ESP IJAST
ESP International Journal of Advancements in Science & Technology
ISSN: 2583-9233 / Volume 4 Issue 1 March 2026 / Page No: 46–52
Paper Id: IJAST-V4I1P106 / Doi: 10.56472/25839233/IJAST-V4I1P106

Original Article

TPS-Eval: Coupled Trust, Privacy, and Security Evaluation of Agentic Clinical AI Pipelines

Saritha Kondapally
Healthcare Technology, Sr. Solution Architecture, Atlanta, GA, USA.

Received Date: 18 February 2026 Revised Date: 26 February 2026 Accepted Date: 02 March 2026

Abstract: *Agentic AI systems that autonomously retrieve patient data, invoke external tools, and maintain cross-session memory introduce safety challenges that extend beyond traditional model-level evaluation. Existing benchmarks assess trust, privacy, and security in isolation, overlooking critical interactions: retrieval strategies influence both answer accuracy and what Protected Health Information (PHI) enters the model context, while memory configurations affect longitudinal reasoning as well as adversarial exposure. We introduce TPS-Eval, a framework that formally defines and jointly evaluates Trust, Privacy, and Security as coupled properties of complete agentic pipelines. We compare six retrieval strategies, from keyword baselines to graph-structured approaches, across three language model backends (GPT-4o, GPT-4o-mini, and Llama-3-8B) using noise-augmented, FHIR-compliant synthetic clinical records. We extend the threat model with two agent-specific attack categories: logic poisoning of knowledge bases and toolchain feedback exploitation. Across ten independent evaluation seeds, graph-based retrieval consistently achieves the highest integrated TPS scores, with rankings robust across model scales. We derive five actionable design principles for safer deployment of agentic AI in clinical decision support systems.*

Keywords*: Agentic AI, Clinical NLP, Graph-RAG, Healthcare AI Safety, PHI Sanitization, Adversarial Robustness, Retrieval-Augmented Generation.*

I. INTRODUCTION

Healthcare represents a high-risk domain for artificial intelligence, where data heterogeneity, longitudinal complexity, and direct clinical impact demand rigorous safety evaluation. Agentic AI systems further increase this complexity: beyond answering isolated queries, they autonomously retrieve patient records, invoke external clinical tools, maintain cross-session memory, and generate outputs that may influence medical decisions, with hallucination posing a recognized risk in clinical contexts [17].

Current evaluation approaches remain fragmented. Accuracy is assessed on question-answer benchmarks, PHI leakage is audited separately, and adversarial robustness is tested in isolation. In operational agentic pipelines, however, these properties are tightly coupled. Retrieval strategies influence both answer correctness and PHI exposure. Memory architectures determine both longitudinal reasoning capability and cross-session attack surface. An architectural choice that improves accuracy may simultaneously increase privacy risk.

To address this gap, we introduce TPS-Eval, a framework that formalizes Trust, Privacy, and Security as coupled system-level properties of agentic pipelines. Our contributions include: (1) composite TPS scoring with formal definitions and calibrated refusal metrics; (2) an extended threat model incorporating logic poisoning and toolchain feedback exploitation specific to agentic workflows; (3) systematic multi-model evaluation across GPT-4o, GPT-4o-mini, and Llama-3-8B with ten-seed statistical analysis; and (4) empirically derived architectural design principles for safe clinical deployment.

While individual components such as Graph-RAG and PHI sanitization are established, this work integrates them into a unified, reproducible system-level evaluation framework tailored to healthcare agentic AI.

II. RELATED WORK

A. Agentic AI and Clinical RAG

Wang et al. [1] survey large-model-based agents, including cooperation paradigms and emerging security concerns. The ReAct framework [2] demonstrated reasoning–action interleaving for tool-augmented language models. In clinical NLP, retrieval-augmented generation (RAG) has become the dominant approach for grounding model outputs in external evidence [3]. Retrieval methods range from keyword-based BM25 [4] to semantic retrieval using BioClinicalBERT [5] and PubMedBERT [6], and more recently to graph-structured retrieval that encodes entity relationships.

However, prior work evaluates retrieval strategies primarily in terms of accuracy and relevance. The implications of retrieval choice for privacy exposure, calibrated refusal behavior, and adversarial robustness in agentic clinical workflows remain largely unexamined.

