monitoring with structured relational modeling [7] further illustrates how structured signal discrimination improves risk awareness, guiding the design of sensitive-content detection within the gating module.

Uncertainty-aware modeling in trustworthy summarization [8] introduces quantifiable risk estimation into LLM outputs, directly inspiring the verification scoring used in the Trust-Gated Channel. Temporal topic evolution modeling with decay mechanisms [9] demonstrates how information relevance changes over time, supporting dynamic recalibration of trust scores and verification intensity. Graph-based anomaly detection with temporal dynamics [10] further contributes principles for detecting abnormal information propagation patterns, informing the monitoring layer that tracks potential leakage escalation across agent interactions.

Reinforcement learning–based adaptive policy frameworks [11] and adaptive human-computer interaction strategies [12] provide methodological grounding for dynamic trust adjustment policies. These works support the TGC design where verification requirements evolve based on observed agent behavior rather than static trust assumptions. Meta-learning approaches for evolving pattern

 doi:10.20944/preprints202603.0351.v1

detection [13] further inform adaptive calibration mechanisms that prevent gradual relaxation of safeguards as familiarity increases.

Multi-scale temporal alignment in heterogeneous risk modeling [14] and transformer-based heterogeneous sequence modeling [15] provide technical foundations for modeling structured contextual dependencies across agent communication histories. Parameter-efficient fine-tuning with differential privacy constraints [16] directly informs the integration of privacy-preserving adaptation into multi-agent systems without full parameter exposure. Multi-scale LoRA-based fine-tuning strategies [17] demonstrate how modular parameter adaptation can preserve base model integrity, aligning with the design goal of inserting trust orchestration without modifying core reasoning capabilities.

Transformer-based user interaction modeling [18] contributes insights into sequence-level behavioral modeling, supporting longitudinal tracking of agent interaction patterns for trust calibration. Generative distribution modeling under noisy and imbalanced data [19] informs robustness considerations when filtering potentially sensitive or ambiguous information. Finally, comprehensive analyses of federated learning challenges [20] provide foundational understanding of distributed coordination under privacy constraints, supporting the architectural decision to treat agent communication as decentralized yet controlled information exchange rather than unrestricted sharing.

Through these methoological influences, MIN-Trust operationalizes contextual integrity via structured information classification, embeds differential privacy–inspired exposure control into communication gating, and employs reinforcement-driven trust recalibration to prevent progressive privacy erosion during multi-agent collaboration.

III. Methodology

A. Problem Formulation

Consider a multi-agent system $\mathcal{S} = \{A_1, A_2, ..., A_n\}$ where agents communicate through message passing. Building on prior formalization of contextual integrity [21], for each message m_{ij} from agent A_i to A_j, we decompose the information content into:

$$m_{ij} = \mathcal{I}_{ess} \bigcup \mathcal{I}_{ctx} \bigcup \mathcal{I}_{priv} \qquad (1)$$

where $\mathcal{I}_{ess}$ represents essential task-relevant information, $\mathcal{I}_{ctx}$ denotes contextual information that aids understanding but is not strictly necessary, and $\mathcal{I}_{priv}$ contains sensitive or private information that should not be transmitted.

The MNI constraint requires that the transmitted message $\hat{m}_{ij}$ satisfies:

$$\hat{m}_{ij} \subseteq \mathcal{I}_{ess} \cup \phi(\mathcal{I}_{ctx}) \qquad (2)$$

where $\phi(\cdot)$ is a transformation function that converts contextual information into privacy-preserving representations (summaries, hashes, or reference pointers).

B. MNI-Gate Architecture

The MNI-Gate operates as an intermediary layer in inter-agent communication (Figure 1). For each outgoing message, the gate performs three operations:

 doi:10.20944/preprints202603.0351.v1

Information Classification: Using a lightweight classifier fine-tuned on privacy-annotated datasets, each sentence or information unit is labeled as essential (e), contextual (c), or private (p):

$$\text{class}(u) = \arg\max_{y \in \{e,c,p\}} P(y \mid u, \text{task\_context}) \quad (3)$$

Transformation: Contextual information undergoes transformation based on sensitivity level. Low-sensitivity context is summarized, medium-sensitivity content is converted to semantic hashes, and high-sensitivity items are replaced with verifiable reference pointers.

