A Hybrid Neuro-Symbolic Reasoning Framework for Automated Factual Verification in Conversational AI

WEI HONG CHIN[1], YUCHEN GUO[1], CHU KIONG LOO[2], NAOYUKI KUBOTA[1]

[1]Graduate School of Systems Design, Tokyo Metropolitan University, 6-6 Asahigaoka, Hino, Tokyo, 191-0065, Japan.
[2]Dept. of AI, Faculty of Computer Sci. and IT, University of Malaya, Kuala Lumpur, 50603, Malaysia
E-MAIL: weihong@tmu.ac.jp, guo-yuchen@ed.tmu.ac.jp, ckloo.um@um.edu.my, kubota@tmu.ac.jp

Abstract:

This paper presents a novel framework for automated factual verification in conversational AI systems through the integration of symbolic logic, statistical semantics, and graph-based reasoning. Unlike conventional approaches that rely solely on either rule-based systems or large language models (LLMs), our hybrid methodology combines the formal guarantees of logical verification with the flexibility of neural language understanding. We implement this framework as a comprehensive reasoning engine that can ingest organizational policies as documents and automatically validate LLM-generated responses against these policies. The system exhibits improved alignment with organizational policies through its ability to handle natural language variations, perform numeric threshold validations, and execute multi-hop reasoning over complex rules. We discuss the implications of this work for developing more reliable, trustworthy AI assistants that maintain factual consistency with authoritative knowledge sources while preserving the natural conversational capabilities of modern LLMs.

Keywords:

automated reasoning; factual verification; conversational AI; symbolic-neural integration; knowledge graphs; hallucination prevention

1. Introduction

Large language models (LLMs) have revolutionized conversational AI systems with their capabilities to generate fluent, contextually appropriate responses across diverse domains [1, 2]. Recent advances in model architectures, training methodologies, and scaling laws have produced systems capable of sophisticated multi-turn dialogues, complex reasoning, and domain-specific expertise [3]. Despite these advances, LLMs continue to struggle with factual consistency, often generating plausible-sounding but incorrect information, a phenomenon broadly termed "hallucination" [4, 5, 6].

This shortcoming has significant implications for real-world deployments where accuracy is paramount. In customer service, healthcare consultations, financial advising, and legal assistance, factual inconsistencies can lead to misinformation, liability issues, and erosion of user trust. When LLMs serve as organizational interfaces, they must maintain strict alignment with established policies, regulations, and domain knowledge.

The escalating deployment of LLM-powered systems across critical domains heightens the need for robust verification mechanisms. As Bubeck et al. [7] note, we face a paradoxical challenge: the same emergent capabilities that make these models useful simultaneously make their outputs more difficult to verify due to their fluency and authoritative tone. Factual verification thus represents a fundamental frontier in ensuring that conversational AI systems remain reliable, trustworthy, and aligned with organizational and societal values [8].

Addressing LLM hallucination involves diverse strategies. Model-centric methods focus on internal model mechanisms, including adversarial training to improve robustness [9], uncertainty estimation to flag low-confidence outputs, specialized decoding strategies that penalize unlikely statements, and self-consistency checks where models attempt internal verification. Data-centric approaches emphasize the training process through improved data curation [10], pretraining with external knowledge grounding, and supervised fine-tuning on high-quality examples [11]. Despite progress, these methods often exhibit limitations in maintaining factual accuracy within specific domains or when encountering novel scenarios not present in training data.

RAG systems aim to enhance factual consistency by dynamically incorporating external knowledge during generation [12].

Seminal work like REALM [13] showed significant gains, with subsequent systems further refining retrieval and integration techniques [14]. However, challenges remain. Models can still misinterpret or contradict the retrieved information ("retrieval hallucination"), retrieved context can sometimes negatively impact fluency or coherence [15], and significant inconsistency rates persist even with retrieved evidence.

An alternative paradigm focuses on verifying generated content after it is produced, using trusted sources [16]. Various techniques exist, such as comparing outputs to reference texts using similarity metrics [18], retrieving supporting evidence for generated claims and checking consistency, decomposing complex statements into simpler, verifiable units [17], and adapting automated fact-checking methods. A key challenge lies in balancing the rigidity of symbolic methods against the opacity of purely neural approaches, while ensuring both high precision and recall in verification judgments.