B. Trustworthy AI and Safety Frameworks

Regulatory and governance frameworks such as the EU AI Act [7] and the NIST AI RMF [8] establish high-level requirements for safety, accountability, and risk management. Model-centric safety research, including factuality and harmlessness evaluation [9], focuses on single-model behavior. Clinical explainability work [10], [11] emphasizes transparency in diagnostic systems.

These frameworks, however, assume static inference settings. They do not model multi-step agent pipelines involving retrieval, tool invocation, and persistent memory. As a result, they do not capture safety interactions that emerge from cross-component coupling in agentic systems.

C. Privacy and Adversarial Threats

Yan et al. [12] and Yao et al. [13] survey privacy risks in large language models, including training data extraction and prompt leakage. Ferrag et al. [14] provide a taxonomy of threats in LLM-agent ecosystems. Perez and Ribas [15] demonstrated large-scale prompt injection vulnerabilities.

Existing work focuses primarily on per-query leakage or prompt-level attacks. Cross-session memory reconstruction risks and knowledge-based logic poisoning in healthcare agent pipelines remain underexplored.

Table 1 positions TPS-Eval relative to these research streams. The distinguishing feature of TPS-Eval is the joint evaluation of coupled Trust–Privacy–Security properties at the full pipeline level, including agent-specific threat modeling across multiple model backends.

Table 1 : Comparison with Related Approaches

Capability	RAG	Safety Frameworks	Privacy Research	Adversarial LLM Work	TPS-Eval
Formal TPS Metrics	✗	Part.	✗	✗	✓
Coupled System Eval	✗	✗	✗	✗	✓
Multi-Model Test	Rare	✗	✗	✗	✓
Agent-Specific Threats	✗	✗	Part.	✓	✓
Memory-Aware Privacy	✗	✗	✓	✗	✓
Noise Augmentation	✗	✗	✗	✗	✓

III. FRAMEWORK DEFINITION

A. Trust: Accuracy and Calibrated Refusal

Let $Q=\{q_1,\ldots,q_n\}$ denote the evaluation query set. For each query, the agent either produces an answer (set A) or issues a refusal (set $R=Q\setminus A$). For each answered query $q_i \in A$, accuracy is defined as:

$$\mathrm{acc}(q_i) = \alpha \cdot \mathrm{EM}(q_i) + (1-\alpha) \cdot \mathrm{SemSim}(q_i)$$

where:

$\mathrm{EM}(q_i)\in\{0,1\}$ is a binary exact-match indicator,

$\mathrm{SemSim}(q_i)$ is the cosine similarity between BioClinicalBERT embeddings of the generated answer and the ground-truth response.

We set $\alpha=0.6$, prioritizing exact clinical correctness while crediting semantically equivalent responses. Sensitivity analysis across $\alpha\in[0.5,0.7]$ produced stable retrieval rankings.

Mean accuracy across answered queries is:

$$MeanAccuracy = \frac{1}{|A|}\sum_{q_i \in A} acc(q_i)$$

Calibrated Refusal Rate (CRR)

In clinical decision support, abstaining when evidence is insufficient is as important as answering correctly.

We define the Calibrated Refusal Rate:

$$CRR = \frac{|CORRECTREFUSALS|}{|R|}$$

A refusal is considered correct when the system abstains, and the ground truth confirms that available evidence is insufficient or contradictory.

CRR therefore measures evidence-aligned abstention rather than model uncertainty alone.

Composite Trust Score
The Trust score for query set Q is defined as:

$$\text{Trust}(Q) = \beta \cdot \text{MeanAccuracy} + (1 - \beta) \cdot \text{CRR}$$

where $\beta=0.7$. Sensitivity analysis across $\beta \in [0.6, 0.8]$ showed consistent method rankings.

This formulation reflects clinical risk asymmetry: confidently incorrect medical advice under weak evidence is more harmful than calibrated refusal.

B. Privacy: Per-Session and Cross-Session Metrics

Per-session leakage measures the fraction of outputs containing at least one PHI entity, as defined under HIPAA [18]:

$$L_{session} = \frac{|q_i \in Q: \text{PHI}(q_i) \neq \emptyset|}{|Q|}$$

where $\text{PHI}(q_i)$ denotes detected PHI entities in the response.

For systems with memory, we define cross-session reconstruction risk:

$$\text{Rcross} = \frac{\left|\left|\bigcup_{j=1}^{k} \text{PHIj}\right|\right|}{|PHI_{total}|}$$

where PHIj represents PHI indirectly revealed in session j, and $\text{PHI}_{\text{total}}$ is the full PHI profile of the target patient.