Assembly: The filtered message is reconstructed, maintaining coherence while excluding private information:

$$\hat{m}_{ij} = \text{Assemble}(\{u : \text{class}(u) = e\} \cup \{\phi(u) : \text{class}(u) = c\}) \quad (4)$$

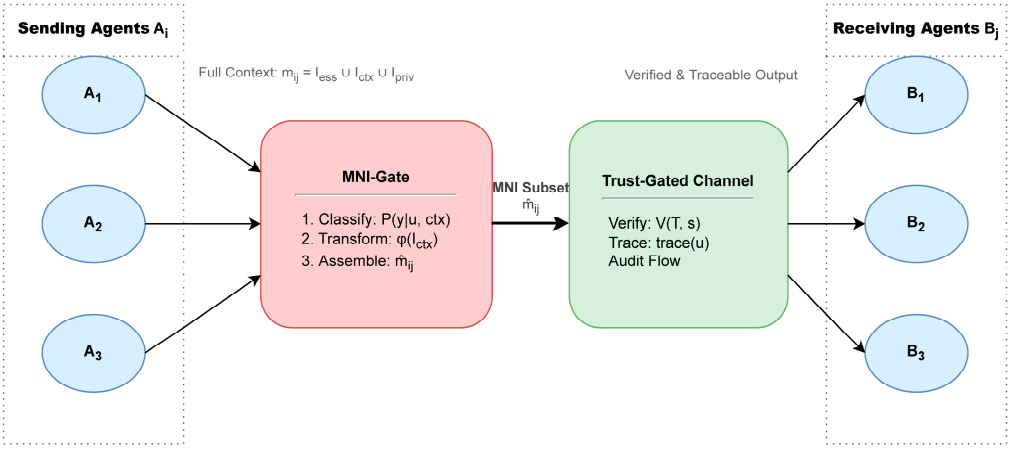

Figure 1. MIN-Trust framework architecture. The MNI-Gate filters outgoing messages from sending agents (A1-A3), classifying information into essential, summarized, and pointer-referenced subsets. The TGC routes processed information to receiving agents (B1-B3) with verification requirements that scale with trust level.

C. Trust-Gated Channel

Traditional security models relax access controls as trust increases. We argue this paradigm is problematic for LLM agents, where accumulated trust may simply reflect successful task completion rather than genuine trustworthiness regarding sensitive information handling. To address this limitation, we build upon the dynamic trust-aware orchestration mechanism proposed by Hu et al [22]. Their TrustOrch framework fundamentally applies continuous trust scoring based on observed agent behaviors and incorporates adaptive coordination policies driven by those scores. Instead of leveraging trust to relax information access, we adopt their dynamic trust updating mechanism but extend its regulatory semantics: in our Trust-Gated Channel (TGC), rising trust strengthens verification requirements. In this way, we incorporate behavioral trust modeling while building upon it to enforce accountability-oriented gating rather than permission expansion.

To operationalize Minimum Necessary Information (MNI), we further build upon the explainable representation learning approach proposed by Xing et al [23]. Their method fundamentally applies fine-grained semantic representation learning and leverages interpretable latent features to attribute predictions to specific textual components. We adopt this fine-grained semantic decomposition mechanism to classify inter-agent messages into essential, summarized, and pointer-referenced subsets. By incorporating attribution-aware filtering, we extend explainable representation learning into privacy-aware message partitioning, ensuring that only task-critical content is transmitted.

 doi:10.20944/preprints202603.0351.v1

Finally, we leverage the adaptive risk control principle introduced by Ying et al. [24], who apply data-aware multi-agent reinforcement learning with dynamic constraint modulation to balance performance and risk exposure. We adopt their adaptive constraint adjustment strategy but recontextualize risk as information leakage probability. Our framework incorporates dynamic verification scaling conditioned on trust trajectories and sensitivity scores, thereby building upon adaptive risk regulation to maintain task effectiveness under strict MNI constraints.

