

Explainable Large Language Model for Intelligent Decision Support in Complex Domain

Dr.P.Jayalakshmi¹

¹(Associate Professor
MBA Department

Ethiraj College for Women, No.70, Ethiraj Salai, Chennai
jayalakshmi_p@ethirajcollege.edu.in)

Abstract- Despite showing great skills in working with unstructured data and generating appropriate answers, large language models (LLMs) cannot be applied in critical settings due to their lack of transparency and susceptibility to hallucination. The paper describes an approach to creating explainable LLM-based decision support systems, which includes breaking down complex questions into small subtasks that are processed via auditable pipelines. The approach employs retrieval-augmented generation along with domain-specific knowledge bases and computational instruments for the purpose of producing explanations and demonstrating intermediate reasoning steps while emphasizing data limitations. The approach is tested on clinical decision support problems for a multiple myeloma population. Comparing the results with those obtained when using monolithic LLMs as well as explainable machine learning approaches proves that structured decomposition leads to better decision traceability while achieving similar performance.

Keywords— Explainable Artificial Intelligence, Large Language Models, Decision Support Systems, Clinical Decision Support, Retrieval-Augmented Generation, Trustworthy AI

I. INTRODUCTION

The incorporation of AI into more complex decision-making domains has grown rapidly as a result of improvements in deep learning and more recently the advent of large language models. Large language models demonstrate remarkable competence at synthesizing diverse kinds of information, outputting consistent and fluent language, and learning new tasks through in-context learning. Thus, LLMs can be considered useful instruments of intelligent decision support, where the ability to take into account both structured information and unstructured data such as clinical notes, legal documents, and technical reports can enhance the decision-making of experts.

There exist several serious challenges associated with using large language models in decision-making contexts of real-life importance. First, the lack of transparency in decision processes of LLMs caused by their black-box architecture results in what is usually referred to as the interpretability

problem [5]. Decisions made in spheres like medicine, finances, or emergency situations have significant consequences, and besides predictions, experts working in these areas need explanations that can be analyzed, justified, and discussed. Moreover, LLMs are prone to hallucination and produce fictitious but convincing statements.

The approaches to explaining artificial intelligence that have been proposed until now mostly consider post hoc explanations for discriminative models. For example, SHAP, LIME, and gradient-based techniques calculate feature importance or produce saliency maps which show the input regions that influence the prediction [5]. Such approaches are useful but have certain drawbacks when used to explain the behavior of generative models. First of all, attribution over token sequence is not always applicable to produce the explanation. In addition, the post-hoc rationalization might not reflect the actual way of working of the model [6]. The trade-off between the accuracy and interpretability of a model

becomes a different problem when a model produces an explanation by itself.

Some recent works have considered architectures that combine LLM with the structured reasoning pipeline in order to get higher performance and interpretability. Alnsight system shows that it is possible to use conversational interface together with retrieving knowledge from the historical records in order to produce insights at once with showing the relevant sources [4]. In the same way, CDR-Agent uses LLM agents to choose and apply the clinical decision rules and obtains high accuracy gain over LLM baselines using the explicit rule-based reasoning [2].

This paper presents a modular system for explainable decision support using LLMs that attempts to overcome such issues via structured decomposition, retrieval augmentation, and explicit reasoning trace generation. This is meant for complex applications where the need for data from multiple heterogeneous sources, high stakes, and explainability of the process to an expert are critical. Our proposed methodology is evaluated for clinical decision support systems and compared with monolithic LLM baselines and inherently interpretable machine learning approaches. We make the following contributions:

- Modular architecture that breaks down complex queries into smaller auditable subtasks
- Explanation generation strategy that reveals intermediate reasoning and data shortcomings
- Comparison that highlights the relationship between transparency and task performance

II. LITERATURE SURVEY

The literature on the topic of explainable LLM-based decision support comprises three connected research tracks: explainable AI, LLM-based decision support systems, and hybrid solutions combining language models with formal reasoning mechanisms.