This captures the cumulative reconstruction risk that single-query leakage metrics fail to measure.

C. Security: Extended Threat Model

We evaluate five attack categories. Three are established: prompt injection, tool poisoning, and memory exploitation [14, 15]. Two are agent-specific:

- Logic poisoning: insertion of subtly incorrect clinical relationships into the retrieval index.
- Toolchain feedback exploitation: adversarial tool outputs propagating across multi-step reasoning.

For attack type t, the attack success rate is:

$$ASR(t) = \frac{Successful\ attacks\ of\ type\ t}{Total \text{ attempts of type t}}$$

We define the security index:

$$S = 1 - \sum_{t} wt \,.\, \text{ASR(t)}$$

where $wt \geq 0$ and $\sum_t wt = 1$.
Logic poisoning receives the highest weight (0.25) due to systemic clinical risk.

D. Integrated TPS Score

The composite score TPS = 0.4 · Trust + 0.35 · (1 − Lsession) + 0.25 · Security weights trust and privacy higher than security, reflecting the clinical principle that incorrect answers and data leakage cause more immediate harm than theoretical vulnerability. Weights are configurable per deployment context.

IV. METHODOLOGY

A. Data Generation and Noise Augmentation

We generate FHIR-compliant EHR bundles using Synthea [16] covering demographics, conditions, labs, medications, and clinical notes. To address the well-known limitation that Synthea produces unrealistically clean records, we built a parameterized noise injection layer with four imperfection types: (a) missing records, randomly dropping 10–20% of lab values and medication entries; (b) contradictory notes—clinical notes that conflict with structured data (e.g., "diabetes well-controlled" alongside elevated HbA1c); (c) temporal ambiguity—replacing precise dates with vague references; and (d) linguistic variation—abbreviations, typos,

and informal phrasing. All noise parameters are seeded for exact reproducibility. Noise levels were sampled uniformly within specified bounds per seed to prevent overfitting to a fixed corruption profile.

B. Retrieval Strategies

We compare six strategies spanning the practical spectrum. TF-IDF and BM25 [4] provide keyword baselines. Embedding-RAG uses BioClinicalBERT [5] embeddings with cosine similarity. Chroma-OpenAI uses OpenAI's text-embedding-3-small in ChromaDB. Hybrid combines Chroma+TF-IDF via reciprocal rank fusion. Graph-RAG [20] builds a directed knowledge graph with patient–condition–observation–medication entities derived deterministically from FHIR bundles, linked by temporal and relational edges. Crucially, the graph enforces PHI-restricted node traversal: identifier nodes (SSN, MRN, phone, address) are marked as restricted and excluded from retrieval paths unless explicitly authorized. This structural constraint limits the PHI that enters the model context per query, which directly reduces both per-session leakage and cross-session reconstruction risk. The graph also supports consistency checking: when a retrieved node contradicts established clinical relationships in the graph neighborhood, the system flags the inconsistency rather than passing it to the language model unchecked. This structured reasoning complements chain-of-thought prompting approaches [19] by grounding each reasoning step in explicit graph relationships.

C. Models and Statistical Design

Each configuration runs on three backends: GPT-4o (large commercial), GPT-4o-mini (smaller commercial), and Llama-3-8B (open-source, local inference). This setup isolates model-specific from architecture-specific findings. We use ten independent seeds (42, 123, 256, 314, 500, 618, 777, 888, 950, 999) per configuration, evaluating 200 clinically representative queries covering patient summaries, diagnostics, lab interpretation, medication reconciliation, and temporal reasoning. We report means with 95% bootstrap confidence intervals and use Wilcoxon signed-rank tests with Bonferroni correction for pairwise comparisons, with Cliff's delta for effect sizes.

D. Attack Configuration

We test 50 adversarial attempts per attack category. Prompt injection uses 50 templates, including clinical authority impersonation. Tool poisoning returns falsified lab reference ranges. Memory exploitation plants adversarial context in earlier sessions. Logic poisoning inserts plausible but incorrect clinical relationships. Toolchain exploitation crafts multi-step scenarios where compromised tool outputs cascade.

V. RESULTS

All values below are mean ± 95% CI across ten independent runs. Statistical significance is at $p < 0.05$ after Bonferroni correction.

