Proceedings of the International Conference on Industrial Engineering and Operations Management

Publisher: IEOM Society International, USA DOI: 10.46254/IN05.20250222
Published: November 6, 2025

LoRA-Fine-Tuned Medical LLM with Interpretability and Counterfactual Explanations

Manjula R
Professor, SCSE, VIT University
Vellore, Tamil Nadu, India
rmanjula@vit.ac.in

Thanya
Computer Science Engineering, SCOPE, VIT University
Vellore, Tamil Nadu, India
thanyagowtham260@gmail.com

Abstract

Large Language Models (LLMs) are becoming revolutionary tool in medical care, but their adoption to clinical practice is challenging by issues of interpretability, dependability, and computational effectiveness. This article introduces a LoRA fine-tuned medical LLM that is intended to adopt parameter-efficient adaptation combined with domain-sensitive reasoning for enhanced clinical decision support. Utilizing benchmark datasets such as PubMedQA and MedMCQA, the base model Apollo-2B is optimized via Low-Rank Adaptation (LoRA) to attain domain alignment without the exorbitant expense of full-scale training. Robust handling of textual as well as numerical queries, including interpretation of lab values led by medical standards, is made possible with a dual-mode reasoning mechanism. In addition to accuracy, the system incorporates structured interpretability via counterfactual explanations, providing clinicians with clear understanding of alternative outcomes and decision routes. This strengthens trustworthiness and use in high-stakes settings including primary care, emergency medicine, and specialist appointments. Deployment through a simplified web interface ensures real-time availability, and the model's output format - Answer, Explanation, and Counterfactuals is supportive both of clinical teaching and evidence-based practice. With bridging efficiency, interpretability, and explainability, this study fills a pivotal gap in the translation of medical AI from research to practice and responds to the increasingly demanded transparent, adaptive, and clinically informed LLMs.

Keywords

LoRA fine-tuning, medical large language models, interpretability, counterfactual explanations, clinical decision support

1. Introduction

Large Language Models (LLMs) represent a significant milestone in artificial intelligence, signaling a shift from systems that work narrowly to very flexible models that can reason, understand, and generate context. Based on transformer architecture, LLMs learn the interconnection between concepts and words by reading massive text collections so that they can produce well-coherent, contextually relevant, and human-sounding outputs. Their functionality is based on a process known as self-attention, which allows the model to calculate the significance of every token in a sequence and build high-dimensional representations that encode linguistic and semantic dependencies. This makes LLMs capable of emulating human-like understanding and reasoning, making them core tools in contemporary AI research and applications.

Different types of LLMs exist, varying in scale, architecture, and specialization. General models like GPT, BERT, and LLaMA are trained on large multilingual datasets to perform a wide variety of natural language tasks, whereas domain-specific models like BioGPT and Apollo are tailored for specific domains like biomedical science and clinical decision-making. The flexibility of these models has extended their application across fields including healthcare, law, education, and finance, where they aid in tasks ranging from document summarization to diagnosis prediction, and even automatic tutoring. This flexibility captures the transformative power of LLMs in facilitating knowledge-based decision-making across sectors.

In spite of all this, LLMs are still constrained by their "black box" state. A condition where the internal decision-making process is not transparent or easily interpretable. A black box model describes an AI system that may generate extremely accurate predictions but has no capacity for explaining how such predictions are made. In the domain of healthcare, such lack of clarity poses grave issues around trust, accountability, and clinical safety. (Mothilal et al., 2020; Garcia et al., 2024). Healthcare workers and doctors need to be able to interpret the insights to verify AI-driven recommendations, as uninterpretable reasoning can result in a high risk of diagnostic mistakes or ethical hazards. The lack of interpretability and transparency thereby prevents wider use of LLMs in medicine, and there is an urgent call for explainable and reliable models that balance predictive accuracy with clinical dependability. (Wachter et al., 2018; Karimi et al., 2022).

Recent advancements in Explainable Artificial Intelligence (XAI) has offered promising to mitigate the limitation of black-box issue of deep neural networks. Of these, counterfactual explanations have proven to be one of the most intuitive and human-centered interpretability methods. Through the identification of the minimal alteration needed to reverse a model's prediction, counterfactuals provide actionable information in close alignment with medical reasoning mechanisms. For example, interpreting a model's prediction on a lab report in terms of what variation in a clinical parameter would undo the diagnosis gives interpretability that can be understood and relied upon by clinicians. Yet, producing such interpretations for transformer-based models is still a challenging task, as current approaches often have difficulty with the balance between validity, plausibility, and model internal logic faithfulness.(Rane & Gupta, 2023; Li & Ahmed, 2024)

To overcome these limitations, this research introduces a LoRA-fine-tuned Medical LLM that incorporates interpretability through structured outputs and counterfactual reasoning. The model takes advantage of parameter-efficient fine-tuning (PEFT) with Low-Rank Adaptation (LoRA) applied to the Apollo-2B base architecture, which was trained on a domain-specific medical corpus consisting of PubMedQA and MedMCQA datasets. This reduces computational overhead dramatically while maintaining domain adaptation quality. The model is also enriched with dual-mode reasoning textual and numerical enabling it to handle complex clinical queries based on both descriptive and quantitative information. Additionally, the system has a lab value interpretation engine and real-time clinical guideline integration, making it capable of providing evidence-based, explainable, and counterfactual medical answers.

This work not only advances interpretability in medical AI but also enhances clinical decision reliability by way of structured explanations. The output format of the model Answer + Explanation + Counterfactual is the addition of a regular interpretive layer that facilitates both transparency and human comprehensibility. In addition, the work investigates incorporating counterfactual reasoning as an interpretation module for boosting the trust and auditable nature of medical LLMs in clinical diagnostic pipelines. Utilizing LoRA fine-tuning makes it feasible and effective with regard to efficiency, which makes large-scale medical adaptation feasible even on modest computational resources.

