Proceedings of Machine Learning Research vol 284:1–4, 2025 19th Conference on Neurosymbolic Learning and Reasoning

Knowledge Augmented Graph Reasoner (KAGR): A Neuro-Symbolic Approach to Instruction Adherence in Healthcare AI

Ravi Bajracharya* RAVI@DATUM.BIO
Aniwaa Owusu-Obeng ANIWAA@DATUM.BIO
Aris Saoulidis ARIS.SAOULIDIS@DATUM.BIO
Xeno Acharya XENO@DATUM.BIO
Chris Wai Hang Lo CHRIS.LOWH@DATUM.BIO
Arun Bajracharya ARUN@DATUM.BIO
Dhurba Bhandari DHURBA.BHANDARI@DATUM.BIO
Datum.md Inc, San Francisco, CA, USA.

Editors: Leilani H. Gilpin, Eleonora Giunchiglia, Pascal Hitzler, and Emile van Krieken

Abstract

AI in healthcare still faces significant challenges in adhering to clinical guidelines, making things contextual to a given patient and integrating information from myriad sources. There is a growing need for safe, accurate, explainable, and transparent AI systems that are based on validated clinical guidelines and evidence sources. At Datum, we have developed a novel approach to overcoming issues of explainability, reliability, and hallucinations in the use of generative AI for clinical decision making. One of the key problems we are addressing is guideline adherence or instruction adherence in language models, which not only enhances explainability but also improves the overall reliability and safety of the system. In this paper, we describe a novel neuro-symbolic reasoning framework called Knowledge Augmented Graph Reasoner (KAGR) which combines a guideline-based reasoning graph, a benchmarking framework and a domain-specific knowledge graph to enable more accurate and explainable AI recommendations that adhere to guidelines in healthcare.

1. Challenge of Guideline Adherence

Clinical Practice Guidelines (CPGs), such as those published by the National Comprehensive Cancer Network (NCCN) or the American Heart Association (AHA), are pivotal in adoption of evidence-based medicine and critical for ensuring safe and accurate clinical decision making Brouwers et al. (2019). Large language models have demonstrated remarkable capabilities in medical natural language processing tasks, including biomedical question-answering and diagnostic reasoning. However, it is particularly challenging to keep language models up-to-date with current standards and changing/updating evidence bases Li et al. (2025). Adapting a model to change requires significant updates to either the training data and/or fine-tuning data, which also introduces the risk of biases in the model. It can also be challenging if the model needs to adhere to different sets of guidelines owing to specific regional specification and standards. The mainstream approach to tackling the adherence problem is through in-context learning for the models, with retrieval

* Corresponding Author

being the favored approach to finding the right context in a methodology called RAG or Retrieval Augmented Generation Lewis et al. (2020). One major challenge with retrieval-based methods is incomplete and/or irrelevant context for reasoning when it comes to guideline adherence, as a result of which the responses may not be very reliable Xu et al. (2024).

2. A Reasoning Framework for Guideline Adherence

One way to approach the issue of guideline adherence is to reframe the problem as a reasoning problem instead of a retrieval problem. A guideline is best represented as a structured decision tree and can be understood as a framework to reason over input data using various recommendation pathways. At Datum.md Inc, we have come up with a unique neuro-symbolic AI architecture called the Knowledge Augmented Graph Reasoner (KAGR), which combines a reasoning framework built on top of practice guidelines with a domain knowledge graph to better help LLMs with guideline adherence. A KAGR architecture can be thought of as a composition of following three components:

2.1. A Reasoning Graph

The first component of KAGR is a guideline-based executable reasoning graph constructed from the guideline. Each reasoning step in the graph can essentially be a question in natural language with a set of possible answers leading to further questions. Refer to Appendix A for more detail on its construction.

2.2. Benchmarking Framework

The second component of KAGR architecture is a benchmarking framework, which generates test patient data sets to evaluate individual flows extracted from the guideline. The benchmarking framework can essentially assess each flow to determine its current performance statistics, which can help in diagnosing potential issues with the flow.

2.3. Biomedical Knowledge Graph

The third component of the KAGR architecture is the biomedical knowledge graph, which serves as a domain knowledge base. It provides context for the language model to reason over the input data using the reasoning framework steps.

3. Trustworthy AI using KAGR

The reasoning capabilities of language models are improving, including their ability to adhere to instructions. However, the rate of hallucinations is still high, even in the latest revisions of the models, making it necessary to add guardrails and other frameworks to ensure instruction adherence. Even then, they fall short for mission-critical domains like healthcare. The Knowledge Augmented Graph Reasoner architecture addresses this problem by equipping subject matter experts with tools needed to build, benchmark and deploy guideline-adherent LLM frameworks and build highly reliable and trustworthy AI with more than 95% correctness in instruction adherence. Refer to Appendix B for performance evaluation and benchmark statistics.

References

Melissa C. Brouwers, Ivan D. Florez, Sheila A. McNair, Emily T. Vella, and Xioamei Yao. Clinical Practice Guidelines: Tools to support high quality patient care. *Seminars in Nuclear Medicine*, 49(2):145–152, 1 2019. doi: 10.1053/j.semnuclmed.2018.11.001. URL https://doi.org/10.1053/j.semnuclmed.2018.11.001.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-Tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP tasks. *Neural Information Processing Systems*, 33:9459–9474, 5 2020. URL https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

Xiaomin Li, Mingye Gao, Yuexing Hao, Taoran Li, Guangya Wan, Zihan Wang, and Yijun Wang. MedGUIDE: Benchmarking Clinical Decision-Making in large language models, 5 2025. URL https://arxiv.org/abs/2505.11613.