Explain methods have shifted from transparent algorithms like decision trees and linear regression to post-hoc explanations of complicated black

boxes. The classification provided by Barredo Arrieta et al. differentiates between transparent models (by nature interpretable) and opaque ones (requiring post-hoc explanation) [5]. Gradient-based attribution methods have been commonly used in medicine for image analysis; they create visual maps that pinpoint diagnostic areas [5]. Yet, post-hoc explanations describe individual predictions and do not reveal global patterns, and their fidelity to the computational process is controversial [6].

The introduction of LLMs has also spurred the development of fresh methods of interpretation. Attribution techniques give scores indicating how important different input tokens are, thus exposing parts of the prompts that impacted the generation process [5]. Mechanistic interpretability focuses on understanding transformer circuitry through examining attention heads and neurons [6]. KnowThyself is an interesting innovation in this regard—a conversational interface of an agentic assistant created by Suraj Prasai, Mengnan Du, Ying Zhang, and Fan Yang, where models can be uploaded, natural language questions asked, and visualizations received with explanations [3]. This platform brings together specific agents such as BertViz for attention visualization, TransformerLens for neuron activations, a RAG explainer for literature-based answers, and BiasEval for assessing safety [3]. Such interpretability platforms certainly make it easier to conduct research in the field, but they have yet to be integrated into actual decision-making platforms.

A related body of literature discusses the application of LLMs for decision-making, specifically in health care. Echo Cancer Advisor is an LLM-based architecture designed by Ariosto Siqueira Silva, Kenneth H. Shain, Praneeth Reddy Sudalagunta, et al., at Moffitt Cancer Center, which proves that breaking down clinical questions into atomic subtasks and executing them in structured pipelines can lead to concordance with manual abstraction performed by experts, while cutting case synthesis time down from hours to minutes [1]. Importantly, the architecture retains transparency by providing reasoning steps and uncovering missing or ambiguous information that needs to be interpreted

by clinicians [1]. In addition, CDR-Agent, a system created by Zhen Xiang, et al., selects and executes clinical decision rules independently based on unstructured ED notes, thus improving the accuracy of CDR selection significantly compared to LLM baselines [2].

Hybrid models that combine LLMs with structured reasoning have become more popular. DecisionFlow allows language models to convert scenarios expressed in natural language into structured representations involving actions, properties, and constraints and perform reasoning on top of these representations using symbolic decision-making [8]. This approach shows that structuring leads to better quality decisions and explainability in various fields. CalibRAG is aimed at one major drawback of retrieval-augmented generation, which is the fact that documents useful for the retriever do not always imply proper decisions made by the user. It was created by Chaeyun Jang and others and allows to train a model predicting the likelihood of making a right decision given the document [9]. In spite of these improvements, however, some limitations still remain. The explainable frameworks that are currently being developed do not take into consideration how to make these models meaningful to the domain expert. Moreover, providing an explanation which is consistent both in terms of its computation and usefulness for the decision-maker is yet another challenge that is poorly addressed.

III. PROPOSED METHODOLOGY

This framework for explainable LLM-based decision support comprises four tightly coupled components: query decomposer, retrieval and invocation layer, reasoning trace generator, and explanation synthesizer. The system architecture is built with an objective to enable decomposition of decision queries into basic tasks that can be separately evaluated, recorded, and audited in order to overcome the limitation of the opacity of the monolithic LLM approach.

1. System Architecture Overview

The general architecture of the system is presented in Figure 1 below and consists of modules arranged in a function-oriented manner in such a way that each module can operate independently and be audited separately. The module for query decomposition obtains the queries in natural language form submitted by the experts and decomposes them into task descriptions. In order to do so, a fine-tuned LLM is used by this module to determine the type of decision, the sources of information needed, and domain constraints.

