Evaluating tool retrieval with policy compliance through knowledge graphs and language models

Zachary Grider[1], Alexander Nelson[1,*]

[1]University of Arkansas, Department of Electrical Engineering and Computer Science, 1 University of Arkansas, Fayetteville, AR 72704, USA

Abstract

Tool-calling is an ability which allows language models to perform actions using remote tools. The Model Context Protocol (MCP) is an interface which standardises tool interfaces. Tools are frequently hosted by large providers and come with a set of policies. It is crucial for local agents in the form of small language models (SLMs) to be able to select tools which respect users' privacy preferences. In this work, we use the OPP-115 dataset to identify attributes and values in segments of policies which address data concerns such as first party sharing, third party sharing, security and retention. We synthesize this dataset with a subset of the ToolBench dataset to form a dataset of tool descriptions and associated policies. We then develop a LLM-based retriever which evaluates tools and associated policies to determine compliance with the user's desired privacy policy. We develop a knowledge graph with structured schema to query and retrieve tools to evaluate triple-based policy representations compared to standard policy text.

Keywords

Language Models, Retrieval-augmented Generation, Policy Compliance, Knowledge Graphs, KG-RAG

1. Introduction

Large language models (LLMs) have become the state-of-the-art (SOTA) for knowledge-related tasks. These models are capable of generating responses based upon their implicit (parametric) knowledge due to the vast corpora of text which they are trained upon. While powerful, LLMs suffer from fixed knowledge and "hallucinations" [1], which prompted research into solutions like retrieval-augmented generation (RAG) [2]. RAG consists of a system which augments the language model by providing an external store which embeds information, and a retriever to match documents to queries. GraphRAG [3] is RAG where a graph is used for storing facts, and has benefits over vanilla RAG [4].

RAG systems enable grounded answers, but LLMs are capable of more than text retrieval tasks. LLMs were trained to perform tool-calling which provides a language model with capabilities such as web search or sending email. Later, the Model-Context Protocol (MCP) was established to provide a standardized way for language models to access external APIs and data sources for tool-calling.

LLMs are powerful but resource-intensive [5]. Small language models (SLMs) which have similar capabilities but require a fraction of the resources [6], which also makes them suitable for on-device tasks [7].

SLMs on mobile devices are common, but they can be intrusive due to sensitive data which puts user privacy at risk. These concerns are amplified using remote tool-calls for handling user data. Previous work shows LLMs become confused when they are presented with many tool choices, though solutions using RAG for storing tool definitions show promise [8]. This may be viable for on device SLMs to prevent performance loss when presented with many tools. Simultaneously, the choices of these tools should respect user privacy policies. There is little work evaluating SLMs for tool-calling with privacy compliance [9], and a lack of exploring capabilities of SLMs for handling scenarios with

RAGE-KG 2026: The Third International Workshop on Retrieval-Augmented Generation Enabled by Knowledge Graphs, co-located with ISWC 2026, October 25–26, 2026, Bari, Italy

*Corresponding author.

zg009@uark.edu (Z. Grider); ahnelson@uark.edu (A. Nelson)

https://ahnelson.uark.edu/ (A. Nelson)

0009-0009-6235-6487 (Z. Grider); 0000-0003-3383-0193 (A. Nelson)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

published 2026-11-04

CEUR-WS.org/Vol-4312/short31.pdf

an overwhelming number of tool calls or "tool overload". To address this, we conduct a preliminary investigation of SLM capabilities in tool overload, user privacy policy compliance, and tool-policy choices. We also compare graph-based (triple) and plaintext representations of policy segments within a GraphRAG framework.

Research Questions:

1. How robust is SLM tool selection under tool overload?
2. How accurate are SLMs at determining user privacy policy compliance?
3. Are SLMs capable of respecting user privacy preferences under tool overload?
4. Does triple-based representation of tool policies in KG-RAG frameworks improve policy compliance?

The primary contributions of this work are:

1. We synthesize a novel knowledge graph consisting of privacy policy segments generated from the OPP-115 dataset, candidate queries, and their relevant tools.
2. We observe the accuracy of SLMs in tool overload scenarios, privacy policy compliance, and privacy policy compliant tool selection. We observe changes based on triple or plaintext representations of candidate policies.