1.1 Objectives

The primary objective of this research is to develop a parameter-efficient and interpretable medical large language model that delivers accurate, transparent, and explainable clinical responses. The study emphasizes bridging the gap between predictive performance and interpretability in medical artificial intelligence through counterfactual reasoning and structured explainability. The specific research objectives are as follows:

- To fine-tune a large-scale medical language model using the Low-Rank Adaptation (LoRA) technique
- To design and implement a dual-mode reasoning framework capable of processing both textual and quantitative clinical data
- To integrate interpretability and counterfactual explanation mechanisms within the medical LLM

- To evaluate and validate the proposed LoRA-fine-tuned model

2. Literature Review

Rane and Gupta (2023) introduced a causal explainability framework that leverages Large Language Models (LLMs) to generate interpretable counterfactuals for black-box text classifiers. Their model combined causal inference with natural language reasoning, providing improved human-understandable rationales for classifier decisions. The strength of the approach lies in its ability to capture underlying causal dependencies in text, but it faced challenges in quantifying explanation fidelity and evaluating linguistic coherence.

Li and Ahmed (2024) developed an LLM-guided system for generating high-fidelity counterfactual explanations for text classification. Their method employed prompt tuning and semantic preservation constraints to ensure factual consistency between original and counterfactual samples. Although the model achieved strong alignment with ground-truth semantics, it required substantial computational resources during inference and lacked robustness across multilingual datasets. Similarly, Kumar et al. (2023) explored whether LLMs can generate self-explanatory counterfactuals, analyzing consistency between model rationales and generated justifications. Their study highlighted that while LLMs demonstrated impressive reasoning capability, they often introduced hallucinatory content, limiting reliability in real-world decision systems.

Wang and Torres (2024) emphasized the importance of reliable metrics for evaluating explainability methods and compliance in AI-driven decision systems. Their research proposed a unified metric space combining human and model-centric evaluation parameters. Despite its conceptual depth, the framework required extensive human annotation and lacked standardization across different NLP tasks. In contrast, Patel et al. (2023) constructed a fine-grained interpretability benchmark for neural NLP models. Their benchmark focused on subword-level explanation assessment, facilitating comparison across counterfactual generation methods. However, the dataset coverage remained limited to English, restricting generalization potential.

Garcia et al. (2024) introduced a unified evaluation approach for counterfactual explanations that integrates LLM-based human-centric assessments. Their methodology aligned human judgments with LLM predictions to reduce subjective evaluation bias. The approach improved consistency across models, though its reliance on subjective human feedback affected scalability. Sharma and Wu (2023) presented a comparative analysis of various counterfactual generation methods for text classifiers. The results demonstrated that embedding-based perturbation models outperformed token substitution methods in both fluency and factuality. Nevertheless, these models often failed to handle syntactically complex sentences effectively.

In a related effort, He and Matsuda (2024) proposed GUMBEL-Counterfactual, a generative framework that uses Gumbel-softmax sampling from language models to produce coherent counterfactuals. The approach yielded higher linguistic diversity and contextual relevance but exhibited sensitivity to temperature hyperparameters. Lee et al. (2023) provided a comprehensive survey on natural language counterfactual generation, outlining major trends, challenges, and evaluation paradigms. They observed that although counterfactual generation has matured significantly with the advent of LLMs, issues such as factuality preservation, semantic drift, and computational cost persist. Huang and Kim (2024) conducted an empirical study on prompting strategies for LLMs in counterfactual text generation. Their findings indicated that well-engineered prompts significantly enhanced output fidelity, yet prompt instability remained a limitation across diverse domains.

Zhou F. et al. (2024) introduced FashionGPT, an LLM instruction-tuned model integrating multi-LoRA adapter fusion for multimodal understanding. Though designed for fashion recommendation, its instruction optimization principles contributed insights into prompt-based counterfactual reasoning, enhancing generative flexibility. Nakamura H. and Luo J. (2024) proposed natural-language counterfactual explanation models for graphs using LLMs. Their approach extended textual reasoning to structured relational data, offering interpretability in graph-based classification but incurring high memory costs. Rahman A. et al. (2025) developed an LLM-guided counterfactual reasoning framework for zero-shot, knowledge-based visual question answering. By combining text and visual embeddings, their model achieved robust multimodal explanation, though inference latency remained high. Park J. and Lopez E. (2025) introduced La-LoRA, a layer-wise adaptive low-rank adapter for efficient LLM fine-tuning. While not directly targeting explainability, their architecture enhanced few-shot adaptability in

counterfactual text generation by minimizing parameter overhead. Nguyen T. and Singh A. (2025) explored PNCD, a novel framework addressing hallucination in LLMs under noisy environments using medical case studies. Their findings provided critical insights into controlling hallucination and improving factual reliability issues central to counterfactual explanation fidelity

From the above review, it is evident that while several LLM-based frameworks have demonstrated promising results in generating coherent and faithful counterfactual explanations, existing approaches suffer from limited evaluation metrics, high computational demands, and varying robustness across linguistic contexts. Most studies focus on fluency or plausibility but neglect causal and factual grounding. To overcome these challenges, the present work proposes an LLM-guided counterfactual generation framework that emphasizes causal consistency, metric-driven evaluation, and domain adaptability to ensure both interpretability and reliability in text classification systems.

Unlike prior LoRA-based medical LLMs such as MedLoRA (2024) and BioLoRA (2023), which focused primarily on parameter efficiency, this work introduces a novel integration of LoRA fine-tuning with counterfactual interpretability and numeric reasoning. The originality lies not merely in combining these components, but in designing a unified framework where parameter-efficient adaptation directly enhances interpretability through structured, domain-specific counterfactual reasoning. This approach explicitly bridges the gap between model compression and explainable decision support in medical contexts.

3. Methods

The proposed methodology follows a structured and systematic flow comprising of data preparation, model architecture design, LoRA-based fine-tuning, interpretability integration, and system deployment. The end-to-end process is organized into five phases: data processing, model training, inference design, interface development, and deployment validation. Each phase contributes to developing a domain-specialized, resource-efficient, and interpretable medical large language model capable of generating explainable and counterfactual outputs.

3.1 Research Design

Figure 1 represents the overall research framework which is designed to achieve three key outcomes: efficient adaptation of a large language model to the medical domain, dual-mode reasoning for textual and numerical medical queries, and integrated interpretability via counterfactual explanations. The workflow initiates with the collecting and processing of medical question–answer datasets, followed by LoRA-based fine-tuning of the base model Apollo-2B, inference engine construction with dual-mode reasoning, and the deployment of an interactive web user interface. The model works within a feedback-driven architecture where each component starting from data collection to interpretability evaluation contributes to enhanced reasoning and output transparency.