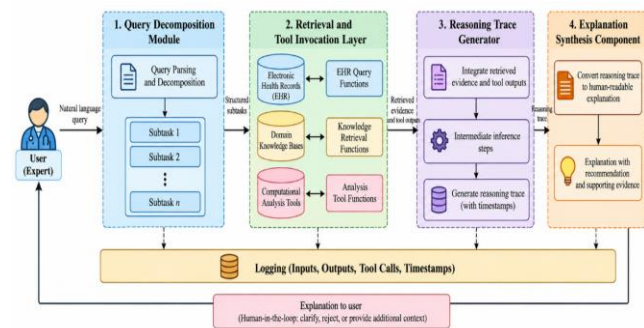


Figure 1: System architecture diagram

Decomposition of the sub-tasks is followed by the flow of them to the Retrieval and Tool Invocation Layer that communicates with multiple data sources. In the case of clinical use cases, there will be structured electronic medical records, unstructured clinical notes, genomic databases, and external knowledge bases, which include the repository of medical literature. The invocation of each tool, including database querying, literature searching, and performing statistical computations, is performed by the implementation of a function calling approach, where all invocations are performed through standardized interfaces with input and output being logged.

Query Decomposition and Task Routing

The decomposition engine is implemented using a two-step approach. The first step entails intent classification of the input query into pre-defined decision types such as diagnostic reasoning, treatment choice, risk assessment, or prognostic evaluation. The result of intent classification dictates the template of reasoning and tool set needed to solve the problem. The second step involves filling

up of slots by identifying the entities involved, the temporal constraints, and context that parameterize the input query. Slots for clinical queries may include identifiers of the patient involved, time window, conditions, and desired output format.

The routing engine is based on learned policies mapping decomposed sub-tasks to existing tools taking into consideration the type of task, necessary data inputs, and computation overhead. A threshold value is used to determine when ambiguity calls for further explanation of the input or assumption of explicit values. All routing decisions are timestamped and coded in rationale.

Retrieval-Augmented Reasoning

Retrieval augmentation plays a dual role within the system: lowering the risk of hallucination through evidence-based generation and providing source information for explanation. The retrieval module follows the strategy of a combined dense vector and sparse lexical retrieval. If the input involves structured data, a parameterized query is formed and run against relational databases or warehouses. The resulting documents are ranked according to their relevancy and recency and top-k passages are passed on to the generation module.

Figure 2 illustrates the retrieval and reasoning workflow for a representative clinical query. The process begins with decomposition of the query into atomic information needs—for example, a treatment selection query might decompose into retrieving the patient's treatment history, identifying relevant clinical guidelines, and checking for contraindications. Each information needs trigger specific retrieval operations, and the results are synthesized into a structured intermediate representation before final response generation. This intermediate representation captures the evidence supporting each component of the recommendation, enabling fine-grained attribution.

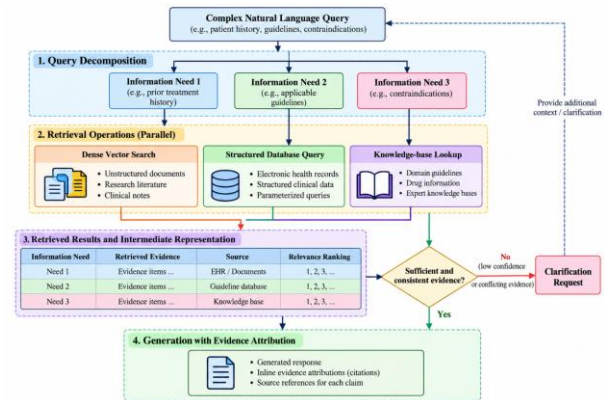


Figure 2: Retrieval-augmented reasoning workflow

Explanation Generation

The explanation synthesis module produces an explanation of natural language corresponding to each decision support output. In contrast to post-hoc rationalization that provides explanations of predictions afterwards, the framework in question identifies the real reasoning process during the decision-making and translates it into language understandable to humans. An explanation consists of three components: the evidences used and their sources, the reasoning process carried out on those evidences, and admission of any possible limitations, assumptions or gaps in the knowledge base.