Section II covers background work regarding privacy-oriented knowledge graphs, tool calling, RAG and KG-RAG, and policy compliance. Section III will cover the methodology in which we generated the synthetic KG, evaluated tool bloat, and evaluated SLMs capabilities to determine compliance with user privacy policies. Section IV will demonstrate the results from the methodological section. Section V will discuss the preliminary findings. The work concludes in Section VI.

2. Background

Knowledge graphs (KGs) [10] are a way to represent data using entities and relations (E_1, R, E_2) where E_1 and E_2 are entities in the KG, and R is a relation. Knowledge graphs have found usage in privacy policy analysis [11] and privacy policy compliance auditing [12]. In [13] the authors use an ontology and an LLM-generated ontology for KG generation based on AI policy documents. They rigorously evaluate multiple LLMs across multiple policy QA tasks, finding the open LLM schema and KG augmentation improve results.

Tool-calling enables language models to perform real-world tasks through API usage. LLMs show impressive capacity for tool use [14], which has motivated the need for both benchmarks [15, 16] and standardized interfaces for LLMs to interact with tools [17]. Tool selection accuracy decreases as available candidate tools increases [18]. RAG-MCP [8] addresses tool overload by indexing MCP tools and retrieval, showing substantial increases in tool accuracy.

Retrieval-augmented generation (RAG) [2] allows LLMs access to grounded documents through a store and retriever. Knowledge graph-based RAG [19] substitutes a semantic store for a knowledge graph for fact-based documents, creating new retrieval approaches [3]. Graph-based RAG has proven successful across in multiple fields, including policy [20] and legal [21] domains. The authors of [22] develop a privacy-protected RAG framework, improving over comparable architectures while preserving privacy.

Privacy policy reasoning is important for ensuring users are respected and aware of what information they are revealing. There are a plethora of privacy policy datasets, including benchmarks for natural language understanding [23] and legal reasoning in LLMs [24]. OPP-115 [25] is a widely used benchmark dataset for privacy policy segment classification annotated by legal experts. PrivComp-KG [26] is a knowledge graph developed using an ontology to enable automated compliance reasoning for privacy policies and was evaluated using RAG against OPP-115. PrivacyBench [27] is a framework evaluating privacy leakage of personal AI in RAG systems, demonstrating privacy risks in LLMs.

3. Methodology

3.1. Dataset

To evaluate the performance of small LMs on tool-calling with privacy compliance, we synthesized the OPP-115 and a subset of the ToolBench dataset. The OPP-115 policy text was cleaned using an HTML parser. Both were selected due to being standardized benchmarks in their respective area. This serves as a representative sample of tools and policies for this pilot study.

3.2. KG Construction/Evaluation Dataset Synthesis

To include tools tools in our KG, we isolated the tools, resulting in 564 unique tools. We defined a relation between queries and their relevant tools.

To create the privacy policy portion of the KG, we extracted the unique segments of each privacy policy segment in the OPP-115 dataset. The OPP-115 dataset offers rich annotation descriptions relating to various category-attribute-value tuples. Each category has a list of attributes and values which further describe each segment. There are 259 unique category-attribute-value triples, which were loaded into the KG and provided relations to segment instances.

For generating the example user privacy preferences, we manually generated short sentences, with the user preferences set covering combinations of three category-attribute-value pairs.

3.3. Tool Overload of SLMs

In this first experiment, we evaluate the accuracy of the SLMs for appropriately selecting the tool provided a set of tools for a given user query. We provide a list of tools ranging from size 10 to 100 with increments of 10. We selected 30 randomly sampled tool requests from the dataset. Each query was bound to one correct tool. We select one tool for correctness in the complete tool listing.

We created three experimental setups: correct tool in a fixed position, in a random position, and no correct tools. We evaluate the effects of tool overload under various conditions, by repeating the same queries but increasing the amount of tools from 10 to 100 in steps of 10. We evaluate the accuracy of the models across tested tool sizes and each tool context setup. For all experiments, the small language models we evaluated were Granite4 with 3 billion parameters [28] and Qwen3 with 8 billion [29] with temperatures of 0.

3.4. Policy Compliance of SLMs

To evaluate policy compliance, We evaluate three conditions: explicit user-tool policy compliance, explicit user-tool policy non-compliance, and tool policy is unrelated, and either use triple or plaintext representations.