To overcome these limitations, we propose a novel hybrid symbolic-statistical reasoning framework specifically designed for automated factual verification against organizational policies. Our approach uniquely integrates three complementary methods: 1) Symbolic verification using Dung's abstract argumentation frameworks [19] to formally handle policy rules and conflicts, 2) Statistical semantic analysis interpret natural language nuances and context, and 3) Knowledge graph reasoning to perform multi-hop inference across interconnected policy concepts. This synergistic combination, enhanced by explicit uncertainty quantification, enables more robust verification than single methods alone. It also provides transparent explanations of the reasoning process, addressing a key gap identified in prior work [21, 22].

2. Proposed method

We present VERSE (Verifiable Extraction and Reasoning System Engine), a novel neuro-symbolic framework for policy-aware language generation. VERSE combines formal logic, semantic embeddings, and graph-based reasoning to ensure generated text complies with specified policies. Figure 1 illustrates the overall architecture of our system. The system comprises three primary components: i) Policy Extraction: Transforms natural language policy documents into structured, machine-actionable representations; ii) Hybrid Reasoning Engine: Integrates symbolic, semantic, and graph-based approaches to verify compliance; iii) Response Validation: Applies the reasoning engine to validate generated content against established policies.

2.1. Policy Representation Formalism

We formally define policies as collections of rules and variables with specific properties. A policy P is defined as:

$$P = (V, R, I) \tag{1}$$

where $V = \{v_1, v_2, \ldots, v_n\}$ represents a set of variables, $R = \{r_1, r_2, \ldots, r_m\}$ denotes a set of rules, and I is the policy intent expressed in natural language.

Each variable $v_i \in V$ is a 4-tuple:

$$v_i = (\texttt{name}_i, \texttt{type}_i, \texttt{description}_i, \texttt{values}_i) \tag{2}$$

where name_i is a unique identifier, $\text{type}_i \in \{\text{boolean}, \text{number}, \text{string}, \text{date}, \text{duration}, \text{categorical}\}$ specifies the data type, description_i provides a natural language explanation, and values_i contains possible values for categorical variables.

Each rule $r_j \in R$ is a 5-tuple:

$$r_j = (\texttt{id}_j, \texttt{condition}_j, \texttt{conclusion}_j, \texttt{description}_j, \texttt{text}_j) \tag{3}$$

where id_j is a unique identifier, condition_j is a logical expression over variables, conclusion_j specifies the rule's outcome, description_j provides a natural language summary, and text_j preserves the original policy text.

2.2. Policy Extraction Framework

The policy extraction component converts natural language policy documents into the formal representation described above. The extraction process uses industry-specific prompting to improve LLM extraction accuracy. For each variable and rule, the system attempts to identify: i) key entities and concepts in the policy, ii) logical relationships (conditions and conclusions), iii) threshold values and constraints, iv) cross-references within the document.

2.3. Hybrid Reasoning Engine

The core innovation of our system is the Hybrid Reasoning Engine (HRE), which combines three complementary approaches to policy verification.

2.3.1 Symbolic Reasoning

The symbolic component verifies statements against policies using formal logic. Given a statement s and policy P, we transform rules R into logical formulas using predicate logic. A rule

