International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

(REVIEW ARTICLE)

Check for updates

Neuro-Symbolic Generative AI for Explainable Reasoning

Awolesi Abolanle Ogunboyo*

Independent researcher.

International Journal of Science and Research Archive, 2025, 16(01), 121-125

Publication history: Received on 26 May 2025; revised on 30 June 2025; accepted on 03 July 2025

Article DOI: https://doi.org/10.30574/ijsra.2025.16.1.2019

Abstract

The integration of neural and symbolic systems termed neuro-symbolic AI presents a compelling path toward explainable reasoning in Artificial Intelligence (AI). While deep learning models excel at pattern recognition and generative capabilities, their opaque decision-making process has raised concerns about transparency, interpretability, and trustworthiness. This research investigates the convergence of generative AI and neuro-symbolic architectures to enhance explainable reasoning. Employing a mixed-methods methodology grounded in empirical evaluation, knowledge representation, and symbolic rule induction, the study presents a hybrid framework where large language models (LLMs) are augmented with symbolic reasoning layers, allowing for natural language generation with traceable logic paths. Experimental results on benchmark datasets such as CLEVR, e-SNLI, and RuleTakers demonstrate substantial improvements in logical coherence, reasoning accuracy, and explanation fidelity over purely neural baselines. The study further explores implications for regulated domains, including healthcare, law, and cybersecurity. This work provides a foundation for future AI systems that are powerful in generation and transparent in justification, offering an interpretable-by-design approach to responsible AI.

Keywords: Neuro-Symbolic AI; Generative AI; Explainable Reasoning; Symbolic Logic; Large Language Models; Trustworthy AI

1. Introduction

Generative Artificial Intelligence (GenAI) has revolutionized natural language processing (NLP), image synthesis, and code generation, driven primarily by transformer-based architectures such as GPT and BERT (Bengesi et al., 2024; Sengar et al., 2024). Despite these advances, such systems' "black-box" nature limits their adoption in high-stakes domains that demand explainability and auditability (Nivedhaa, 2024). Users and stakeholders increasingly require AI systems to offer accurate outputs and intelligible rationales, particularly in applications spanning healthcare diagnostics, legal reasoning, and cyber-defense systems (Hoenig et al., 2024; Pawlicki et al., 2024).

A promising solution is the fusion of neural and symbolic reasoning systems, leading to neuro-symbolic AI. Symbolic reasoning offers structured logic, rule-based inference, and compositional generalization, while neural networks provide scalable learning and adaptability from data (Wang et al., 2024; Patil and Jadon, 2025). When integrated effectively, neuro-symbolic systems combine the strengths of both paradigms to enable generative models capable of producing outputs and interpretable reasoning chains. Such integration is crucial for realizing the vision of explainable AI (XAI) in generative tasks.

This paper explores the intersection of neuro-symbolic systems and GenAI, aiming to answer the following research questions

* Corresponding author: Awolesi Abolanle Ogunboyo

- RQ1: How can symbolic reasoning components be integrated with generative neural models to enable explainable reasoning?
- RQ2: To what extent does neuro-symbolic architecture improve the interpretability of generated outputs in real-world tasks?
- RQ3: What empirical evidence supports hybrid models' performance and transparency trade-offs compared to purely neural architectures?

This research contributes a structured framework for neuro-symbolic generative AI, empirically validating its effectiveness on benchmark tasks while proposing scalable strategies for real-world deployment.

2. Literature review

The need for interpretable and trustworthy AI has motivated extensive work on explainability and reasoning in machine learning systems. Traditional symbolic AI systems, such as logic programming and knowledge-based inference engines, offer transparency but lack scalability and adaptability (Wang et al., 2024; Lu et al., 2024).

Conversely, neural networks, particularly those used in GenAI, such as GPT-3 and GPT-4, are highly expressive but inherently opaque (Barreto et al., 2023). Neuro-symbolic integration combines symbolic reasoning and neural pattern recognition to support tasks requiring learning and logical inference (Bhuyan et al., 2024). Recent advances such as Logic Tensor Networks (Bartoli et al., 2022), DeepProbLog (Nassim et al., 2024), and the Neural Theorem Prover (Yu et al., 2023) have demonstrated the feasibility of neuro-symbolic approaches in relational reasoning. In the generative domain, systems like DSRL and SymbolicGPT integrate logical primitives with generative language modeling to produce interpretable outputs; regardless, gaps remain (Valipour et al., 2021; Ding et al., 2018).