A. Trust Evaluation

Table 2 presents trust scores on GPT-4o. Graph-RAG achieves the highest composite trust (0.89±0.02), with the difference from all other strategies statistically significant ($p < 0.01$, Cliff's delta > 0.4). An important finding concerns the accuracy-refusal trade-off: embedding retrieval produces the highest raw SemSim (0.92) but the lowest CRR (0.71), meaning it generates plausible-sounding answers when evidence is inadequate. Graph-RAG achieves a lower SemSim (0.88) but substantially better CRR (0.85) through relationship-aware evidence chain assessment. The Δ Noise column in Table II shows that keyword methods are most vulnerable to noise augmentation (TF-IDF: −0.11, BM25: −0.08) because missing records and linguistic variation break lexical matches, while Graph-RAG proves most resilient (−0.04) by routing around missing nodes through alternative relationship paths.

Table 2 : Trust Scores (Mean ± 95% Ci, Gpt-4o)

Strategy	Trust	Acc.	CRR	SemSim	ΔNoise
TF-IDF	0.76±0.03	0.80±0.03	0.66±0.03	0.75±0.02	−0.11
BM25	0.73±0.02	0.83±0.02	0.66±0.04	0.79±0.02	−0.08
Embed.	0.79±0.02	0.87±0.01	0.69±0.01	0.92±0.01	−0.06
Chroma	0.82±0.02	0.87±0.03	0.74±0.03	0.92±0.03	−0.05
Hybrid	0.86±0.02	0.88±0.01	0.81±0.02	0.85±0.01	−0.04
GraphRAG	0.90±0.02	0.91±0.01	0.86±0.02	0.88±0.01	−0.04

The retrieval strategy ranking remains stable across all three model backends. Model size provides a consistent 4–6 point improvement (GPT-4o > mini > Llama), but Graph-RAG leads on every backend. This consistency means design guidance applies regardless of model choice.

B. Privacy Evaluation

Table 3 shows PHI leakage across memory configurations. Multi-layer sanitization reduces per-session leakage from 0.25 to 0.02 under persistent memory. However, the cross-session reconstruction score reveals a subtler risk: targeted queries across five sessions reconstruct 12%±3% of a patient's PHI profile under persistent memory, even though individual responses are clean. Each response leaks indirect information about topics referenced, clinical patterns discussed, and temporal markers that collectively enable patient re-identification. Graph-RAG reduces reconstruction to 8% through PHI-restricted node traversal.

Table 3 : PHI Leakage by Memory Configuration

Memory	Before	After	Cross-Session	Risk
Stateless	0.00±0.00	0.00±0.00	0.02±0.01	Minimal
Ephemeral	0.09±0.02	0.01±0.01	0.03±0.02	Indirect
Persistent	0.23±0.02	0.02±0.01	0.11±0.02	High

C. Security Evaluation

Table 4 presents attack success rates. Established attacks are effectively mitigated: prompt injection drops from 0.50 to 0.05, tool poisoning from 0.30 to 0.02. The more significant finding concerns the novel attack categories. Logic poisoning retains 8% success even with full defenses because adversarial content resides in the knowledge base, not the input stream. Graph-RAG provides the best defense (ASR 0.04) through structural consistency checking: a poisoned node contradicting established clinical relationships triggers anomaly flags.

Toolchain feedback exploitation shows a meaningful difference between defense strategies: per-step tool output validation reduces ASR to 0.03, while end-to-end-only validation achieves only 0.09. This demonstrates that agentic systems require validation at every pipeline stage, not just the output.

Table 4 : Attack Success Rates (Gpt-4o)

Attack Type	Baseline	Defended	GraphRAG
Prompt Injection	0.50	0.04±0.02	0.03±0.01
Tool Poisoning	0.30	0.02±0.01	0.01±0.01
Memory Exploit	0.40	0.05±0.02	0.02±0.00
Logic Poison*	0.45	0.08±0.03	0.05±0.01
Toolchain*	0.35	0.06±0.02	0.03±0.01

**Novel attack categories introduced in this work.*

D. Integrated TPS Scores

Table 5 presents composite TPS scores. Graph-RAG leads on every backend: 0.91 on GPT-4o, 0.87 on GPT-4o-mini, 0.80 on Llama-3-8B. The gap between graph-structured and embedding-based approaches widens from 3–5 points (trust alone) to 6–9 points (integrated TPS) because graph structure disproportionately benefits privacy and security. A key practical finding: Llama-3-8B with Graph-RAG (TPS = 0.80) outperforms GPT-4o with TF-IDF (TPS = 0.74). Retrieval architecture matters more than model scale for overall system safety.