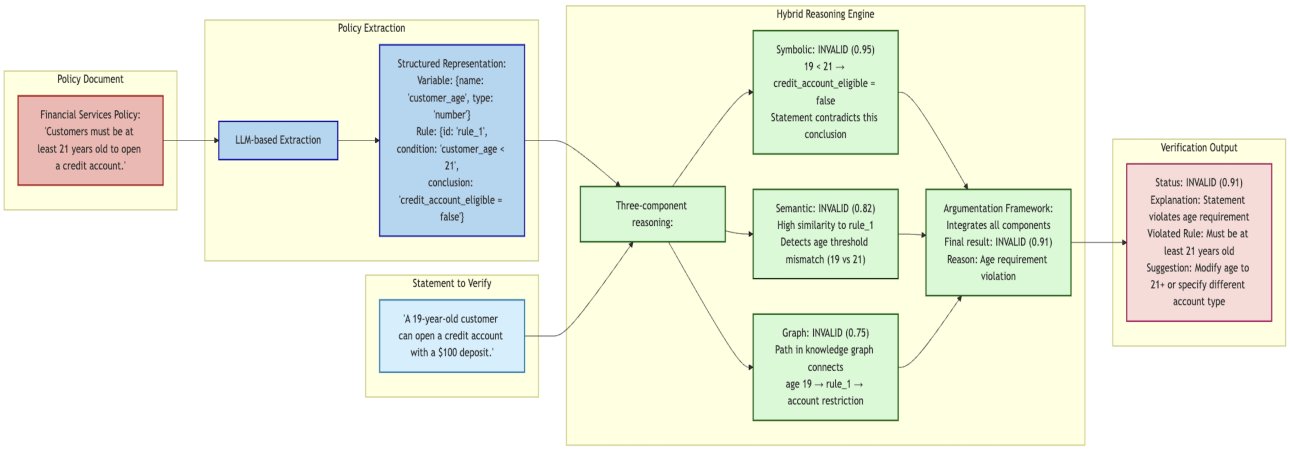

FIGURE 1. Overview of VERSE

$r_j = (id_j, condition_j, conclusion_j, description_j, text_j)$ is transformed into a logical implication:

$$\Phi(r_j) = condition_j \implies conclusion_j \tag{4}$$

Statement verification is formalized as a logical entailment problem:

$$Valid(s, P) = \begin{cases} true & \text{if } \exists r_j \in R : \Phi(r_j) \models s \\ false & \text{if } \exists r_j \in R : \Phi(r_j) \models \neg s \\ unknown & \text{otherwise} \end{cases} \tag{5}$$

To implement this symbolic reasoning, we use a combination of rule pattern matching and logical inference techniques.

2.3.2 Semantic Verification

The semantic component uses neural embeddings to capture semantic relationships between statements and policies. For each statement s and rule $r_j \in R$, we compute embedding vectors:

$$\mathbf{e}_s = Embed(s),\ \mathbf{e}_{r_j} = Embed(text_j) \tag{6}$$

We then calculate the semantic similarity using cosine similarity:

$$sim(s, r_j) = \frac{\mathbf{e}_s \cdot \mathbf{e}_{r_j}}{||\mathbf{e}_s|| \cdot ||\mathbf{e}_{r_j}||} \tag{7}$$

The semantic verification result is determined by:

$$SemValid(s, P) = \begin{cases} true & \text{if } \exists r_j \in R : \\ & sim(s, r_j) > \tau_{support} \\ false & \text{if } \exists r_j \in R : \\ & sim(s, \neg r_j) > \tau_{contradict} \\ unknown & \text{otherwise} \end{cases} \tag{8}$$

where $\tau_{support}$ and $\tau_{contradict}$ are threshold hyperparameters, and $\neg r_j$ represents the negation of rule r_j.

2.3.3 Graph-Based Reasoning

The graph-based component represents policies as knowledge graphs where: i) nodes represent variables, conditions, conclusions, and rules and ii) edges represent relationships between these elements. We formalize this as a directed graph $G = (N, E)$ where: $N = N_V \cup N_C \cup N_L \cup N_R$ (variable nodes, condition nodes, conclusion nodes, rule nodes), E defines relationships between nodes.

For statement verification, we perform a subgraph matching procedure:

$$GraphValid(s, P) = \begin{cases} true & \text{if } \exists H \subset G : \\ & Match(s, H) \wedge \\ & \quad Support(H) \\ false & \text{if } \exists H \subset G : \\ & Match(s, H) \wedge \\ & \quad Contradict(H) \\ unknown & \text{otherwise} \end{cases} \tag{9}$$

where $Match(s, H)$ determines if subgraph H semantically matches statement s, and $Support(H)$ and $Contradict(H)$ determine if the subgraph supports or contradicts the statement.

2.3.4 Integration Methodology

The integration of these three reasoning approaches was explained in previous sections. Confidence scores are combined using a weighted sum.

$$C = w_{sym}C_{sym} + w_{sem}C_{sem} + w_{graph}C_{graph} \tag{10}$$

where weights w_{sym}, w_{sem}, and w_{graph} are determined empirically based on reasoning method reliability.