To extract appropriate policies which fit the criteria, we had to create queries which allowed the mutual exclusion modeling. Policy compliance in our work is expressed as

$$T_p \succcurlyeq U_p \iff \forall c \in \mathscr{C}(U_p),\ T_p \succcurlyeq c, \tag{1}$$

where T_p is the policy of a tool, U_p is the user policy, and $\mathscr{C}(U_p)$ is the set of all conditions required for compliance of a given user policy, where $\succcurlyeq$ denotes policy compatibility where the LLM determines if the tool policy is compliant, and is validated against the existence of the set of category-attribute-value triples for the given policy segment within the knowledge graph. Given these conditions, there are 29 user privacy preference queries corresponding to the seven user privacy preference texts. We attempted to select 10 compliant policy segments for each query, however four queries returned no results from the knowledge graph and others had less than 10 unique segments.

3.5. Tool-with-Policy Compliance

For tool and policy compliance, we reuse user queries used in 3.3 for the KG-RAG system, and used the tool segments in 3.4 as tool policies. Tool policies are randomly sampled from the set of available OPP-115 policy segments. We evaluate four conditional setups: correct tool with policy compliance, incorrect tools with one compliant policy, incorrect policies with one correct tool, and incorrect tools and policies. A relevant tool is randomly retrieved from the KG. We evaluate all seven user policies for each query, and create the tool-policy pairs for each condition. Queries which had no segments are excluded. Due to the composition of user policies, user policies may be satisfied by multiple KG queries. We selected 20 tool-policy pairs for the experiment. When there are few segments, we randomly resample from selected segments for complete pairs.

3.6. Metrics

The metric reported is accuracy. Since this is a retrieval task based on whether the correct tool, policy, or combination is chosen, this creates binary outputs suitable for accuracy measurements.

3.7. Hardware/Versions/Implementation Details

All experiments were performed on a Dell XPS using an RTX3080 graphics card. Experiments were ran on Ubuntu 22.04 using WSL2. Repository: https://github.com/AESIRLab/ToolAndPolicyCompliance-WithKGs .

3.8. Describing Tool Privacy Policies Respecting User Privacy Preferences

Privacy policy preference P_4 states: "The API must not allow first-party collection or use. The API must not allow third-party sharing or collection." This represents a direct composition of privacy policies P_1 and P_2. In our knowledge graph, the corresponding category-attribute-value triple which complies with P1 is "First Party Collection/Use"-"Does/Does Not"⟶"Does Not" which in plaintext is the policy segment does not allow first party collection or use, meaning it respects the user's privacy preference by not collecting or using data with first parties. It also explicitly excludes policies which also include the triple "First Party Collection/Use"–"Does/Does Not"⟶"Does", which means first party data collection or sharing does occur according to the interpretation of the text, violating the user's privacy preferences. This is necessary as some triples have overlapping annotations in the dataset, meaning explicit exclusion is necessary.

The second case is explicitly policy non-compliance which is a policy segment annotated with "First Party Collection/Use"–"Does/Does Not"⟶"Does" and explicitly excludes "First Party Collection/Use"–"Does/Does Not"⟶"Does Not" due to the overlapping annotation triples mentioned above.

The third case for policies is that a segment does not mention text corresponding to a user's privacy preferences. This segment could have any annotation where the category "First Party Collection/Use" and attribute "Does/Does Not" is not mentioned. For this work, this is considered as a failure of user privacy preference as the segment does not explicitly mention satisfying the user privacy preference or implicit policy non-compliance.

A policy which complies with P2 has the triple "Third Party Sharing/Collection"–"Does/Does Not"⟶"Does Not" which means third party sharing does not occur, respecting the user's privacy preferences. Similarly, "Third Party Sharing/Collection"–"Does/Does Not"⟶"Does" triples are excluded from the positive set due to the ability for segments to be annotated with both instances. For the explicitly non-compliant policies, the inverse annotations are used. Similarly to P1, scenarios in which a policy segment is annotated with no "Third Party Sharing/Collection" categories and no "Does/Does Not" attributes are considered not mentioned, and are considered as implicitly non-compliant.