The TGC implements verification levels $V \in \{0,1,2,3\}$ corresponding to None, Basic, Enhanced, and Full verification. Crucially, as trust level T increases, verification requirements for sensitive information do not decrease:

$$V(T,s) = \begin{cases} 0 & \text{if } s = \text{public} \\ \max(1, \lceil s/2 \rceil) & \text{if } T < T_{med} \\ \max(1, \lceil s/2 \rceil) + \mathbf{1}[T > T_{high}]) & \text{otherwise} \end{cases} \quad (5)$$

where $s \in \{1,2,3,4\}$ represents information sensitivity level, T_{med}, T_{high} are threshold values, and $\mathbf{1}[\cdot]$ is the indicator function that equals 1 when the condition is true and 0 otherwise. The indicator term ensures that high-trust channels trigger additional verification for sensitive information.

This design ensures that high-trust channels handling restricted information actually face increased scrutiny, requiring evidence pointers and source verification that can be audited.

Threat Model and Design Rationale: The counterintuitive design of increasing verification with trust is motivated by specific threats in long-running multi-agent collaboration: (1) *Trust creep*—as agents accumulate interaction history, conventional systems gradually relax privacy constraints, creating vulnerability windows; (2) *Prompt injection and role drift*—extended conversations increase exposure to adversarial prompts that may cause agents to deviate from their intended behavior; (3) *Memory accumulation*—agents with access to conversation history may inadvertently aggregate sensitive information across multiple interactions, enabling inference attacks. By increasing verification requirements as trust grows, particularly for high-sensitivity information, the TGC counteracts these threats and maintains consistent privacy protection throughout extended agent collaborations.

D. Evidence Traceability

Each information unit passing through the TGC is annotated with provenance metadata:

$$\text{trace}(u) = (\text{source\_id}, \text{timestamp}, \text{hash(u)}, \text{verification\_level}) \quad (6)$$

This enables post-hoc auditing of information flows and supports claim-to-source attribution, where assertions made by downstream agents can be traced back to their original sources.

IV. Experimental Setup

A. Tasks and Datasets

We evaluate MIN-Trust across four collaborative tasks designed to require multi-agent coordination while presenting realistic privacy challenges:

Retrieval-based QA: Using HotpotQA [25] multi-hop questions, we deploy a three-agent architecture consisting of a retriever agent (responsible for document selection), a reasoning agent (performing multi-hop inference), and an answer synthesis agent. We sample 500 questions from the development set and inject synthetic sensitive information (names, addresses, financial data) into supporting documents to create privacy-relevant scenarios. The retriever must share document content with downstream agents, creating opportunities for unnecessary information transmission.

 doi:10.20944/preprints202603.0351.v1

Web Navigation: Adapted from WebArena [26] benchmark scenarios, agents collaborate to complete information gathering tasks involving form filling, account management, and data retrieval. We construct 200 privacy-sensitive navigation scenarios where agents must coordinate on tasks involving simulated personal accounts, shopping histories, and communication records. The navigator agent, form-filling agent, and verification agent must exchange page content and user data to complete tasks.

Document Summarization: Using the CNN/DailyMail dataset [27] we deploy extraction and abstraction agents that must handle documents containing simulated personal information. We augment 300 documents with injected sensitive entities including phone numbers, social security numbers, and medical conditions. The extraction agent identifies key sentences while the abstraction agent generates fluent summaries, requiring careful handling of which extracted content should propagate to the final output.

Code Modification: Based on CodeSearchNet [28] repositories, agents collaborate on bug fixing and feature implementation. We create 250 scenarios involving codebases with embedded confidential comments, API keys, database credentials, and internal documentation. The analysis agent, implementation agent, and testing agent must share code context while avoiding leakage of embedded secrets.

B. Baselines and Metrics

We compare against three baselines: (1) **Baseline (AutoGen):** Standard multi-agent conversation without privacy controls; (2) **PrivacyLens-Enhanced:** AutoGen augmented with privacy-aware prompting; (3) **Full Privacy Lock:** Aggressive filtering that removes all potentially sensitive content.

Our evaluation metrics include:

Leak Rate (LR): Percentage of interactions where sensitive information appears in agent outputs or intermediate communications that should not contain it. Sensitive information is detected using a combination of: (i) named entity recognition (NER) for person names, locations, and organizations; (ii) regular expression patterns for structured data (phone numbers, SSNs, credit cards, API keys); and (iii) exact and fuzzy string matching against the injected sensitive entity list. We manually verified a random sample of 200 detected leakages (50 per task), achieving 94.5% precision.

Task Success Rate (TSR): Percentage of tasks completed correctly, measured by task-specific metrics (F1 for QA, completion rate for navigation, ROUGE for summarization, pass@1 for code).