Zhentao Xu, Mark Jerome Cruz, Matthew Guevara, Tie Wang, Manasi Deshpande, Xiaofeng Wang, and Zheng Li. Retrieval-Augmented Generation with Knowledge Graphs for Customer Service Question Answering. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2905–2909, 7 2024. doi: 10.1145/3626772.3661370. URL https://arxiv.org/abs/2404.17723.

Appendix A. Reasoner Graph - More Detail

We use a topological ordering of questions, linked to one another through the answers, that progress the flow forward until a final decision prompt is encountered as shown in figure 1.At any point in the reasoning graph, a language model can determine if the given question has enough context to be able to answer the question or if the current context is insufficient or lacks the right information. In such a situation, the model could prompt the user to pick one of the answers to move the flow forward thereby creating an interactive chat-like conversation with the user.

Appendix B. Results

We generated test patient datasets to benchmark the performance of KAGR in terms of guideline adherence correctness, and we compared the performance across GPT and Claude models with a RAG framework setup and a KAGR framework setup. We also compared the baseline performance of foundational models particularly the GPT3.5, GPT4o and Claude 3.5 Sonnet models. We can see in figure 2 that the KAGR framework can easily push the correctness metric for adherence above 95% in the benchmark. For this benchmark, we took the guidelines from CPIC (Clinical Pharmacogenomics Implementation Consortium) and prepared a knowledge graph for the pharmacogenomics (PGx) domain. Most of the mistakes made tend to be related to misinterpretation of the input data by the language model or lack of correct domain knowledge to answer the questions in the framework.

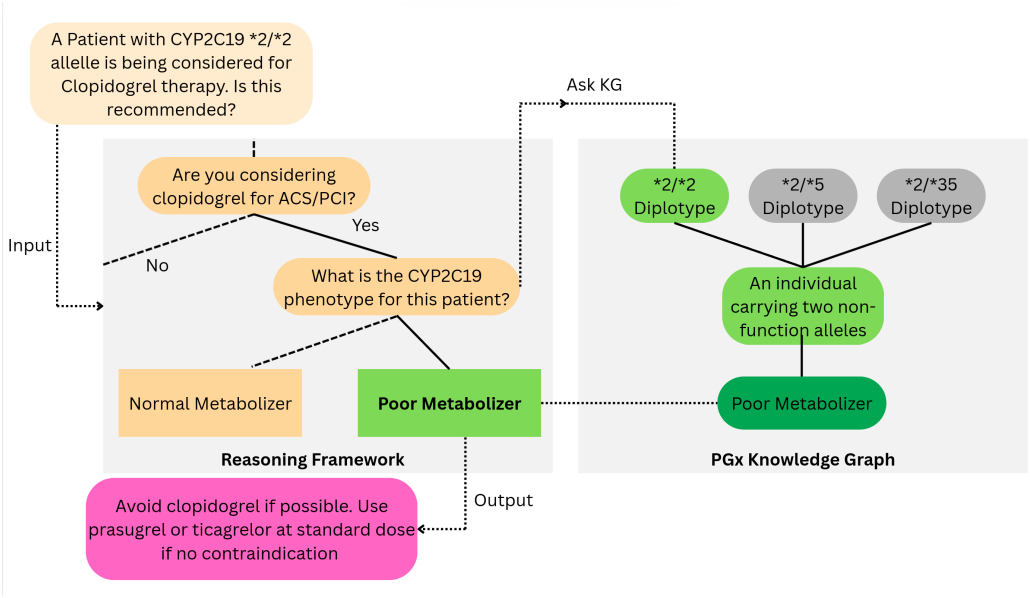

Figure 1: A reasoning step is basically a question with a set of possible answers which in turn can lead to next step

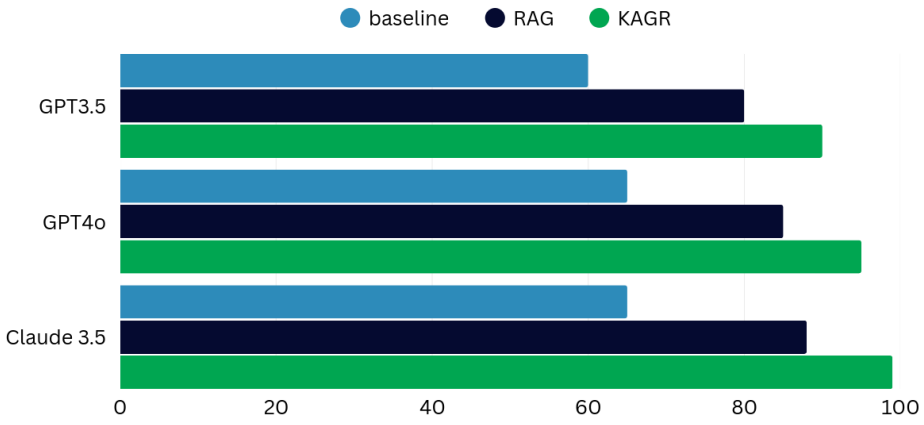

Figure 2: A benchmark to compare guideline adherence to guideline based on correctness across baseline models, a RAG framework and the KAGR framework for the CPIC guideline for Clopidogrel and CYP2C19 gene