Since P4 is a composition of P1 and P2, a policy segment which satisfies a user preference P4 must have the category–attribute–value triples "First Party Collection/Use"–"Does/Does Not"⟶"Does Not" and "Third Party Sharing/Collection"–"Does/Does Not"⟶"Does Not" and explicitly not include 'First Party

Collection/Use"–"Does/Does Not"→"Does" and "First Party Collection/Use"–"Does/Does Not"→"Does". An explicitly non-compliant policy could either include the "First Party Collection/Use"–"Does/Does Not"→"Does" triple or the "Third Party Sharing/Collection"–"Does/Does Not"→"Does" meaning either first party collection and use or third party collection occurs, and either case violates the user's privacy preferences. Implicit non-compliance in this scenario is a policy segment which does not mention all of the possible triples in the explicit cases.

Examples can be found in B.

4. Experiments and Results

4.1. Tool Overload Evaluation

To evaluate tool overload scenarios, we extracted 30 randomly sampled tool requests and evaluated on tool positions and incorrect tools, with available tools ranging from 10-100 incrementing by 10. There were 1767 records, with 33 missing from Qwen's experiments due to no response. The results are shown in Table 1.

Table 1
Tool Overload Evaluation Accuracy

	Granite			Qwen		
Tool Limit	CT+FP	CT+RP	FT	CT+FP	CT+RP	FT
10	63.3	73.3	10.0	80.0	56.7	90.0
20	46.7	73.3	6.7	60.0	63.3	90.0
30	56.7	63.3	3.3	60.0	73.3	83.3
40	60.0	76.7	3.3	63.3	66.7	86.7
50	50.0	73.3	6.7	66.7	50.0	70.0
60	53.3	46.7	3.3	60.0	56.7	76.7
70	63.3	43.3	0.0	53.3	70.0	66.7
80	46.4	53.3	0.0	71.4	50.0	71.4
90	50.0	40.0	0.0	66.7	60.0	56.7
100	46.7	50.0	0.0	70.0	63.3	50.0

$^{CT+FP}$ Correct Tool Present, Fixed Position
$^{CT+RP}$ Correct Tool Present, Random Position
^{FT} Incorrect Tools

We observe a general downwards trend among the accuracy of tool retrieval as the tools available in the prompt succeed. Qwen's capability to correctly ignore all tools is an interesting observation. Granite performs oppositely with catastrophic accuracy loss. Qwen has improved accuracy for choosing correct tools.

4.2. Policy Selection Evaluation

To evaluate policy compliance capability, we extracted 10 segments to determine noncompliance with user policies. There are 29 KG queries. This dataset has 346 records as some queries have no valid segment. The results are shown in 2.

We observe high accuracy evaluating non-compliance in both models using triples and plaintext. Granite's accuracy fails where compliant policies use triples. Qwen's overall accuracy is low. Both models increase accuracy on plaintext policies. [1]

[1]P7 is excluded from the compliant policy results due to no queries in the KG satisfying the constraints.

Table 2
Policy Compliance Accuracy

Compliant Policy	User Preference	Granite Plaintext	Granite Triples	Qwen Plaintext	Qwen Triples
Compliant[1]	**P1**	50.0	0.0	40.0	70.0
	P2	60.0	0.0	50.0	60.0
	P3	20.0	0.0	50.0	0.0
	P4	80.0	0.0	30.0	50.0
	P5	0.0	0.0	0.0	0.0
	P6	0.0	0.0	100.0	0.0
Non-compliant	**P1**	90.0	95.0	70.0	95.0
	P2	95.0	95.0	55.0	70.0
	P3	85.0	100.0	85.0	95.0
	P4	73.9	91.3	78.3	95.7
	P5	78.9	100.0	100.0	94.7
	P6	81.2	100.0	100.0	93.8
	P7	92.3	100.0	92.3	92.3

4.3. Tool and Policy Selection Evaluation

To evaluate the performance for tool-policy pairs given a user query and user privacy policy, we sampled 5 queries from 4.1. We constructed four experiments with 20 candidate tools and up to 10 candidate segments. We evaluated the four experiments of tool-policy agreement for plaintext and triple policies. This dataset consisted of 944 records, since 26 of the queries failed due to the LLM service experiencing a request timeout. The results are found in Table 3.

Both models have high accuracy with correct tools and policies. Qwen has high accuracy given the correct tool-incorrect policy. Accuracy increases on incorrect tools regardless of policy or representation. Granite accuracy collapses completely. It has low to zero accuracy on the 3 negative experiments, performing worse with triples representation.