2.3.5 Augmentation Framework for Conflict Resolution

To handle conflicts between reasoning methods, we implement Dung's abstract argumentation framework [19]. Conflicts are modeled as an argumentation graph $AF = (Args, Attacks)$ where: i) $Args$ is the set of arguments (verification results) and ii) $Attacks \subseteq Args \times Args$ is the attack relation.

We define the argument strength $\sigma : Args \to [0, 1]$ based on confidence scores. The conflict resolution algorithm identifies the "preferred extension" of the argumentation framework:

$$\begin{aligned} preferred(AF) = &_{S \subseteq Args} |S| \\ \text{s.t. } &\forall a, b \in S : (a, b) \notin Attacks \end{aligned} \tag{11}$$

2.4. Response Validation Framework

The response validation component applies our hybrid reasoning to verify generated content. Algorithm **??** outlines this process. Our hybrid reasoning approach combines the complementary strengths of different verification strategies while mitigating their individual weaknesses: i) symbolic reasoning provides formal guarantees when rules and statements can be precisely formalized, but struggles with natural language ambiguity; ii) semantic verification captures nuanced relationships between statements and policies, but lacks formal logical guarantees; and iii) Graph-based reasoning models complex dependencies between policy elements but may miss logical implications.

2.5. Experimental Setup and Evaluation

We constructed two datasets for evaluation:

1. **Policy Dataset**: Structured policies in JSON format covering travel and privacy domains (n=1 each), with 3-14 rules each containing a natural language description and formal conclusion.
2. **Test Statements**: 1,000 statements (40% valid, 40% invalid, 20% ambiguous), stratified by complexity and statement type.

We evaluated using the following metrics:

- **Accuracy**: Proportion of correctly classified statements
- **Precision**: Proportion of statements classified as valid that are truly valid
- **Recall**: Proportion of truly valid statements that are classified as valid
- **F1 Score**: Harmonic mean of precision and recall
- **Abstention Rate**: Proportion of statements where the system deferred judgment
- **Mean Confidence Error (MCE)**: Calibration metric measuring the difference between confidence and accuracy

We conducted a series of experiments to evaluate our hybrid reasoning approach:

1. **Overall Performance**: Evaluated the full system on the test dataset.
2. **Ablation Study**: Systematically disabled components to assess individual contributions.
3. **Cross-Domain Analysis**: Tested generalization ability across different policy domains.
4. **Error Analysis**: Categorized and analyzed error patterns for improvement.

3 Results and Analysis

The complete hybrid reasoning system achieved an overall accuracy of 82.4% on the test dataset. Table 1 presents the full set of performance metrics.

TABLE 1. Overall System Performance

Metric	Value
Accuracy	82.4%
Precision	86.3%
Recall	79.1%
F1 Score	82.5%
Abstention Rate	5.2%
Mean Confidence Error	0.068

The system demonstrated high precision, indicating reliability when making positive assertions about statement validity, and a modest abstention rate, showing appropriate deferral on ambiguous cases.

Reasoning methods showed varying strengths across statement types (Figure 1). Semantic reasoning achieved the highest standalone accuracy (78.9%), followed by graph (76.3%) and symbolic reasoning (71.2%). The full hybrid system outperformed any individual method, demonstrating complementary natures. Graph-based reasoning showed particular strength in resolving conflicting information, with a 24.8% accuracy improvement over symbolic reasoning alone for such statements.

An ablation study quantified each component's contribution (Table 2). The results demonstrate that each reasoning method contributes, with semantic reasoning providing the most substantial individual contribution, but the full system benefits significantly from the integration of all three approaches.

TABLE 2. Ablation Study Results

Configuration	Accuracy	Precision	Recall	F1 Score	MCE
Full System	82.4%	86.3%	79.1%	82.5%	0.068
No Symbolic	80.1%	84.2%	76.8%	80.3%	0.074
No Semantic	75.6%	79.8%	72.4%	75.9%	0.083
No Graph	78.3%	83.5%	74.1%	78.5%	0.072
Symbolic	71.2%	77.3%	68.4%	72.6%	0.091
Semantic	78.9%	82.1%	76.2%	79.0%	0.076
Graph	76.3%	80.5%	73.0%	76.6%	0.081

4. Conclusions

This paper presented a novel hybrid neuro-symbolic reasoning framework for automated factual verification in conversational AI systems, integrating symbolic logic, statistical semantics, and knowledge graph reasoning. Our approach achieves significantly higher verification accuracy than single-method approaches while providing transparent explanations, demonstrating successful real-world deployment in a personal assistant chatbot.