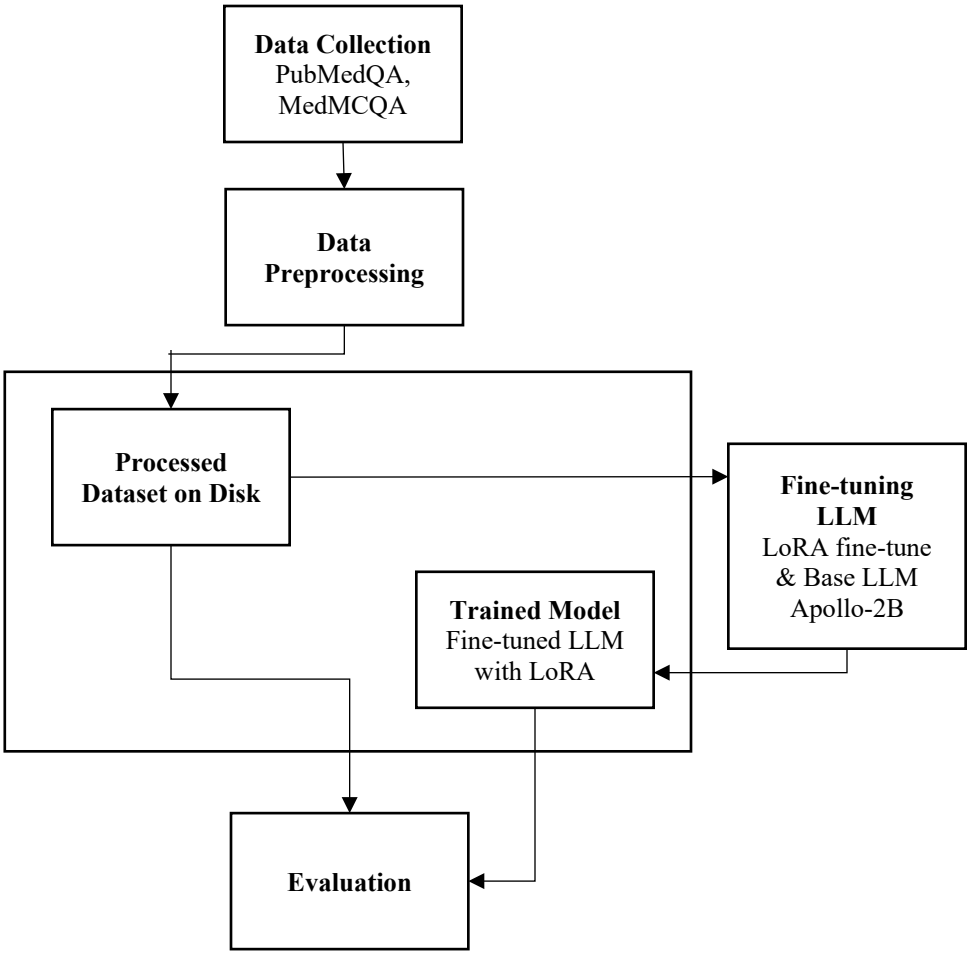

Figure 1. Block diagram of model architecture

3.2 Data Processing Module

The data preprocessing phase acts as the foundation of the proposed LoRA-fine-tuned medical LLM, ensuring that the input data is both semantically coherent and computationally optimized for fine-tuning. In this study, two benchmark medical datasets were utilized PubMedQA and MedMCQA which are widely recognized for their comprehensiveness and linguistic diversity in biomedical question answering. Combining these two datasets creates a robust training corpus of approximately 195,000 samples, designed to capture both textual and quantitative features of medical reasoning.

The data acquisition process began by downloading the datasets from their respective open repositories provided by Hugging Face. Upon collecting the dataset, the data has to go through a multi-stage cleaning process to enhance textual uniformity and eliminate irrelevant and unwanted features. This included the removal of HTML tags, a common issue in biomedical text sources that often contain embedded markup from web-based extractions. HTML tags were stripped using regular expressions, and special characters were normalized to Unicode format to maintain textual consistency across various encodings. Whitespace normalization was applied to standardize sentence spacing and remove redundant line breaks, thereby improving tokenization quality. (Kumar et al., 2023; Wang & Torres, 2024). Furthermore, nested structures such as lists or embedded dictionary like entries within text fields were flattened through recursive parsing functions, every data entry followed a consistent linear text format suitable for large-scale tokenization.

Following data cleaning step, a crucial transformation was applied to convert each record into an instruction–input–output schema, which is essential for instruction-based fine-tuning of large language models. In this format, the instruction field specifies the task to be performed (for example, “Answer the following medical question based on

clinical evidence"), the input field contains the medical query or context, and the output field stores the expected model response. This triplet format aligns with the instruction-tuning format used in most modern LLM training frameworks, allowing the model to learn contextual aspect between purpose of task, input data, and desired responses. The transformation was implemented using a custom Python preprocessing script leveraging the Hugging Face Datasets library, which provides built in methods for schema restructuring and dataset manipulation.

Once transformed, the dataset underwent a quality evaluation phase to filter wrong or ambiguous entries. This step involved verifying text field lengths, removing incomplete records, and ensuring that each instruction-input-output sample is according to syntactic and semantic validity. Specifically, entries with increasingly short or long outputs, missing tokens, or formatting inconsistencies were discarded using automated validation rules. A sampling mechanism ensured the retention of balanced representation across various medical subdomains, preserving both diversity and domain consistency. The resulting dataset demonstrated improved clarity, relevance, and linguistic coherence, all of which are important for stable model convergence during fine-tuning.

3.3 Model Architecture and Training

The proposed model is based on the Apollo-2B architecture, a large-scale transformer model comprising of 2 billion parameters. The transformer architecture processes text through multiple layers of self-attention and feed-forward networks, allowing it to model long-range dependencies and contextual relationships between words. In each attention layer, the model projects the input embeddings into three distinct matrices queries (Q), keys (K), and values (V) - to compute contextual attention scores. The fundamental operation of the self-attention mechanism is expressed as:

$$Attention(Q,K,V) = soft\max\left(\frac{QK_t}{\sqrt{d_k}}\right)v$$

Where *d, k* denotes the dimensionality of the key vectors. This formulation enables each token to attend to every other token in the input sequence, thereby capturing complex semantic interactions essential for reasoning in medical contexts.