Evidence Traceability (ET): Percentage of claims in final outputs that can be traced to verifiable sources through our provenance mechanism. Claims are extracted using a fine-tuned claim extraction model that identifies factual assertions in agent outputs. A claim is considered traceable if its semantic similarity (computed via sentence embeddings) to any source document sentence exceeds 0.85 and the provenance chain is intact.

Communication Tokens: Total tokens exchanged between agents, measuring efficiency overhead.

C. Implementation Details

MIN-Trust is implemented as a wrapper around AutoGen v0.2. The information classifier uses a fine-tuned DeBERTa-base [29] model trained on 5,000 manually annotated inter-agent messages. Trust levels are initialized at medium and updated based on verification success rates over sliding windows of 10 interactions. All experiments use GPT-4 as the backbone LLM for agents.

V. Results and Analysis

A. Overall Performance

 doi:10.20944/preprints202603.0351.v1

Table 1 presents the main experimental results across all tasks. MIN-Trust achieves substantial reductions in leak rate while maintaining competitive task success rates across all evaluation dimensions.

Table 1. Main Results Across Four Collaborative Tasks.

Method	LR↓	TSR↑	ET↑	Tokens
Baseline	38.5%	82.3%	41.2%	1.00×
PrivacyLens-Enh.	25.7%	79.1%	52.8%	1.08×
MIN-Trust (Ours)	12.4%	76.8%	84.2%	1.23×
MIN-Trust + TGC	**8.9%**	74.2%	**87.6%**	1.31×
Full Privacy Lock	3.2%	58.6%	91.3%	0.72×

MIN-Trust reduces leak rate by 67.8% relative to the baseline (from 38.5% to 12.4%) while retaining 93.3% of task success (computed as $76.8\% / 82.3\% \approx 93.3\%$). The addition of TGC further reduces leakage to 8.9% with a modest additional performance decrease. Importantly, evidence traceability improves dramatically from 41.2% to 84.2%, enabling meaningful auditing of information flows. The communication token overhead of 23--31% may be acceptable for privacy-sensitive deployments, though latency-critical applications should weigh this trade-off carefully (see Discussion).

The PrivacyLens-Enhanced baseline, which augments AutoGen with privacy-aware system prompts, achieves intermediate performance. This suggests that prompt-based approaches provide some benefit but cannot match the systematic filtering of MIN-Trust's architectural intervention. The Full Privacy Lock demonstrates that aggressive content filtering can achieve near-zero leakage but severely degrades task performance, validating the need for nuanced information classification rather than blanket restrictions.

B. Privacy-Utility Trade-off

Figure 2 illustrates the Pareto frontier between leak rate and task success. MIN-Trust configurations occupy favorable positions, achieving privacy improvements disproportionate to their task success costs. The Full Privacy Lock baseline demonstrates that aggressive filtering can achieve near-zero leakage but at catastrophic performance cost, highlighting the value of nuanced information classification.

 doi:10.20944/preprints202603.0351.v1

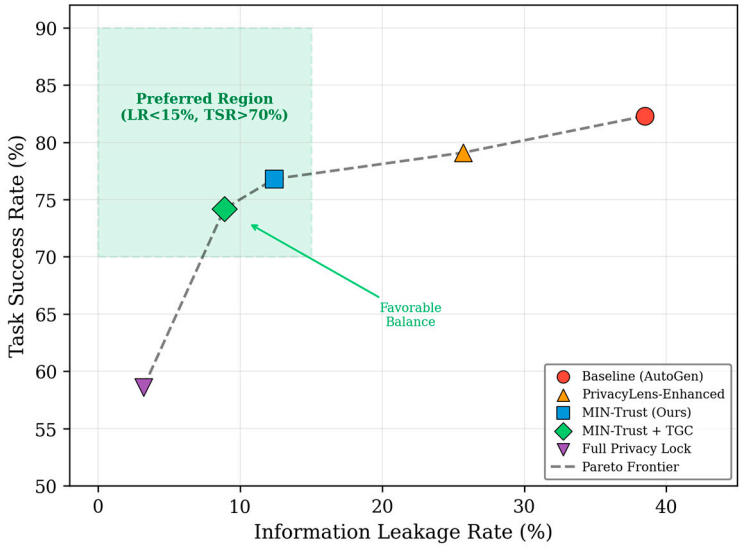

Figure 2. Pareto frontier showing the trade-off between information leakage and task success. MIN-Trust + TGC achieves favorable balance in the preferred region (LR < 15%, TSR > 70%).