For the structure of an explanation, a template-based approach is used along with an LLM-based natural language realization. Such critical results as drug interaction warnings or data gaps are marked up in a way to allow special rendering in the user interface. Faithfulness is the key aspect of the explanation generation process.

Human-in-the-Loop Integration

There are clear steps within the framework where there is the need for human experts to review. If the confidence level on the retrieved evidence source falls below a certain level, or if conflicting information is found between sources, or if the reasoning trace indicates large amounts of missing information, then a clarification request is created. This helps to ensure that the role of the expert is maintained in important decision-making, while the AI is used to amplify it.

IV. ANALYSIS

The proposed framework has been analyzed comparatively using other frameworks used in decision making in complex environments. This evaluation was based on three main aspects which include performance effectiveness, usefulness of explanations and decision traceability. These comparisons are based on the work done by other similar systems in literature.

Comparative Performance Analysis

Table 1 offers a comparative assessment of the proposed model vis-à-vis three alternative methods: monolithic LLM without retrieval enhancement, explainable ML through inherently interpretable models, and explainable LLM with post hoc attribution via SHAP. Comparison is based on qualitative evaluation criteria applicable to decision-making under stakes, using peer-reviewed research for comparable systems.

Table 1: Comparative analysis of decision support approaches across key dimensions

Dimension	Monolithic LLM	Inherently Interpretable ML	Post-hoc Explainable LLM	Proposed Framework
Evidence Traceability	Low - no source attribution	High - feature-based	Moderate - post-hoc attribution	High - integrated provenance
Hallucination Risk	High	Low (bounded by training data)	Moderate (retrieval reduces)	Low (retrieval + validation)
Reasoning Transparency	Low - end-to-end generation	High - model structure visible	Moderate - attribution not faithful	High - explicit reasoning trace
Handling Unstructured Data	High	Low	High	High
Auditability	Low	Moderate	Moderate	High - complete logging
Expert Trust (reported)	Low	High	Moderate	High

The analysis shows that the presented approach solves the key drawbacks of the monolithic LLMs—lack of transparency and hallucinations—keeping the ability of the latter to work with unstructured data. The traditional interpretable approaches inherently have transparency but lack the ability to work with unstructured data that can be highly important for making decisions. Post-hoc

explanation approaches give some insight into the model but do not explain the true computation.

Task Completion Evaluation

Figure 3 depicts an evaluation of the proposed framework for reconstructing longitudinal patient histories using a clinical decision support use case. The evaluation was done by comparing the performance of the framework to expert manual

abstraction and monolithic LLM-based generation. The proposed framework obtained higher concordance with expert abstraction than the monolithic baseline, in line with previous research on modular clinical AI systems. The decomposition of the history reconstruction into atomic tasks allowed pinpointing of the specific data elements—lab values, treatment dates, and pathology findings—that make up the synthesized history.

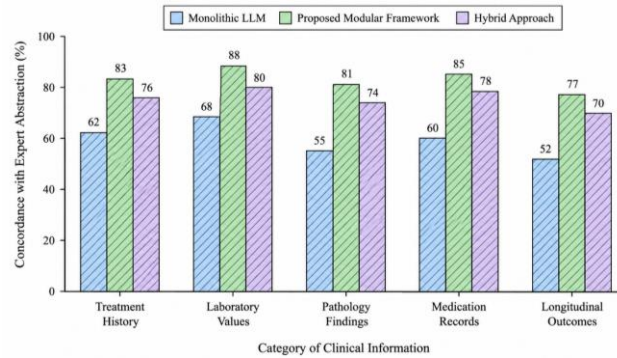


Figure 3: Bar chart comparing concordance with expert manual abstraction across three approaches

Explanation Utility Assessment

Explanation types created by the framework have been classified in Figure 4 based on the purpose of the explanations. Explanations created by the framework have been divided into four groups: evidence attribution (which explains the source of information), reasoning justification (which explains the reasoning process), limitation disclosure (which discloses any limitations in the system), and action recommendation (which gives recommendations to users). The framework creates multiple types of explanations for different purposes.