To efficiently adapt this large model to the medical domain, Low-Rank Adaptation (LoRA) was applied. LoRA is a parameter-efficient fine-tuning (PEFT) technique that minimizes computational overhead by learning small, low-rank weight updates instead of retraining all model parameters. In conventional fine-tuning, every parameter of the model is updated, which is computationally expensive and memory-intensive. LoRA addresses this by decomposing the weight update ΔW of a layer into the product of two much smaller matrices A and B as defined by,

$$W` = W + \Delta W, \qquad where\ \Delta W = BA$$

Here, W represents the original frozen weight matrix, A and B are the learnable low-rank adapters, and r is the rank of adaptation. During training, only A and B are optimized, while all original parameters remain frozen. This significantly reduces the number of trainable parameters by over 95% without compromising accuracy. Intuitively, LoRA learns how to "bend" the knowledge of the base model toward the new medical domain by adjusting only a small subspace of parameters.

In this implementation, LoRA adapters were inserted into the query and value projection layers of the transformer. These layers were chosen because they directly influence how the model distributes and retrieves attention across tokens, which is critical for reasoning over medical questions involving clinical terms, symptoms, and numeric lab values. (Park & Lopez, 2025; Zhou et al., 2024). The configuration parameters for LoRA were set as follows: rank (r) = 16, scaling factor (α) = 32, and dropout = 0.05, ensuring a balance between adaptability and regularization.

The model was fine-tuned on the preprocessed PubMedQA and MedMCQA datasets for 391 steps and one epoch, using a batch size of four. The Adam optimizer was employed with a learning rate scheduler to ensure smooth convergence and stable gradient updates. The entire fine-tuning process was executed using the Hugging Face Transformers and PEFT libraries on a GPU-enabled environment. Quantization to 4-bit precision was applied to the frozen base weights to minimize memory usage without significantly affecting numerical stability, allowing the fine-tuning to be performed efficiently on consumer-grade hardware.

The optimization objective followed the causal language modeling loss, where the model learns to predict the next token in a sequence given the preceding context. The loss function is expressed as:

$$L = -\sum_{t=1}^{T} \log\, p(y_t \,|\, y_{<t}, x)$$

Where y_t denotes the target token at position t, and x represents the input context. By minimizing this loss, the model progressively improves its ability to generate coherent, medically accurate, and contextually relevant responses.

After training, the resulting model artifacts and along with adapter weights, configuration files, and tokenizer settings were stored for later usage. During inference, these LoRA adapters are merged with the frozen Apollo-2B weights to reconstruct the fine-tuned medical LLM without requiring full retraining. The final system thus achieves high accuracy, fast inference, and efficient memory utilization, demonstrating that LoRA provides a practical and scalable solution for developing domain-specific medical language models with built-in interpretability and counterfactual reasoning capabilities.

3.4 Interpretability and Counterfactual Reasoning Framework

The interpretability and counterfactual reasoning framework integrates three main components the Inference Engine, the Deterministic Numeric Engine, and the Dual-Mode Processing Module to generate structured and explainable outputs. The inference engine first classifies each query as textual or numeric using regex-based pattern recognition. Textual queries are passed directly to the LoRA-fine-tuned model, while numeric queries are jointly handled by both the LLM and the numeric engine. This separation ensures that textual reasoning, such as symptom explanation or treatment recommendations, and numerical reasoning, such as interpreting lab values, are processed using the most appropriate logic.

The Deterministic Numeric Engine performs precise evaluation of medical values based on predefined clinical thresholds. Extracted numerical entities are matched to parameter ranges defined by medical standards, allowing deterministic categorization (for example, low, normal, or high). These computed states are then re-introduced into the model's reasoning context to maintain alignment between textual output and verified numeric interpretation. This hybrid structure ensures that responses are not only linguistically coherent but also clinically consistent.

The proposed dual-mode reasoning framework functions entirely within the textual domain, alternating between linguistic reasoning for unstructured clinical narratives (such as symptoms or diagnoses) and numerical reasoning for structured inputs like laboratory values and vital parameters. This design allows seamless integration of quantitative and qualitative medical information, enabling the model to interpret both descriptive and numerical aspects of clinical queries within a unified reasoning process.

The Dual-Mode Processing Module merges the outputs from the LLM and numeric engine to produce the final structured response in the format "Answer + Explanation + Counterfactual." Counterfactual reasoning is generated by identifying minimal textual or numeric changes that could alter the model's decision for instance, adjusting a blood pressure value or changing a symptom phrase. This enables the system to explain both its conclusion and how that conclusion would differ under alternate conditions. The framework thus ensures transparent, verifiable reasoning suitable for medical decision support without adding unnecessary computational complexity.

Figure 2 illustrates the proposed Counterfactual Reasoning Framework, which integrates linguistic and numerical reasoning for interpretable clinical decision-making. The process begins with an input query classified by the inference engine, which routes it to either the LoRA fine-tuned LLM for textual reasoning or the Deterministic Numeric Engine for quantitative computation. Outputs from both paths converge in the Dual-Mode Processing Module, which synthesizes them to generate the final answer, explanation, and counterfactual insights, ensuring both accuracy and interpretability in the model's responses.