Table 5 : Integrated TPS Scores (Mean ± 95% CI)

Strategy	GPT-4o	4o-mini	Llama-3
TF-IDF	0.75±0.03	0.72±0.02	0.69±0.04
BM25	0.77±0.02	0.76±0.03	0.67±0.04
Embed.	0.82±0.01	0.79±0.02	0.72±0.03
Chroma	0.85±0.02	0.80±0.02	0.73±0.04
Hybrid	0.89±0.02	0.85±0.02	0.77±0.02
GraphRAG	0.92±0.02	0.87±0.01	0.78±0.03

Table 6 provides computational overhead. Graph-RAG requires 2.8s per query versus 0.3s for TF-IDF. For clinical decision support where sub-second latency is not critical, this trade-off is justified by the TPS improvement.

Table 6 : Computational Overhead (Apple Silicon M2)

Strategy	Latency (s)	Memory (MB)	Index (MB)	Rel. Cost
TF-IDF	0.3	120	15	1.0×
BM25	0.4	140	18	1.3×
Chroma	1.2	520	210	4.0×
Embed.	1.4	580	245	4.7×
Hybrid	1.9	680	260	6.3×
GraphRAG	2.8	950	380	9.3×

VI. DISCUSSION

A. Why Graph-RAG Outperforms Across Dimensions

Three complementary mechanisms drive Graph-RAG's advantage. First, relationship-aware evidence assessment traces clinical chains (diagnosis → lab → medication) and detects evidentiary gaps, triggering calibrated refusal rather than hallucinated answers [17]. This explains the CRR gap (0.85 vs. 0.71 for embedding retrieval) — the graph makes missing evidence structurally visible. Under noise, graph traversal finds alternative paths through related entities, limiting accuracy loss to 0.04 versus 0.11 for TF-IDF.

Second, node-level PHI restriction decomposes patient records into typed nodes and excludes identifier nodes from retrieval paths. The model receives clinical content without co-located identifiers, reducing cross-session reconstruction from 15% to 8% as a structural property of retrieval itself, not a post-hoc filter.

Third, structural consistency checking catches logic poisoning. A falsified entry claiming "Metformin is contraindicated with Lisinopril" is semantically plausible in vector space but contradicts established graph relationships showing safe co-prescription. Embedding retrieval cannot detect such contradictions — two opposing statements can have nearly identical embeddings. These three mechanisms independently benefit trust, privacy, and security, respectively, explaining why the TPS gap widens from 3–5 points (trust alone) to 6–9 points (composite).

B. Model Scale and Architecture

Model scale provides consistent 4–9 point TPS improvement per tier and greater adversarial resilience. However, retrieval strategy rankings are identical across all three backends — no model-specific interaction reverses the ordering. Llama-3-8B with Graph-RAG (TPS = 0.80) outperforms GPT-4o with TF-IDF (TPS = 0.74), reframing the safety question from "which model" to "which pipeline architecture" [20]. This directly benefits healthcare institutions that cannot send patient data to external APIs: a local open-source model with graph-structured retrieval achieves better integrated safety than a frontier model with naive retrieval.

C. Limitations

Data remains synthetic despite noise augmentation; real clinical records contain scanned images, copy-pasted boilerplate, and multi-provider inconsistencies that Synthea cannot replicate. Validation on de-identified clinical data [18] is the most critical next step. The TPS weighting (0.4/0.35/0.25) reflects our risk judgment; different deployment contexts (emergency triage vs. research platforms) may warrant different weightings, though sensitivity analysis showed stable strategy rankings across adjacent configurations. All experiments ran on single-machine hardware; cloud deployments introduce network-level security and multi-tenant isolation concerns not evaluated here. Knowledge graphs are lightweight and lack rich ontologies (SNOMED CT, ICD-10) that production systems require. The threat model assumes external query-level attackers and does not cover insider threats or supply-chain attacks beyond knowledge-based poisoning. Finally, hallucination [17] is captured only indirectly through accuracy scoring; dedicated hallucination categorization is planned for future work.