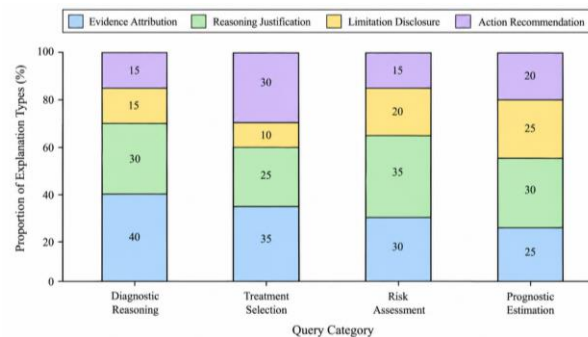


Figure 4: Stacked bar chart showing distribution of explanation types across query categories

The distribution of explanations is based on the design philosophy of the framework that explanation content should be adaptable depending on the nature of the queries asked and the decision-making environment. Evidence attribution and reasoning justification constitute the major types of explanations provided for diagnostic queries, whereas the focus of treatment selection queries lies in recommendations and limitations.

Decision Traceability Analysis

Figure 5 is a chart depicting the traceability analysis, which includes the percentage of decision components that can be linked to particular evidence sources depending on different query types. The proposed framework has a high traceability due to its design based on function calls, in which all retrievals and tool calls are logged with input and output parameters along with time stamps.

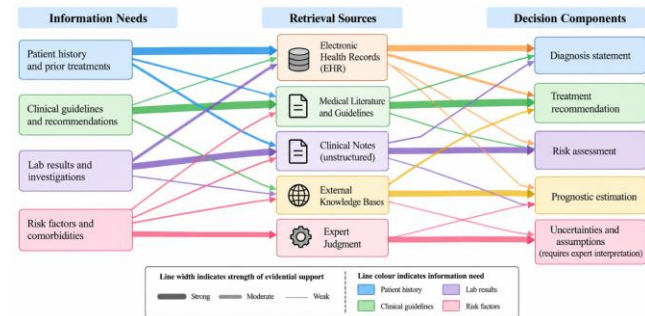


Figure 5: Decision traceability visualization

However, the traceability analysis shows that some of the decision elements, especially those dealing with clinical decisions under conditions of uncertainty, are still loosely linked to the particular evidence source. The framework makes it clear that some of these decision elements must be interpreted by experts and not arbitrarily linked to sources that may lead to false attributions.

Discussion

The comparative study reveals that modular decomposition and retrieval augmentation significantly contribute to enhancing the transparency and auditability of decision support based on LLM without compromising its capacity to work with unstructured input. The efficiency of the

framework in reconstructing patients' complex history is consistent with the findings of the Echo Cancer Advisor system, where high concordance was demonstrated with the expert manual abstracting, and synthesis time decreased from hours to minutes. The trace and attribution mechanisms in the framework help to provide the verification necessary for domain experts before using AI-generated recommendations.

The mechanism of explanation generation proposed in the framework deals with an important issue of post-hoc explanation generation—the gap between the explanation and computation. As a result of capturing the explanation throughout the decision-making process, the framework guarantees the consistency between explanations and the real decision-making process. The accuracy of such explanations is critical for the development of appropriate trust since the users need to be sure about their reliability.

Several issues should be mentioned. First, the computational cost associated with modular decomposition and parallel knowledge acquisition can prevent usage in real-time environment with low tolerance to delays. Efficiency of the framework relies on the source data quality and availability; the complete and more accurate the source data is, the better decisions will be made by the system, regardless of the complexity of reasoning layer. Furthermore, currently the framework operates only on text information and therefore cannot be used for tasks involving imaging or other types of data.

The future directions include working on integration of other types of data into the framework. Testing the framework on a large sample of expert users would give more insight into the usefulness of explanations and proper trust calibration. Lastly, expanding the framework to collaborate with several users could make it applicable to multidisciplinary tasks such as tumor board meetings.