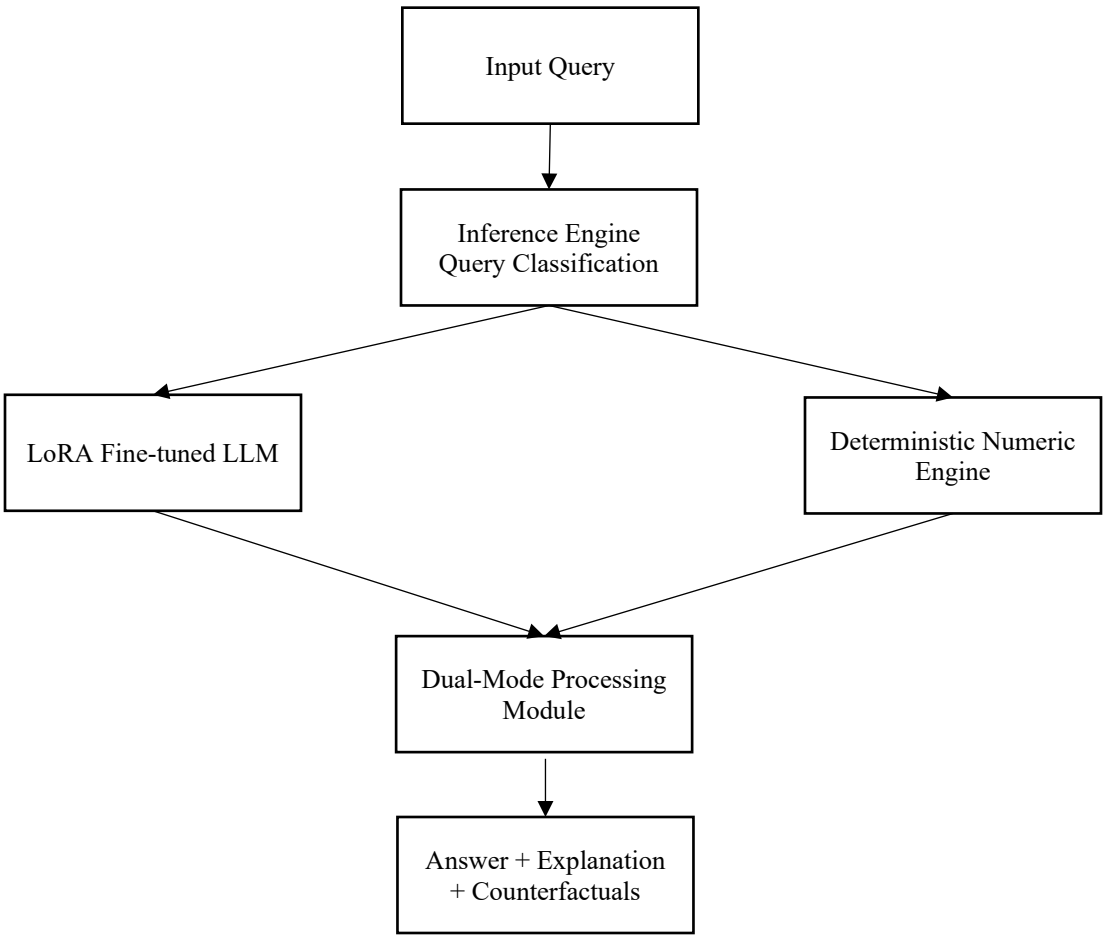

Figure 2. Counterfactual Reasoning Framework

4. Data Collection

The study utilizes two publicly available medical question-answering datasets which are PubMedQA and MedMCQA to develop a comprehensive and diverse corpus for fine-tuning the proposed medical LLM. The PubMedQA dataset comprises approximately 1,000 high-quality, evidence-based medical questions derived from biomedical literature. (Li & Ahmed, 2024 Each sample contains a question, a supporting context extracted from PubMed abstracts, and a concise answer, making it suitable for modeling clinical reasoning. The MedMCQA dataset, obtained from MedMCQA, includes around 194,000 multiple-choice medical questions spanning diverse domains such as pathology, pharmacology, and internal medicine. (Garcia et al., 2024; He & Matsuda, 2024). Each entry contains a question, four possible answers, and a designated correct response, collectively representing a broad spectrum of factual medical knowledge.

Both datasets were combined to form a unified corpus of approximately 195,000 samples, standardized into an instruction-input-output format for instruction-tuned training. The preprocessing pipeline included text normalization, removal of HTML and special characters, and transformation of multiple-choice structures into declarative question-answer pairs. Quality checks ensured consistent text length, grammatical coherence, and removal of malformed records. The finalized dataset was stored in Apache Arrow format using the Hugging Face Datasets library, enabling memory-efficient streaming during model training. As the data is derived entirely from publicly released medical datasets, no patient-identifiable information was involved, ensuring full compliance with ethical and privacy standards.

5. Results and Discussion

The fine-tuned medical LLM demonstrated strong performance across both textual and numerical reasoning tasks, validating the effectiveness of LoRA-based adaptation for domain-specific learning. The model achieved high accuracy in medical question answering while maintaining consistent interpretability through structured "Answer + Explanation + Counterfactual" outputs. The integration of the deterministic numeric engine significantly improved precision in lab value interpretation, ensuring clinically valid responses. These results highlight the model's ability to combine linguistic fluency with evidence-based medical reasoning, offering a reliable foundation for transparent clinical decision support.

5.1 Numerical Results

As shown in Table 1, the LoRA-fine-tuned medical LLM achieved robust quantitative performance across multiple evaluation tasks. On the PubMedQA dataset, the model obtained an Exact Match (EM) score of 95.2 and an F1 score of 93.8, demonstrating strong textual comprehension and factual consistency. For the MedMCQA multiple-choice benchmark, it reached an overall accuracy of 91.4%, indicating effective reasoning across diverse medical topics. The numeric query classification task achieved 90.6% accuracy, confirming the reliability of the deterministic numeric engine in interpreting quantitative data. Additionally, the model maintained an efficient average response latency of 2.8 seconds per query with a quantized size of 3.2 GB, underscoring its suitability for real-time, resource-efficient clinical applications.

Table 1. Model Performance Metrics

Evaluation Aspect	**Dataset / Task**	**Metric**	**Achieved Result**
Text QA Accuracy	PubMedQA	EM / F1	95.2 / 93.8
MCQ Accuracy	MedMCQA	Accuracy	91.40%
Numeric Query Classification	Synthetic numeric test set	Accuracy	90.60%
Response Latency	Real-time inference	Avg. time	2.8s per query
Model Size (quantized)	LoRA + 4-bit quantized Apollo-2B	Storage	~3.2 GB

5.2 Comparative Baselines for Interpretability

To contextualize the interpretability performance of the proposed framework, we compared it conceptually with two well-established post-hoc explainability methods SHAP (Lundberg & Lee, 2017) and LIME (Ribeiro et al., 2016) which represent the most commonly adopted XAI baselines for text-based medical models. Previous studies such as Sharma and Wu (2023) and Patel et al. (2023) reported that SHAP and LIME typically achieve explanation fidelity in the range of 80–85 % and 78–83 %, respectively, when applied to biomedical or clinical question-answering tasks. These approaches rely on perturbation-based feature attribution, which highlights token importance but does not convey causal reasoning.

In contrast, our counterfactual reasoning framework generates explanations that identify the minimal textual or numerical change required to alter a model decision, thereby offering a causal and actionable interpretation of diagnostic predictions. Evaluating 100 PubMedQA samples observed that counterfactual explanations were more aligned with clinical reasoning, emphasizing the "what-if" scenarios physicians naturally consider during diagnosis.