5. Discussion

In 4.1, we observe a comparable accuracy for returning correct tools. However, when incorrect tools are present Qwen responds more accurately when tools are incorrect. Granite exhibits a persistent bias towards selecting a tool regardless of the presence of correct tools.

In 4.2, both models are accurate at determining non-compliance of policies regardless of policy representation. Conversely, accuracy for both models is notably lower in evaluating compliant policies. A more interesting initial finding is that structured triple representation had consistent lower accuracy. This suggests that current SLMs struggle with inline triple reasoning.

For 4.3, Granite's accuracy collapses for scenarios where no solution exists. We hypothesize this behavior is transitive from 4.1, as Granite returns tools when none are valid. While Qwen has consistently higher accuracy, it does demonstrate a similar pattern. This is exhibited in the presence of incorrect policies the correct tool causes accuracy to drop regardless of policy representation. Further, triple representations consistently impair compliance verification in both models, while remaining robust for detecting non-compliance.

5.1. Limitations and Scope

This work focuses on characterizing failure modes in SLMs under privacy and tool overload constraints. The OPP-115 dataset was cleaned, however artifacts remain from annotations. The evaluation currently was intentionally bounded to a set of high-precision queries and policy pairs to analyze model behaviors

Table 3
Policy and Tool Evaluation Accuracy

Experiment	User Preference	Granite Plaintext	Granite Triples	Qwen Plaintext	Qwen Triples
CT+CP	**P1**	44.4	80.0	60.0	22.2
	P2	55.6	55.6	77.8	66.7
	P3	70.0	70.0	80.0	90.0
	P4	66.7	70.0	40.0	60.0
	P5	66.7	88.9	33.3	50.0
	P6	62.5	80.0	75.0	33.3
CT+IP	**P1**	0.0	0.0	30.0	80.0
	P2	0.0	0.0	30.0	50.0
	P3	0.0	0.0	50.0	20.0
	P4	0.0	0.0	100.0	100.0
	P5	0.0	0.0	20.0	80.0
	P6	0.0	0.0	20.0	60.0
IT+CP	**P1**	37.5	0.0	100.0	100.0
	P2	10.0	0.0	88.9	90.0
	P3	10.0	0.0	100.0	90.0
	P4	0.0	0.0	100.0	100.0
	P5	22.2	0.0	100.0	100.0
	P6	20.0	0.0	90.0	90.0
IT+IP	**P1**	10.0	0.0	80.0	100.0
	P2	10.0	11.1	100.0	80.0
	P3	0.0	0.0	100.0	100.0
	P4	20.0	10.0	100.0	100.0
	P5	30.0	0.0	100.0	100.0
	P6	0.0	0.0	90.0	100.0

$^{CT+CP}$ Correct Tool for Query and Compliant Tool Policy
$^{CT+IP}$ Correct Tool for Query and Non-compliant Tool Policy
$^{IT+CP}$ Incorrect Tool for Query and Compliant Tool Policy
$^{IT+IP}$ Incorrect Tool for Query and Non-compliant Tool Policy

in specific failure conditions. Future work can scale this benchmark across the complete OPP-115 and ToolBench datasets and incorporate larger frontier models as upper-bound baselines and report more robust statistical variance across evaluation runs. Finally, a small number of execution records failed due to model formatting output errors (e.g. failing to follow response constraints) and were omitted from accuracy calculations as opposed to manual disambiguation to prevent user-determined bias.

6. Conclusion

In this preliminary work, we have created a KG with user queries relating to tool calls, and policy segments represented in both plaintext and KG triples with KG-RAG. We evaluated the accuracy of SLMs under tool bloat scenarios, user privacy policy compliance scenarios for increasingly complex user policies, and in various tool-with-policy reasoning experiments. We demonstrate discrepancies in models for selecting tools, and hypothesize models focus more on tool-correctness than policy-compliance, which can expose a privacy risk in SLMs for tool-calling and compliance reasoning. This demonstrates the need for more research on SLM capabilities in privacy scenarios where user data is at risk. Future work involves more rigorous semantic policy-tool evaluation over the whole dataset, and comparing the performance of local-capable SLMs with foundation models.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 2237945 and 2113903.

Declaration on Generative AI

The author(s) used GPT-5.5 for: content enhancement, formatting assistance, grammar and spelling check. The authors reviewed the generative content and take full responsibility for the content of the publication.