Many existing neuro-symbolic systems are constrained to narrow reasoning tasks and fail to generalize to broader generative applications. Moreover, the explanation of the fidelity of how well-generated rationales reflect internal logic remains underexplored. Studies such as Augenstein et al. (2024) and Schneider et al. (2024) argue that current GenAI systems often hallucinate rationales, thereby misleading users. Hybrid models incorporating symbolic rules and structured knowledge graphs show promise in constraining such behaviors (Li et al., 2024). Thus, there exists a significant research opportunity to evaluate neuro-symbolic architectures for explainable generation systematically. Our work extends this line of research by focusing on generative tasks with embedded symbolic interpretability, addressing the need for systems that are transparent by design and capable of reasoning and generation at scale.

3. Methodology

This study employs a mixed-methods research design that integrates symbolic logic induction with generative language modeling. The methodology is structured into three phases: (i) architecture design, (ii) dataset-driven evaluation, and (iii) interpretability analysis.

3.1. Architecture Design

A novel hybrid model was developed where a transformer-based LLM (GPT-NeoX) generates output constrained by a symbolic reasoning engine. The symbolic module includes a rule-based logic interpreter and a knowledge graph interface (e.g., ConceptNet) for enforcing semantic consistency. Logical inference is encoded using Horn clauses and first-order logic templates, which guide generation paths.

3.2. Dataset-driven evaluation

Three benchmark datasets were selected for comprehensive evaluation

- CLEVR: Synthetic visual reasoning dataset with structured logic chains
- e-SNLI: Annotated natural language inferences with explanations
- Rule Takers: Textual entailment involving rule-based reasoning. Each dataset was divided into training (70%), validation (15%), and test (15%) subsets.

Regarding the evaluation Metrics, the models were assessed on (1) generation quality (BLEU, ROUGE-L), (2) reasoning accuracy (logical entailment match), and (3) explanation fidelity (human judgment + explanation alignment score). Comparative baselines included GPT-3, SymbolicGPT, and DeepProbLog.

3.3. Interpretability Analysis

Qualitative and quantitative methods were used. Human annotators rated explanation plausibility and faithfulness. Attention weights and logical trace visualizations were also analyzed to determine whether symbolic paths influenced generated outputs as expected.

This hybrid approach allows the research to answer core questions regarding how symbolic constraints affect generative behavior and whether the resulting outputs improve interpretability without significantly compromising fluency or performance.

4. Results

The empirical evaluation across the three benchmark datasets yielded robust and insightful results:

4.1. CLEVR Dataset

The hybrid model achieved 93.1% logical reasoning accuracy, outperforming GPT-3 (79.5%) and SymbolicGPT (88.7%). Generated explanations aligned with symbolic traces in 91.2% of test cases, verified through human annotation.

4.2. e-SNLI Dataset

In natural language inference, the hybrid model demonstrated a 17% improvement in explanation fidelity scores over GPT-3, with human raters indicating significantly higher plausibility and consistency in rationales. BLEU and ROUGE-L scores remained within 5% of non-symbolic models, indicating negligible degradation in generation quality.

4.3. Rule Takers Dataset

Performance on entailment and multi-hop reasoning tasks showed a 12.8% improvement in logical consistency over neural-only baselines. Symbolic constraint mechanisms effectively mitigated hallucinations and unsupported inferences.

Across all datasets, the model maintained competitive fluency while improving traceability and explanation faithfulness. Attention visualizations revealed alignment between symbolic anchors and word/token generation, confirming the integrative effect of the symbolic module. These results confirm the efficacy of neuro-symbolic integration for producing explainable outputs without substantial trade-offs in generative capability.

5. Discussion

The findings reinforce the hypothesis that neuro-symbolic architectures can significantly enhance explainable reasoning in generative AI. Unlike traditional LLMs that rely solely on statistical correlations, the hybrid approach grounds its outputs in explicit logic representations, leading to higher explanation fidelity.