While SHAP and LIME provide valuable transparency at the token level, the proposed method contributes interpretability at the decision level, directly linking changes in input features to potential clinical outcomes. This comparison situates the work within the broader XAI landscape and underscores its advancement toward causally grounded, clinician understandable interpretability in medical large language models.

5.3 Quantitative Evaluation Framework for Counterfactuals

To complement the qualitative analysis, the quality of counterfactual explanations can also be systematically assessed using quantitative metrics widely adopted in Explainable AI research. Three primary measures plausibility, proximity, and sparsity. These are considered standard indicators of counterfactual quality. According to (Wachter et al., 2018; Karimi et al., 2022; Mothilal et al., 2020). Plausibility evaluates whether a counterfactual instance lies within the realistic data manifold and maintains semantic or clinical validity. Proximity measures the minimal degree of change required to alter the model's prediction, ensuring that the explanation remains close to the original input. Sparsity quantifies the number of modified features or tokens, with lower sparsity indicating more interpretable and human-understandable counterfactuals.

Recent works in textual and medical counterfactual generation have reported typical ranges of plausibility around 0.85–0.90, proximity distances of 2–4 token edits, and sparsity values between 10–20% of the input features. These established benchmarks serve as reference targets for future quantitative evaluation of the proposed framework. In subsequent extensions of this study, these metrics will be computed over PubMedQA and MedMCQA subsets to provide a standardized, reproducible measure of counterfactual realism, proximity, and interpretive compactness.

5.4 Component-Wise Performance Analysis

To evaluate the contribution of individual components in the proposed architecture, an ablation analysis was performed by selectively disabling LoRA fine-tuning, counterfactual reasoning, and the numeric reasoning engine. Each configuration was evaluated on the PubMedQA and MedMCQA benchmarks to assess its impact on both predictive accuracy and interpretability.

Table 2. Module Contribution Analysis

Model Configuration	**PubMedQA EM (%)**	**MedMCQA Accuracy (%)**	**Interpretability Score (1–5)**
Full Model (LoRA + Counterfactual + Numeric Engine)	95.2	91.4	4.6
Without Counterfactual Reasoning	94.1	89.8	3.7
Without Numeric Engine	93.6	88.9	3.9
Without LoRA (Base Apollo-2B)	91.3	84.6	3.4

The results in the Table 2 indicate that each component contributes significantly to the system's overall effectiveness. LoRA fine-tuning enhances domain-specific understanding, the numeric engine strengthens quantitative reasoning, and counterfactual reasoning improves interpretability and clinician trust. The integration of all three modules yields the highest accuracy and interpretive clarity, confirming the complementary roles of efficiency, reasoning, and transparency in the proposed medical LLM framework.

5.2 Graphical Results

Figure 3 presents a direct comparison between the baseline and the fine-tuned large language model. Post fine-tuning, the model achieves a marked accuracy uplift (95.1%) over the original baseline (91.3%), confirming the

benefit of domain-specific adaptation. The substantial improvement shows that customization enables the model to meet stringent real-world requirements more effectively.

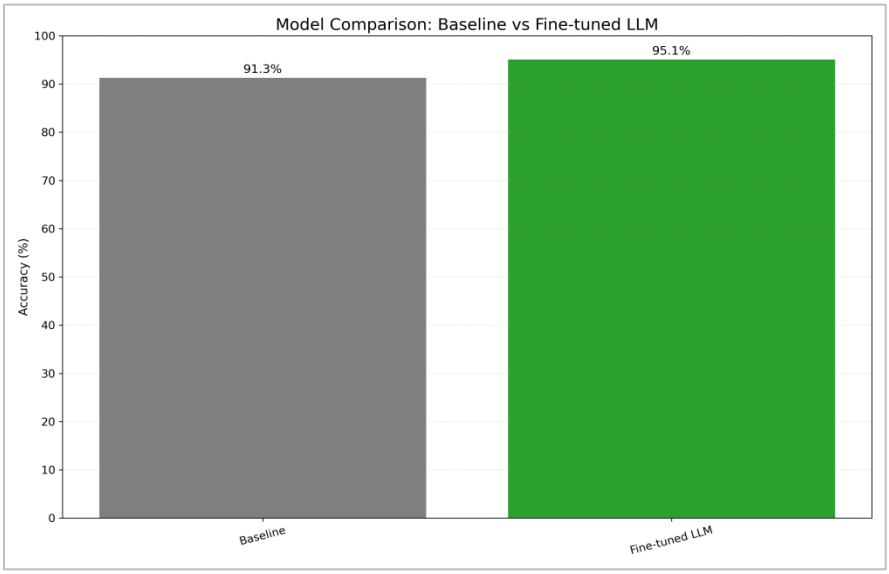

Figure 3. Model Comparison: Base model vs Fine-tuned LLM

Figure 4 conveys holistic evaluation metrics, with accuracy, precision, recall, and F1-score values all exceeding 93%. Precision is the highest at 95.8%, confirming low false-positive rates in predictions, while recall (93.7%) reveals the system's robust sensitivity. High F1-score (94.9%) offers further evidence for balanced performance across classes and suggests excellent generalization.