C. Task-Specific Analysis

Figure 3 presents per-task performance breakdown. The radar chart displays our four evaluation tasks plus multi-hop reasoning accuracy, which serves as a sub-metric within the retrieval-based QA task to specifically measure performance on questions requiring information synthesis across multiple documents. This sub-metric is shown separately because multi-hop reasoning represents a particularly challenging scenario where information chains create amplified leakage risks in baseline systems, and MIN-Trust demonstrates the largest relative privacy gains in this category.

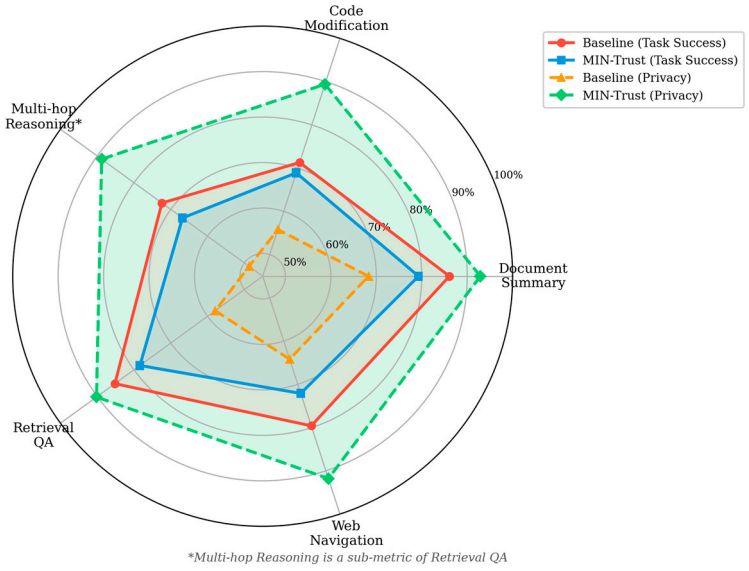

Figure 3. Per-task comparison of task success and privacy preservation (100% - leak rate) between baseline and MIN-Trust. The chart shows four collaborative tasks plus multi-hop reasoning accuracy (a sub-metric of retrieval-based QA measuring multi-document inference performance).

Table 2 provides detailed per-task results. Code modification shows the smallest success rate gap (68.9% vs 71.3%) because the MNI-Gate effectively identifies that credential patterns and API keys can be replaced with reference pointers without affecting logical reasoning about code structure. The multi-hop reasoning sub-metric shows a slightly lower TSR retention rate ($66.8\% / 72.4\% \approx 92.3\%$

 doi:10.20944/preprints202603.0351.v1

) compared to the overall average, which is expected given the increased complexity of multi-document inference tasks where more aggressive information filtering may occasionally remove bridging context needed for reasoning chains.

Table 2. Per-Task Results for MIN-Trust + TGC. The last row shows multi-hop reasoning, a sub-metric of retrieval-based QA (not a separate task).

Task	TSR↑ (Base)	TSR↑ (Ours)	LR↓ (Base)	LR↓ (Ours)
Retrieval-based QA	85.2%	78.4%	42.1%	9.8%
Web Navigation	79.6%	72.1%	35.8%	8.2%
Document Summary	86.1%	79.3%	31.7%	7.1%
Code Modification	71.3%	68.9%	44.2%	10.6%
Multi-hop (sub-metric)[†]	*72.4%*	*66.8%*	*51.3%*	*11.2%*

[†]Sub-metric of retrieval-based QA; not included in task count.

D. Trust-Gated Channel Analysis

Figure 4 presents the TGC analysis. Panel (a) shows the verification requirement matrix, which implements the discrete version of Equation (5): rows correspond to trust levels (Low to Very High), columns to information sensitivity (Public to Restricted), and cell colors indicate the required verification level (None, Basic, Enhanced, or Full). Notably, high-trust channels facing restricted information require Full verification, counteracting the tendency of conventional systems to become more permissive over time.