References

[1] A. T. Kalai, O. Nachum, S. S. Vempala, E. Zhang, Why language models hallucinate, 2025. URL: https://arxiv.org/abs/2509.04664. arXiv:2509.04664.

[2] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, Retrieval-augmented generation for knowledge-intensive nlp tasks, Advances in neural information processing systems 33 (2020) 9459–9474.

[3] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, D. Metropolitansky, R. O. Ness, J. Larson, From local to global: A graph rag approach to query-focused summarization, 2025. URL: https://arxiv.org/abs/2404.16130. arXiv:2404.16130.

[4] X. Zhu, Y. Xie, Y. Liu, Y. Li, W. Hu, Knowledge graph-guided retrieval augmented generation, in: L. Chiruzzo, A. Ritter, L. Wang (Eds.), Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, Albuquerque, New Mexico, 2025, pp. 8912–8924. URL: https://aclanthology.org/2025.naacl-long.449/. doi:10.18653/v1/2025.naacl-long.449.

[5] N. Jegham, M. Abdelatti, C. Y. Koh, L. Elmoubarki, A. Hendawi, How hungry is ai? benchmarking energy, water, and carbon footprint of llm inference, 2025. URL: https://arxiv.org/abs/2505.09598. arXiv:2505.09598.

[6] F. Wang, Z. Zhang, X. Zhang, Z. Wu, T. Mo, Q. Lu, Wang, A comprehensive survey of small language models in the era of large language models: Techniques, enhancements, applications, collaboration with llms, and trustworthiness, ACM Transactions on Intelligent Systems and Technology 16 (2025) 1–87.

[7] W. Chen, Z. Li, M. Ma, Octopus: On-device language model for function calling of software apis, Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies 3 (2025) 329–339.

[8] T. Gan, Q. Sun, RAG-MCP: Mitigating Prompt Bloat in LLM Tool Selection via Retrieval-Augmented Generation, 2025. URL: https://arxiv.org/abs/2505.03275, _eprint: 2505.03275.

[9] V. Carbune, N. Wichers, O. Riva, Enabling privacy-compliant and context-aware tool use by artificial intelligence agents, https://www.tdcommons.org/dpubs_series/8272, 2025. Technical Disclosure Commons, Defensive Publications Series.

[10] A. Hogan, E. Blomqvist, M. Cochez, C. D'amato, G. D. Melo, C. Gutierrez, S. Kirrane, J. E. L. Gayo, R. Navigli, S. Neumaier, A.-C. N. Ngomo, A. Polleres, S. M. Rashid, A. Rula, L. Schmelzeisen, J. Sequeda, S. Staab, A. Zimmermann, Knowledge graphs, ACM Computing Surveys 54 (2021) 1–37. URL: http://dx.doi.org/10.1145/3447772. doi:10.1145/3447772.

[11] H. Cui, R. Trimananda, A. Markopoulou, S. Jordan, PoliGraph: Automated privacy policy analysis using knowledge graphs, in: 32nd USENIX Security Symposium (USENIX Security 23), USENIX Association, Anaheim, CA, 2023, pp. 1037–1054. URL: https://www.usenix.org/conference/usenixsecurity23/presentation/cui.

[12] H. Liu, N. Zhang, Z. Su, A knowledge graph framework for purpose-driven privacy policy