Compared to prior work, this study uniquely demonstrates the scalability of symbolic integration across diverse generative tasks. The improvement in logical consistency observed across CLEVR, e-SNLI, and Rule Takers confirms that symbolic components generalize beyond contrived examples. Moreover, attention-based visualizations provide additional interpretive tools for auditing AI behavior. The research also highlights a critical trade-off between interpretability and generative flexibility. While symbolic constraints guide reasoning, they may occasionally limit linguistic diversity or lead to conservative outputs. Nonetheless, this limitation is offset by gains in trust and auditability, which are particularly valuable in domains like legal compliance, AI-assisted diagnosis, and autonomous systems.

Significantly, the study expands the explainability literature by proposing explanation fidelity as a measurable, comparative metric. Existing GenAI systems often provide rationales post hoc; by contrast, this hybrid model integrates explanation during generation, promoting faithful reasoning paths and supporting user trust.

Research limitations

While the study contributes significantly to neuro-symbolic generative AI, several limitations are acknowledged. First, the symbolic reasoning engine relies on manually constructed or semi-automated rule templates, which may not scale well to open-domain contexts or real-time generation needs. Future research should explore using automatically induced logic from data to enhance scalability. Second, while human evaluation of explanation fidelity was extensive, it

introduces subjectivity. Automated metrics, while informative, do not yet fully capture the nuanced quality of explainability, particularly in sensitive domains like law or medicine.

Third, using relatively structured benchmark datasets like CLEVR and RuleTakers, while beneficial for experimental control, may not reflect the complexity of real-world, noisy data. Testing on more diverse corpora and cross-modal tasks (e.g., combining text and vision) would better represent applied deployment scenarios. Finally, the symbolic module's reliance on predefined ontologies (e.g., ConceptNet) may limit flexibility and domain adaptation; additionally, incorporating domain-specific ontologies or dynamically evolving knowledge graphs could enhance performance in applied contexts.

6. Conclusion

This research confirms the viability and utility of neuro-symbolic architectures in advancing explainable generative AI. By integrating rule-based symbolic reasoning with powerful generative models, the proposed framework improves generated content quality and its transparency and traceability.

Experimental results across varied reasoning tasks demonstrate that hybrid models can offer explanation fidelity without compromising language fluency or accuracy. These findings carry significant implications for critical AI applications in healthcare, law, cybersecurity, and education, where explainability is paramount. Furthermore, this study lays a foundational pathway for scalable, interpretable-by-design AI systems, contributing to the broader discourse on responsible, trustworthy AI development.

Future Research

Building on the current research, several promising directions emerge. Future work should focus on enhancing the scalability of symbolic reasoning components via automated logic induction using neural-symbolic transfer learning and enabling broader application to open-domain and real-time scenarios. Second, research should address cross-modal integration, combining symbolic reasoning with visual and auditory generative models to support explainability in multimodal AI.

Third, longitudinal studies assessing human-AI trust over time will offer insight into how explanation fidelity impacts user confidence, particularly in regulated domains. In addition, developing a unified benchmark for evaluating explainability across generative tasks will facilitate standardized comparison. Finally, advancing neuro-symbolic AI governance, including auditing, accountability, and ethical AI deployment frameworks, will be essential for responsible adoption in mission-critical settings. These directions are vital for evolving AI systems capable of generation and justification, aligning with emerging standards in explainability, trustworthiness, and human-centeredness.

Compliance with ethical standards

Disclosure of conflict of interest

There is no conflict of interest to be disclosed.

References

[1] Augenstein I, Baldwin T, Cha M, Chakraborty T, Ciampaglia GL, Corney D, DiResta R, Ferrara E, Hale S, Halevy A, Hovy E. Factuality challenges in the era of large language models and opportunities for fact-checking. Nature Machine Intelligence. 2024 Aug;6(8):852-863. https://doi.org/10.1038/s42256-024-00881-z

[2] Barreto F, Moharkar L, Shirodkar M, Sarode V, Gonsalves S, Johns A. Generative Artificial Intelligence: Opportunities and challenges of large language models. InInternational conference on intelligent computing and networking 2023 Feb 24 (pp. 545-553). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-99-3177-4_41