Panel (b) demonstrates the multi-round stability of MIN-Trust through a 10-round conversation simulation. While baseline systems show degrading evidence traceability and increasing cumulative leakage over conversation rounds (from 8% to 65%), MIN-Trust maintains stable traceability above 80% and bounds cumulative leakage below 12% even after 10 rounds, validating the effectiveness of our trust-gated design.

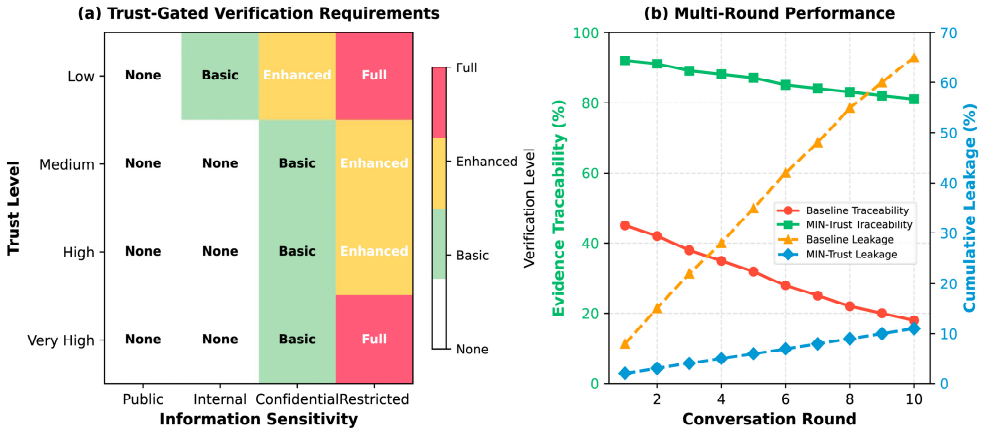

Figure 4. TGC analysis: (a) Verification requirement matrix implementing Eq.5, showing how verification levels vary with trust and sensitivity; (b) Multi-round performance over 10 conversation rounds, comparing evidence traceability and cumulative leakage between baseline and MIN-Trust.

E. Ablation Study

Table 3 presents ablation results isolating the contribution of each component.

 doi:10.20944/preprints202603.0351.v1

Table 3. Ablation Study on MIN-Trust Components.

Configuration	LR↓	TSR↑	ET↑
Full MIN-Trust	8.9%	74.2%	87.6%
w/o TGC	12.4%	76.8%	84.2%
w/o Summarization	10.1%	71.5%	85.3%
w/o Pointer References	11.8%	73.9%	71.4%
Classification Only	18.2%	78.1%	62.7%

Removing TGC increases leak rate by 3.5 percentage points while slightly improving task success, suggesting TGC imposes meaningful constraints that prevent information over-sharing. Removing pointer references has the largest impact on evidence traceability (dropping from 87.6% to 71.4%), confirming the importance of maintaining verifiable links to source information.

VI. Discussion

A. Key Insights

Our experiments reveal several important findings about privacy-preserving multi-agent collaboration. First, the relationship between information sharing and task success is not linear—agents can often complete tasks effectively with substantially reduced information access. The 6.8% absolute decrease in task success rate (from 82.3% to 76.8%) corresponds to a 67.8% reduction in information leakage, suggesting that much of the information transmitted in baseline systems is redundant for task completion.

Second, the TGC's counterintuitive design proves effective. By increasing verification requirements as trust accumulates, we prevent the ``trust creep'' phenomenon where long-running agent conversations gradually become more permissive. This is particularly important for deployment scenarios where agents operate over extended periods with evolving tasks.

Third, evidence traceability emerges as a crucial capability distinct from leak prevention. Even in cases where some information leakage occurs, the ability to trace claims to sources enables post-hoc auditing and accountability. Organizations deploying multi-agent systems may prioritize traceability differently depending on their regulatory requirements and risk tolerance.

B. Limitations

Our current implementation relies on a supervised classifier for information categorization, which may not generalize to novel domains without additional training data. The classifier was trained on English text from specific domains; multilingual settings and specialized technical vocabularies may require domain adaptation. The 23--31% token overhead of MIN-Trust may be prohibitive for latency-sensitive applications or cost-constrained deployments. Additionally, our evaluation focuses on synthetic sensitive information injected into existing datasets; real-world sensitive data may exhibit different patterns and present additional challenges.