compliance auditing, in: X. Tang, H. Xia, V. N. Huynh, X. Li (Eds.), Knowledge and Systems Sciences, Springer Nature Singapore, Singapore, 2026, pp. 69–83.
[13] W. Baldwin, S. Ghanavati, Knowledge graph representations for llm-based policy compliance reasoning, 2026. URL: https://arxiv.org/abs/2604.27713. arXiv:2604.27713.
[14] Y. Qin, S. Hu, Y. Lin, W. Chen, N. Ding, G. Cui, Sun, Tool learning with foundation models, ACM Computing Surveys 57 (2024) 1–40.
[15] Y. Qin, S. Liang, Y. Ye, K. Zhu, L. Yan, Y. Lu, Y. Lin, X. Cong, X. Tang, B. Qian, S. Zhao, L. Hong, R. Tian, R. Xie, J. Zhou, M. Gerstein, d. li, Z. Liu, M. Sun, Toolllm: Facilitating large language models to master 16000+ real-world apis, in: B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, Y. Sun (Eds.), International Conference on Learning Representations, volume 2024, 2024, pp. 9695–9717. URL: https://proceedings.iclr.cc/paper_files/paper/2024/file/28e50ee5b72e90b50e7196fde8ea260e-Paper-Conference.pdf.
[16] Z. Shi, Y. Wang, L. Yan, P. Ren, S. Wang, D. Yin, Z. Ren, Retrieval models aren't tool-savvy: Benchmarking tool retrieval for large language models, in: W. Che, J. Nabende, E. Shutova, M. T. Pilehvar (Eds.), Findings of the Association for Computational Linguistics: ACL 2025, Association for Computational Linguistics, Vienna, Austria, 2025, pp. 24497–24524. URL: https://aclanthology.org/2025.findings-acl.1258/. doi:10.18653/v1/2025.findings-acl.1258.
[17] Anthropic, Model context protocol specification, https://modelcontextprotocol.io, 2024. Accessed: 2026-07-25.
[18] S. G. Patil, T. Zhang, X. Wang, J. E. Gonzalez, Gorilla: Large language model connected with massive apis, Advances in Neural Information Processing Systems 37 (2024) 126544–126565.
[19] D. Sanmartin, Kg-rag: Bridging the gap between knowledge and creativity, 2024. URL: https://arxiv.org/abs/2405.12035. arXiv:2405.12035.
[20] A. Maharjan, U. Yadav, Chunking, Retrieval, and Re-ranking: An Empirical Evaluation of RAG Architectures for Policy Document Question Answering, ArXiv abs/2601.15457 (2026). URL: https://api.semanticscholar.org/CorpusID:284934963.
[21] M. R. Kabir, R. M. Sultan, F. Rahman, M. R. Amin, S. Momen, N. Mohammed, S. Rahman, LegalRAG: A Hybrid RAG System for Multilingual Legal Information Retrieval, 2025 International Joint Conference on Neural Networks (IJCNN) (2025) 1–8. URL: https://api.semanticscholar.org/CorpusID:278000677.
[22] Y. Ning, M. Xu, J. Wen, Q. Pi, Y. Zhu, M. Zhong, Qian, Privacy-protected retrieval-augmented generation for knowledge graph question answering, Proceedings of the AAAI Conference on Artificial Intelligence 40 (2026) 32573–32581.
[23] A. Shankar, A. Waldis, C. Bless, M. Andueza Rodriguez, L. Mazzola, Privacyglue: A benchmark dataset for general language understanding in privacy policies, Applied Sciences 13 (2023). URL: https://www.mdpi.com/2076-3417/13/6/3701. doi:10.3390/app13063701.
[24] N. Guha, J. Nyarko, D. Ho, C. Ré, A. Chilton, A. Chohlas-Wood, A. Peters, Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models, Advances in neural information processing systems 36 (2023) 44123–44279.
[25] S. Wilson, F. Schaub, A. Balayn, D. Shah, L. DeCathuelaz, C. J. Cowdell, L. Schafer, N. Wang, N. Godbole, S. Salem, M. Sokol, N. Sadeh, N. Sadeh, The creation and analysis of a website privacy policy corpus, in: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016, pp. 1330–1340.
[26] L. Garza, L. Elluri, A. Kotal, A. Piplai, D. Gupta, A. Joshi, Privcomp-kg : Leveraging knowledge graph and large language models for privacy policy compliance verification, 2024. URL: https://arxiv.org/abs/2404.19744. arXiv:2404.19744.
[27] S. Mukhopadhyay, S. Reddy, S. Muthukumar, J. An, P. Kumaraguru, Privacybench: A conversational benchmark for evaluating privacy in personalized ai, 2025. URL: https://arxiv.org/abs/2512.24848. arXiv:2512.24848.
[28] I. Granite Team, Ibm granite 4.0 language models, https://github.com/ibm-granite/granite-4.0-language-models, 2025.
[29] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, C. Zheng, D. Liu,

F. Zhou, F. Huang, F. Hu, H. Ge, H. Wei, H. Lin, J. Tang, J. Yang, J. Tu, J. Zhang, J. Yang, J. Yang, J. Zhou, J. Zhou, J. Lin, K. Dang, K. Bao, K. Yang, L. Yu, L. Deng, M. Li, M. Xue, M. Li, P. Zhang, P. Wang, Q. Zhu, R. Men, R. Gao, S. Liu, S. Luo, T. Li, T. Tang, W. Yin, X. Ren, X. Wang, X. Zhang, X. Ren, Y. Fan, Y. Su, Y. Zhang, Y. Zhang, Y. Wan, Y. Liu, Z. Wang, Z. Cui, Z. Zhang, Z. Zhou, Z. Qiu, Qwen3 technical report, 2025. URL: https://arxiv.org/abs/2505.09388. arXiv:2505.09388.