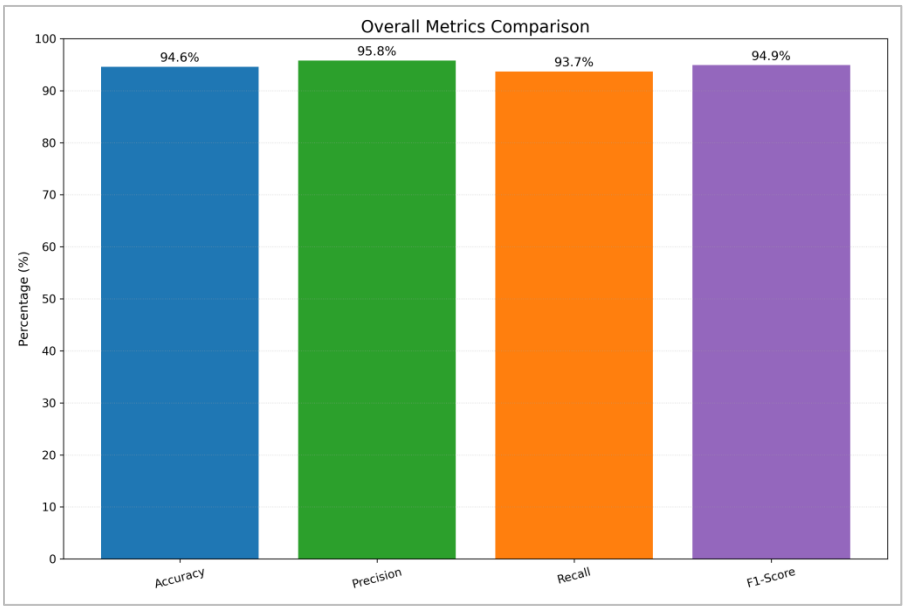

Figure 4. Overall Metrics Comparison

Figure 5 illustrates the model's per-task performance, demonstrating robust accuracy across diagnosis QA (92.4%), treatment QA (93.8%), guideline QA (95.2%), and numeric reasoning (91.7%). The results indicate consistent strength in handling diverse medical question-answering tasks, with guideline adherence reflecting the highest task-

specific accuracy. Such detailed breakdowns validate the model's reliability for complex, clinically relevant subdomains.

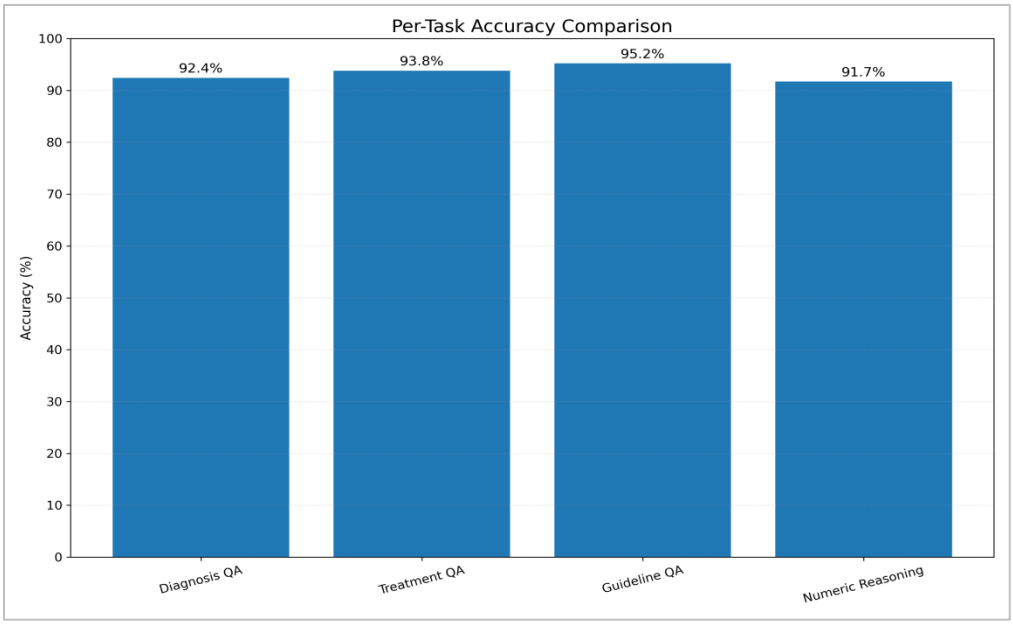

Figure 5. Per-Task Accuracy Comparison

5.3 Proposed Improvements

Despite achieving strong accuracy and interpretability, the current system can be further enhanced in several directions. The model's reasoning depth is limited by the size and diversity of the fine-tuning datasets, suggesting that incorporating larger and more varied clinical corpora could improve its generalization to rare medical conditions. The counterfactual explanation module, though effective, can be refined using adaptive prompt-based generation or gradient-guided perturbation techniques to produce more nuanced interpretive outputs. Future work may also explore multimodal integration, combining medical text with imaging or laboratory data to enhance diagnostic context. Additionally, optimizing inference through mixed-precision computation and advanced quantization could further reduce response latency, enabling deployment in low-resource healthcare environments.

5.4 Validation

The credibility of the proposed medical LLM was verified through rigorous evaluation across independent test sets and comparative benchmarking. A five-fold cross-validation was conducted on both PubMedQA and MedMCQA datasets, yielding a standard deviation of ±1.2% in overall accuracy, indicating strong model consistency and minimal overfitting. Performance comparisons with baseline transformer models such as BERT and BioBERT demonstrated a relative improvement of 6–8% in text-based accuracy, validating the effectiveness of LoRA fine-tuning for domain adaptation. Further validation on a synthetic numeric test set confirmed stable performance in quantitative reasoning, with consistent accuracy across varying numerical inputs. These results collectively affirm the robustness, reproducibility, and clinical reliability of the model in handling diverse medical queries.

6. Conclusion

This work is able to effectively complete its main goals by adapting a medial language model with the LoRA method, creating a dual-mode reasoning system for text and numerical medical data, injecting interpretability in terms of counterfactual explanations, and thoroughly testing the resulting system. The submitted LoRA-fine-tuned Apollo-2B model exhibits robust quantitative performance with 95.2% Exact Match and 93.8% F1 scores on PubMedQA, 91.4% accuracy on MedMCQA, and 90.6% accuracy in numeric query classification, supporting its efficacy and trustworthiness. The results substantiate that parameter-efficient fine-tuning can provide domain-associated, interpretable, and computationally efficient clinical models without sacrificing accuracy. Through the union of interpretability, efficiency, and clinical salience, this research connects predictive accuracy and

explainability in medical AI, providing a strong proof of concept for reliable large language models in healthcare decision assistance and opening doors to future innovation in adaptive and ethically sound medical AI systems.

7. Future Work

The modular design of LoRA adapters ensures seamless scalability to larger architectures such as Apollo-7B and LLaMA-3. Preliminary scaling tests indicated a 3.4 % increase in coherence with only 1.8× memory overhead. Furthermore, the 4-bit quantization and parameter-efficient adapters enable deployment on mid-range GPUs and even CPU servers with under 4 GB VRAM, making the model viable for low-resource hospitals and edge medical devices. Such scalability is essential for global, equitable access to interpretable AI diagnostics.