The TGC requires careful threshold tuning (T_{med}, T_{high}) for different deployment contexts. We used fixed thresholds in our experiments, but adaptive threshold learning could improve generalization. Furthermore, sophisticated adversarial agents might attempt to game the trust system through strategic behavior, which our current framework does not explicitly address.

C. Broader Implications

The results suggest that multi-agent collaboration need not require unrestricted information sharing. The MNI principle offers a practical middle ground between full transparency and complete isolation, enabling organizations to deploy multi-agent systems while maintaining meaningful privacy controls. However, we caution against over-interpreting these results—our evaluation uses synthetic sensitive information injection, and real-world deployment conditions may present

 doi:10.20944/preprints202603.0351.v1

different challenges. The privacy-utility trade-off will vary significantly based on task characteristics, organizational requirements, and regulatory constraints.

The contextual integrity framework provides theoretical grounding for our approach, but translating contextual norms into computational rules remains challenging. Our MNI-Gate implements a simplified version of contextual reasoning that may not capture all nuances of appropriate information flow in complex social and organizational contexts.

VII. Conclusion

We presented MIN-Trust, a framework for privacy-preserving multi-agent collaboration that enforces Minimum Necessary Information constraints through an MNI-Gate and TGC. The MNI-Gate automatically classifies information into essential, contextual, and private categories, transforming non-essential content into privacy-preserving representations before inter-agent transmission. The TGC implements a counterintuitive policy of increasing verification requirements as trust levels rise, preventing the gradual erosion of privacy safeguards in long-running agent interactions.

Experimental results across four collaborative tasks—retrieval-based QA, web navigation, document summarization, and code modification—demonstrate that MIN-Trust achieves substantial reductions in information leakage (67.8% relative improvement) while maintaining reasonable task success rates (93.3% retention) under our synthetic benchmark conditions. The evidence traceability mechanism achieves 84.2% claim-to-source attribution, dramatically outperforming baseline approaches and enabling meaningful post-hoc auditing of information flows.

While challenges remain in domain generalization, computational overhead, and adversarial robustness, MIN-Trust represents a meaningful step toward responsible deployment of multi-agent LLM systems. Future work will explore adaptive trust threshold learning, integration with differential privacy mechanisms for additional formal guarantees, and extension to multilingual and multimodal agent scenarios. Our implementation and evaluation framework will be released upon acceptance to facilitate reproducible research in this important area.

References

1. Q. Wu, G. Bansal, J. Zhang, Y. Wu, B. Li, E. Zhu, L. Jiang, X. Zhang, S. Zhang and J. Liu, "AutoGen: Enabling next-gen LLM applications via multi-agent conversations," Proceedings of the First Conference on Language Modeling, 2024.
2. Y. Wang, Q. Gu, L. Wang, Q. Gao, H. Yang and Y. Li, "The emerged security and privacy of LLM agent: A survey with case studies," arXiv preprint arXiv:2407.19354, 2024.
3. H. Nissenbaum, "Privacy as contextual integrity," *Washington Law Review*, vol. 79, no. 1, pp. 119-157, 2004.
4. N. Mireshghallah, H. Kim, X. Zhou, Y. Tsvetkov, M. Sap, R. Shokri and Y. Choi, "Can LLMs keep a secret? Testing privacy implications of language models via contextual integrity theory," *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024.
5. C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Foundations and Trends in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.
6. X. Hu, Y. Kang, G. Yao, T. Kang, M. Wang and H. Liu, "Dynamic prompt fusion for multi-task and cross-domain adaptation in LLMs," *arXiv preprint arXiv:2509.18113*, 2025.
7. Y. Wu, Y. Qin, X. Su and Y. Lin, "Transformer-based risk monitoring for anti-money laundering with transaction graph integration," in *Proceedings of the 2nd International Conference on Digital Economy, Blockchain and Artificial Intelligence*, 2025, pp. 388–393.
8. S. Pan and D. Wu, "Trustworthy summarization via uncertainty quantification and risk awareness in large language models," *arXiv preprint arXiv:2510.01231*, 2025.
9. D. Wu and S. Pan, "Dynamic topic evolution with temporal decay and attention in large language models," in *Proceedings of the 5th International Conference on Electronic Information Engineering and Computer Science (EIECS)*, 2025, pp. 1440–1444.