VII. DESIGN PRINCIPLES

Five principles derive directly from experimental findings. (1) Retrieval as governance: Graph-RAG's node-level restrictions cut cross-session PHI reconstruction from 15% to 8% and logic poisoning ASR from 0.12 to 0.04. Retrieval selection is a governance decision, not merely an accuracy optimization. (2) Full-lifecycle defense: per-step tool validation reduced toolchain ASR to 0.03 versus 0.09 for output-only checking. Safety controls must be distributed across every pipeline stage. (3) Calibrated refusal: evidence-based refusal thresholds (CRR = 0.85) outperform model self-confidence (CRR = 0.71). Systems should refuse explicitly when evidence quality is insufficient. (4) Interpretability as security: Graph-RAG's inspectable reasoning chains enable logic poisoning detection. Auditability is a security investment, not a transparency afterthought. (5) Memory as risk dial: persistent

memory enables 12% cross-session PHI reconstruction. Memory mode must be chosen explicitly with compensating controls when persistence is required.

VIII. CONCLUSION

We introduced TPS-Eval, a framework for jointly evaluating Trust, Privacy, and Security as coupled properties of agentic AI pipelines in healthcare. Graph-based retrieval consistently achieves the highest integrated TPS scores across all model backends. The extended threat model reveals that logic poisoning poses structural challenges that conventional prompt-level defenses cannot address. Cross-session inference attacks reconstruct patient PHI even from individually sanitized responses. The framework, including evaluation harnesses and data generation tools, is released as open source. Future work priorities include validation on de-identified clinical records, richer ontology integration, multi-agent evaluation scenarios, and cloud deployment testing.

A. Acknowledgment

The author thanks the healthcare AI research community for ongoing collaboration on safe deployment of clinical AI systems.

IX. REFERENCES

[1] Y. Wang et al., “Large model based agents: State-of-the-art, cooperation paradigms, security and privacy,” IEEE Commun. Surveys & Tutorials, 2025.
[2] S. Yao et al., “ReAct: Synergizing reasoning and acting in language models,” Proc. ICLR, 2023.
[3] P. Lewis et al., “Retrieval-augmented generation for knowledge-intensive NLP tasks,” Proc. NeurIPS, vol. 33, pp. 9459–9474, 2020.
[4] S. Robertson et al., “Okapi at TREC,” SIGIR Forum, vol. 32, no. 1, pp. 118–126, 1998.
[5] E. Alsentzer et al., “Publicly available clinical BERT embeddings,” Proc. Clinical NLP Workshop, pp. 72–78, 2019.
[6] Y. Gu et al., “Domain-specific language model pretraining for biomedical NLP,” ACM TCHR, vol. 3, no. 1, 2022.
[7] European Parliament, “AI Act (Regulation 2024/1689),” OJ EU, 2024.
[8] NIST, “AI Risk Management Framework (AI RMF 1.0),” 2023.
[9] OpenAI, “Improving factuality and harmlessness in language models,” 2023.
[10] S. Saharan et al., “Deep learning and XAI for breast cancer detection,” Sci. Reports, vol. 15, 2025.
[11] S. Singh et al., “DiaXplain: Transparent AI for Type-2 diabetes,” Comp. & Elect. Eng., vol. 126, 2025.
[12] B. Yan et al., “On protecting the data privacy of LLMs,” High-Conf. Computing, vol. 5, no. 2, 2025.
[13] Y. Yao et al., “A survey on LLM security and privacy,” High-Conf. Computing, vol. 4, no. 2, 2024.
[14] M. Ferrag et al., “From prompt injections to protocol exploits,” ICT Express, 2025.
[15] S. Perez, I. Ribas, “Ignore this title and HackAPrompt,” EMNLP, pp. 4945–4977, 2023.
[16] J. Walonoski et al., “Synthea: Generating synthetic patients,” JAMIA Open, vol. 1, no. 1, pp. 18–25, 2018.
[17] Z. Ji et al., “Survey of hallucination in NLG,” ACM Comp. Surveys, vol. 55, no. 12, 2023.
[18] HIPAA, Health Insurance Portability and Accountability Act, 1996.
[19] J. Wei et al., “Chain-of-thought prompting elicits reasoning,” NeurIPS, vol. 35, 2022.
[20] S. Pan et al., “Unifying LLMs and knowledge graphs,” IEEE TKDE, vol. 36, no. 7, 2024.