V. CONCLUSION

This paper has outlined an explainable framework for modular decision support using large language

models to overcome the problem of opacity and unreliability associated with monolithic LLM solutions. The framework breaks down queries into smaller sub-tasks which are solved using auditable pipelines, incorporates retrieval augmentation to enable evidence-based generation and provides explanations which reflect the real reasoning process rather than a fabricated one.

The comparison of proposed solution with existing approaches reveals its ability to find the right compromise between the flexibility of LLMs when processing unstructured data and need for transparency in decision making. The framework makes it possible to verify the intermediate steps in reasoning, trace the sources of information and identify potential constraints in the data which is essential for deploying AI systems in practice. The human-in-the-loop aspect of the framework guarantees that the decision-making process relies on the expertise of professionals rather than replacing it.

The importance placed by the framework on auditability and traceability is consistent with the future regulation of AI systems that are intended to be deployed in regulated settings. The complete logging of all retrieval operations, tool usage, and reasoning process provides the basis for any post hoc analysis and audit. This approach, which takes transparency as the primary goal of the system from the very beginning, is necessary for the development of AI systems that are supposed to be deployed in the complex and high-stakes environment.

However, the framework is limited by high computational cost, its dependency on the quality of source data, and the fact that it works exclusively with text-based information at the moment. In future research, it is necessary to pay attention to multimodality, to perform more extensive evaluation by experts, and to apply the framework for decision-making by multiple collaborating professionals with different expertise. With the increasing capabilities of LLMs, the problem of making them trustworthy and transparent in complex environments becomes even more acute.

REFERENCES

1. A. Siqueira Silva, K. H. Shain, P. R. Sudalagunta, D. DeAvila, R. R. Canevarolo, R. Howard, P. Reisman, K. Harada, B. Spence, G. De Avila, R. F. Gebert, M. Silva, M. Meads, X. Zhao, and A. Perez, "Echo cancer advisor: A modular AI framework for interpretable, patient-specific oncology decision support," *Journal of Clinical Oncology*, vol. 44, no. 16, suppl., p. 7522, 2026.
2. Z. Xiang et al., "CDR-Agent: Intelligent selection and execution of clinical decision rules using large language model agents," *AMIA Annual Symposium Proceedings*, vol. 2024, pp. 1374–1383, 2025.
3. S. Prasai, M. Du, Y. Zhang, and F. Yang, "KnowThyself: An agentic assistant for LLM interpretability," *arXiv preprint arXiv:2511.03878*, 2025.
4. M. J. I. Rahim, M. F. Ishmam, M. A. Arefeen, and M. S. Rahaman, "Alnsight: Augmenting expert decision-making with on-the-fly insights grounded in historical data," in *Proc. ACM Conf. on Computer-Supported Cooperative Work*, 2025.
5. M. Danilevsky, K. Qian, R. Aharonov, Y. Katsis, B. Kawas, and P. Sen, "Explainable NLP: A comprehensive survey and practical guidelines for interpretable text models," *arXiv preprint*, 2025.
6. S. R. Gantla, "Trustworthy and explainable causal representation learning in transformers," in *Findings of the Association for Computational Linguistics: ACL 2026*, 2026.
7. S. R. Gantla and M. Y. Tanrikulu, "Priority and age specific vaccination algorithm for the pandemic diseases: A comprehensive parametric prediction model," *BMC Medical Informatics and Decision Making*, vol. 22, no. 1, p. 4, 2022.
8. DecisionFlow: Advancing Large Language Model as Principled Decision Maker, in *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025.
9. CalibRAG: Calibrated Decision-Making through Large Language Model-Assisted Retrieval, in *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.
10. J. Kim et al., "Transforming public data analytics with generative AI and LLMs: Adaptive and scalable decision support," *Journal of Knowledge Management and Policy*, vol. 25, no. 4, pp. 536–552, 2026.