A. Preference ID to User Privacy Preference Text Mapping

Table 4
Preference Text to Preference ID for Experiments 2 and 3

Preference ID	User Preference Text
P1	The API must not allow first party collection or use.
P2	The API must not allow third party sharing or collection.
P3	The API must delete or anonymize the user's data after a specified amount of time.
P4	The API must not allow first party collection or use. The API must not allow third party sharing or collection.
P5	The API must not allow first party collection or use. The API must delete or anonymize the user's data after a specified amount of time.
P6	The API must delete or anonymize the user's data after a specified amount of time. The API must not allow third party sharing or collection.
P7	The API must not allow first party collection or use. The API must not allow third party sharing or collection. The API must delete or anonymize the user's data after a specified amount of time.

B. Policy Segment Examples for User Preference P4

Compliant policy segment

OUR EMAIL POLICY SIDEARM Sports complies with the federal CAN-SPAM Act of 2003. We will not send you marketing messages if you have opted out. We will not share, sell, rent, swap or authorize any third party to use your email address for commercial purposes without your permission. If you feel you have received an email from us in error, please contact us at sidearm@sidearmsports.com.

Explicitly non-compliant segment failing first-party collection and use

ng> Submitting a Resume > If you are interested in applying for a job with Caribou Coffee, we collect the information asked for in the online application, and we use that information solely for the purpose of matching your experience and preferences with the job openings we have, and someone from Caribou Coffee may contact you to ask for your resume and to discuss potential employment opportunities. The information you submit is used solely for the purposes of recruiting and is not shared with any third party.

Explicitly non-compliant segment failing third-party collection and use

About Our Web Surveys r> The Library uses various types of online forms and surveys to collect opinions and feedback from a random sample of visitors. Primarily, the Library uses the ForeSee Results' American Customer Satisfaction Index (ACSI) online survey on an ongoing basis to obtain feedback and data on visitors' satisfaction with various Library of Congress web properties. The ACSI survey invitation pops up for a random sample of visitors; it is voluntary and does not collect personal information. The Library may also use other third party forms and surveys allowing you to provide feedback. All information you submit is voluntary.

Implicitly non-compliant segment

When we send you such information by email, you will always have an immediate opportunity to decline further communication. Please use the contact specified in the email so that we can remove you from the list that contains your name. If you are uncertain about whom to contact, you may contact our webmaster at webmaster@si.edu.

C. Prompt Templates Used

C.1. Tool Bloat Prompt for 4.1

```
"""\
Objective:
Given tool descriptions and a user query choose the correct tool which satisfies the query.
If all tools fail the constraints, return nothing.
Otherwise, return the api name, tool name, and reason.

Tool descriptions:
{{tools}}

User Query:
{{query_str}}
```

```
Constraints:
1. The tool's use must align with the user query or it is rejected.

Chosen Tool: \
"""
```

C.2. Policy Compliance Prompt for 4.2

```
"""\
Objective:
Given a user's privacy policy and a policy text, answer if the
policy text respects the user's privacy policy.

User Privacy Policy:
{{user_policy}}

Policy:
{{policy_text}}

Constraints:
1. You must answer "yes" or "no" and provide a reason.

Answer: \
"""
```

C.3. Tool Policy-User Policy Compliance Prompt for 4.3

```
"""\
Objective:
Given tool descriptions with policies, a user query, and a user privacy policy, choose the corre
If all tools fail the constraints, return nothing.
Otherwise, return the complete tool name, complete tool api name, complete tool policy, and reas

Tool descriptions and policies:
{{tools_and_policies}}

User Query:
{{query_str}}

User Privacy Policy:
{{user_policy}}

Constraints:
1. The tool's privacy policy must align with the user's or it is rejected.
2. The tool's use must align with the user query or it is rejected.

Chosen Tool With Policy: \
"""
```