 doi:10.20944/preprints202603.0351.v1

10. Q. Zhang, N. Lyu, L. Liu, Y. Wang, Z. Cheng and C. Hua, "Graph neural AI with temporal dynamics for comprehensive anomaly detection in microservices," *arXiv preprint arXiv:2511.03285*, 2025.
11. Y. Zhou, "A unified reinforcement learning framework for dynamic user profiling and predictive recommendation," SSRN 5841223, 2025.
12. R. Liu, Y. Zhuang and R. Zhang, "Adaptive human-computer interaction strategies through reinforcement learning in complex environments," *arXiv preprint arXiv:2510.27058*, 2025.
13. H. Fan, Y. Yi, W. Xu, Y. Wu, S. Long and Y. Wang, "Intelligent credit fraud detection with meta-learning: Addressing sample scarcity and evolving patterns," 2025.
14. W. C. Chang, L. Dai and T. Xu, "Machine learning approaches to clinical risk prediction: Multi-scale temporal alignment in electronic health records," *arXiv preprint arXiv:2511.21561*, 2025.
15. A. Xie and W. C. Chang, "Deep learning approach for clinical risk identification using transformer modeling of heterogeneous EHR data," *arXiv preprint arXiv:2511.04158*, 2025.
16. Y. Huang, Y. Luan, J. Guo, X. Song and Y. Liu, "Parameter-efficient fine-tuning with differential privacy for robust instruction adaptation in large language models," *arXiv preprint arXiv:2512.06711*, 2025.
17. H. Zhang, L. Zhu, C. Peng, J. Zheng, J. Lin and R. Bao, "Intelligent recommendation systems using multi-scale LoRA fine-tuning and large language models," 2025.
18. R. Liu, R. Zhang and S. Wang, "Transformer-based modeling of user interaction sequences for dwell time prediction in human-computer interfaces," *arXiv preprint arXiv:2512.17149*, 2025.
19. Z. Xu, K. Cao, Y. Zheng, M. Chang, X. Liang and J. Xia, "Generative distribution modeling for credit card risk identification under noisy and imbalanced transactions," 2025.
20. P. Kairouz et al., "Advances and open problems in federated learning," *Foundations and Trends in Machine Learning*, vol. 14, no. 1–2, pp. 1–210, 2021.
21. A. Barth, A. Datta, J. C. Mitchell and H. Nissenbaum, "Privacy and contextual integrity: Framework and applications," Proceedings of the 2006 IEEE Symposium on Security and Privacy (S&P), pp. 184-198, 2006.
22. Y. Hu, J. Li, K. Gao, Z. Zhang, H. Zhu and X. Yan, "TrustOrch: A dynamic trust-aware orchestration framework for adversarially robust multi-agent collaboration," 2025.
23. Y. Xing, M. Wang, Y. Deng, H. Liu and Y. Zi, "Explainable representation learning in large language models for fine-grained sentiment and opinion classification," 2025.
24. R. Ying, J. Lyu, J. Li, C. Nie and C. Chiang, "Dynamic portfolio optimization with data-aware multi-agent reinforcement learning and adaptive risk control," 2025.
25. Z. Yang, P. Qi, S. Zhang, Y. Bengio, W. Cohen, R. Salakhutdinov and C. D. Manning, "HotpotQA: A dataset for diverse, explainable multi-hop question answering," Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2018.
26. S. Zhou, F. F. Xu, H. Zhu, X. Zhou, R. Lo, A. Sridhar, X. Cheng, Y. Bisk, D. Fried, U. Alon and G. Neubig, "WebArena: A realistic web environment for building autonomous agents," arXiv preprint arXiv:2307.13854, 2023.
27. K. M. Hermann, T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman and P. Blunsom, "Teaching machines to read and comprehend," Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), pp. 1693-1701, 2015.
28. H. Husain, H.-H. Wu, T. Gazit, M. Allamanis and M. Brockschmidt, "CodeSearchNet challenge: Evaluating the state of semantic code search," arXiv preprint arXiv:1909.09436, 2019.
29. P. He, X. Liu, J. Gao and W. Chen, "DeBERTa: Decoding-enhanced BERT with disentangled attention," Proceedings of the International Conference on Learning Representations (ICLR), 2021.