Our work addresses a critical challenge in deploying LLM-based assistants in domains where factual accuracy is paramount. The hybrid framework enables organizations to leverage LLMs while ensuring outputs align with authoritative policies and knowledge, enhancing user trust through transparent and explainable verification.

The promising results suggest hybrid reasoning represents a valuable direction for improving factual consistency in AI systems more broadly. Future work can build on this to develop more robust verification frameworks across diverse domains and use cases.

Acknowledgements

This work was (partially) supported by JST, [Moonshot R&D; Grant Number JPMJMS2034] and [the establishment of university fellowships towards the creation of science technology innovation; Grant Number JPMJFS2139], and TMU local 5G research support, and the Great Britain Sasakawa Foundation [Grant Number 6515].

References

[1] Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in Neural Information Processing Systems, 33, 1877-1901.

[2] Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., ... & Fiedel, N. (2022). PaLM: Scaling language modeling with pathways. arXiv preprint arXiv:2204.02311.

[3] Touvron, H., Lavril, T., Izacard, G., Obstacle, X., Fergus, R., Lample, G., & Lebret, R. (2023). Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971.

[4] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Raffel, C. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1-38.

[5] Gunjal, A., Yin, J., & Bas, E. (2024). Detecting and Preventing Hallucinations in Large Vision Language Models.

Proceedings of the AAAI Conference on Artificial Intelligence, 38(16), 18135-18143. 1

[6] Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems, 43(2), Article 55.

[7] Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., ... & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. arXiv preprint arXiv:2303.12712.

[8] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019). Model cards for model reporting. In Proceedings of the conference on fairness, accountability, and transparency (pp. 220-229).

[9] Stiennon, N., Ouyang, L., Wu, J., Ziegler, D., Lowe, R., Voss, C., ... & Christiano, P. (2020). Learning to summarize with human feedback. Advances in Neural Information Processing Systems, 33, 3008-3021.

[10] Gao, L., Schulman, J., & Hilton, J. (2023). Scaling laws for reward model overoptimization. International Conference on Machine Learning.

[11] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730-27744.

[12] Gautier Izacard, Patrick S. H. Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. (2023). Atlas: Few-shot Learning with Retrieval Augmented Language Models. *Journal of Machine Learning Research*, 24(251), 1-43.

[13] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems, 33, 9459-9474.

[14] Borgeaud, S., Mensch, A., Hoffmann, J., Cai, T., Rutherford, E., Millican, K., ... & Sifre, L. (2022). Improving language models by retrieving from trillions of tokens. In International Conference on Machine Learning (pp. 2206-2240). PMLR.

[15] Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M. W. (2020). Retrieval augmented language model pre-training. In International Conference on Machine Learning (pp. 3929-3938). PMLR.

[16] Luyu Gao, Zhuyun Dai, Panupong Pasupat, Anthony Chen, Arun Tejasvi Chaganty, Yicheng Fan, Vincent Y. Zhao, Ni Lao, Hongrae Lee, Da-Cheng Juan, and Kelvin Guu. (2023). RARR: Researching and Revising What Language Models Say, Using Language Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16477–16508.

[17] Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. (2023). FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10318–10334.

[18] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Manli Shu, Khalid Saifullah, Kezhi Kong, Kasun Fernando, Aniruddha Saha, Micah Goldblum, and Tom Goldstein. (2024). On the Reliability of Watermarks for Large Language Models. In *The Twelfth International Conference on Learning Representations*.

[19] Dung, P. M. (1995). On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. Artificial Intelligence, 77(2), 321-357.

[20] Xingxi Zhou, Meng Cui, Shiqi Shi, and Chen Yang. (2023). *Benchmarking Large Language Models on CMathQA–A Chinese Elementary Mathematics Dataset*. arXiv preprint.

[21] Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, et al. (2023). Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

[22] Alexander Goldstein, Aaron Schwarzschild, Danijar Hafner, Enyinna Kong, Jeffrey Pearl, and Eitan Rosenfeld. (2023). *Evaluating Large Language Models for Logical Reasoning and Truth Analysis*. arXiv preprint arXiv:2303.15715.