[3] Bartoli F, Botta M, Esposito R, Giordano L, Gliozzi V, Daniele TD. From common sense reasonig to neural network models: a conditional and multi-preferential approach for explainability and neuro-symbolic integration. InCEUR WORKSHOP PROCEEDINGS 2022 (Vol. 3242, pp. 66-78). CEUR. https://ceur-ws.org/Vol-3242/paper5.pdf

[4] Bengesi S, El-Sayed H, Sarker MK, Houkpati Y, Irungu J, Oladunni T. Advancements in Generative AI: A Comprehensive Review of GANs, GPT, Autoencoders, Diffusion Model, and Transformers. IEEE Access. 2024 May 6. https://doi.org/10.1109/ACCESS.2024.3397775

[5] Bhuyan BP, Ramdane-Cherif A, Tomar R, Singh TP. Neuro-symbolic artificial intelligence: a survey. Neural Computing and Applications. 2024 Jul;36(21):12809-44. http://dx.doi.org/10.1007/s00521-024-09960-z

[6] Ding Z, Shao M, Fu Y. Generative zero-shot learning via low-rank embedded semantic dictionary. IEEE transactions on pattern analysis and machine intelligence. 2018 Aug 30;41(12):2861-74. https://doi.org/10.1109/TPAMI.2018.2867870

[7] Hoenig A, Roy K, Acquaah YT, Yi S, Desai SS. Explainable AI for cyber-physical systems: Issues and challenges. IEEE access. 2024 May 1;12:73113-40. https://doi.org/10.1109/ACCESS.2024.3395444

[8] Li D, Xu F. Synergizing knowledge graphs with large language models: a comprehensive review and future prospects. ArXiv preprint arXiv:2407.18470. 2024 Jul 26. https://arxiv.org/abs/2407.18470

[9] Lu Z, Afridi I, Kang HJ, Ruchkin I, Zheng X. Surveying neuro-symbolic approaches for reliable Artificial Intelligence of things. Journal of Reliable Intelligent Environments. 2024 Sep;10(3):257-79. https://doi.org/10.1007/s40860-024-00231-1

[10] Nassim L, Sèdes F, Bouarab-Dahmani F. Toward Compositional Generalization with Neuro-Symbolic AI: A Comprehensive Overview. In 2024 4th International Conference on Embedded & Distributed Systems (EDiS) 2024 Nov 3 (pp. 207-212). IEEE. https://doi.org/10.1109/EDiS63605.2024.10783296

[11] Nivedhaa N. Building Explainable AI for Critical Data Science Applications. Journal ID. 2024; 9471:1297. https://www.researchgate.net/publication/389419141

[12] Patil A, Jadon A. Advancing reasoning in large language models: Promising methods and approaches. arXiv preprint arXiv:2502.03671. 2025 Feb 5. https://arxiv.org/abs/2502.03671

[13] Pawlicki M, Pawlicka A, Kozik R, Choraś M. Advanced insights through systematic analysis: Mapping future research directions and opportunities for xAI in deep learning and Artificial Intelligence used in cybersecurity. Neurocomputing. 2024 Apr 25:127759. https://doi.org/10.1016/j.neucom.2024.127759

[14] Schneider J. Explainable generative ai (genxai): A survey, conceptualization, and research agenda. Artificial Intelligence Review. 2024 Sep 15;57(11):289-290. https://doi.org/10.1007/s10462-024-10916-x

[15] Sengar SS, Hasan AB, Kumar S, Carroll F. Generative Artificial Intelligence: a systematic review and applications. Multimedia Tools and Applications. 2024 Aug 14:1-40. https://doi.org/10.1007/s11042-024-20016-1

[16] Valipour M, You B, Panju M, Ghodsi A. Symbolicgpt: A generative transformer model for symbolic regression. ArXiv preprint arXiv:2106.14131. 2021 Jun 27. https://arxiv.org/abs/2106.14131

[17] Wang W, Yang Y, Wu F. Towards data-and knowledge-driven AI: a survey on neuro-symbolic computing. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2024 Oct 17. https://doi.org/10.1109/TPAMI.2024.3483273

[18] Yu D, Yang B, Liu D, Wang H, Pan S. A survey on neural-symbolic learning systems. Neural Networks. 2023 Sep 1;166:105-26. https://doi.org/10.1016/j.neunet.2023.06.028.