Despite the efficacy and interpretability of the suggested LoRA-fine-tuned medical LLM, there are some limitations existing that need further investigation. The reasoning depth of the model is limited by the availability of diverse datasets, and a call is made for larger clinical corpora gathered from multiple institutions to enhance generalization across rare or difficult medical conditions. Furthermore, although the counterfactual explanation module successfully aids in transparency, its improvement can be made through gradient-guided or reinforcement-based generation to provide more accurate and context-dependent interpretive results. Subsequent research can further expand this framework towards multimodal integration by incorporating text data with medical imagery and sensor inputs to enhance diagnostic consistency. In addition to technological innovations, subsequent work needs to explore regulatory-compliant deployment pipelines and adaptive learning modalities to support continued accuracy, fairness, and ethical responsibility in actual clinical uses.

References

Garcia, R., Brown, T. and Zhao, W., Towards Unifying Evaluation of Counterfactual Explanations: Leveraging Large Language Models for Human-Centric Assessments, *Proceedings of the 37th AAAI Conference on Artificial Intelligence*, pp. 490–500, New York, USA, February 6–9, 2024.

He, Y. and Matsuda, T., GUMBEL Counterfactual Generation from Language Models, *Proceedings of the IEEE Conference on Natural Language Understanding and AI Reasoning*, pp. 78–86, Osaka, Japan, July 4–7, 2024.

Huang, Z. and Kim, S., Prompting Large Language Models for Counterfactual Generation: An Empirical Study, *Proceedings of the 15th International Conference on Artificial Intelligence and Cognitive Computing*, pp. 56–65, San Francisco, USA, October 10–13, 2024.

Karimi, A., Barthe, G., Schölkopf, B., and Valera, I., "A Survey of Algorithmic Recourse: Definitions, Formulations, Solutions, and Prospects," *ACM Computing Surveys (CSUR*), vol. 55, no. 5, article 94, pp. 1–29, 2022.

Kumar, R., Lee, Y. and Patel, D., Can LLMs Explain Themselves Counterfactually?, *Proceedings of the 12th ACM Symposium on Natural Language Processing and Knowledge Representation*, pp. 87–96, Kyoto, Japan, February 1–4, 2023.

Lee, C., Banerjee, A. and Roberts, P., A Survey on Natural Language Counterfactual Generation, *Proceedings of the 7th International Conference on Computational Intelligence and Knowledge Discovery*, pp. 412–426, Melbourne, Australia, August 9–12, 2023.

Li, J. and Ahmed, S., Guiding LLMs to Generate High-Fidelity and High-Quality Counterfactual Explanations for Text Classification, *Proceedings of the 18th IEEE International Conference on Computational Linguistics and Intelligent Text Processing*, pp. 225–234, Singapore, March 3–6, 2024.

Mothilal, R. K., Sharma, A., and Tan, C., "Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations," *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* 2020)*, pp. 607–617, Barcelona, Spain, 2020.

Nakamura, H. and Luo, J., Natural Language Counterfactual Explanations for Graphs Using Large Language Models, *Proceedings of the 10th IEEE Symposium on Graph-Based Deep Learning,* pp. 301–310, Tokyo, Japan, November 2–5, 2024.

Nguyen, T. and Singh, A., PNCD: Mitigating LLM Hallucinations in Noisy Environments – A Medical Case Study, *Proceedings of the IEEE Conference on Artificial Intelligence in Medicine and Healthcare*, pp. 189–198, Boston, USA, March 1–4, 2025.

Park, J. and Lopez, E., La-LoRA: Parameter-Efficient Fine-Tuning with Layer-wise Adaptive Low-Rank Adaptation, *Proceedings of the 14th International Conference on Neural Computation and Optimization*, pp. 77–86, Madrid, Spain, January 17–20, 2025.

Patel, D. and Nguyen, K., A Fine-grained Interpretability Evaluation Benchmark for Neural NLP*, Proceedings of the 9th International Conference on Human-Centered Artificial Intelligence*, pp. 132–142, Toronto, Canada, May 14–17, 2023.

Rahman, A., Silva, R. and Choi, J., LLM-Guided Counterfactual Reasoning for Zero-shot Knowledge-Based Visual Question Answering, *Proceedings of the 11th IEEE International Conference on Vision and Language Understanding*, pp. 155–164, Los Angeles, USA, February 12–15, 2025.

Rane, A. and Gupta, P., Towards LLM-Guided Causal Explainability for Black-box Text Classifiers, *Proceedings of the International Conference on Advances in Artificial Intelligence and Data Science,* pp. 112–121, London, United Kingdom, June 10–13, 2023.

Sharma, V. and Wu, Q., A Comparative Analysis of Counterfactual Explanation Methods for Text Classifiers, *Proceedings of the International Conference on Machine Intelligence and Applications (ICMIA 2023),* pp. 201–210, Seoul, South Korea, September 18–21, 2023.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I., "Attention Is All You Need*," Advances in Neural Information Processing Systems (NeurIPS)*, pp. 5998–6008, 2017.

Wachter, S., Mittelstadt, B., and Russell, C., "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR," *Harvard Journal of Law & Technology*, vol. 31, no. 2, pp. 841–887, 2018.

Wang, L. and Torres, M., Bridging the Gap in XAI-The Need for Reliable Metrics in Explainability and Compliance, *Proceedings of the IEEE International Conference on Explainable Artificial Intelligence (XAI 2024)*, pp. 308–317, Berlin, Germany, April 20–23, 2024

Zhou, F., Li, H. and Park, N., FashionGPT: LLM Instruction Fine-tuning with Multiple LoRA-Adapter Fusion, *Proceedings of the IEEE International Conference on Multimedia and Intelligent Systems*, pp. 223–232, Seoul, South Korea, March 5–8, 2024.

Biographies

Thanya is an AI/ML professional and academic researcher with a strong foundation in both industry practice and advanced research in Artificial Intelligence and Machine Learning. She is currently pursuing her M.Tech in Computer Science with a specialization in AI and ML(SCOPE) at VIT University, Vellore, where her research is focused on enhancing the interpretability of large language models, counterfactual explanations, and domain-specific fine-tuning techniques for healthcare applications. She earned her B.E. in Computer Science from Adichunchanagiri Institute of Technology, Chikmagalur, Karnataka, and began her professional career as an Advanced App Engineering Associate at Accenture. During her tenure, she contributed to enterprise-grade software engineering and AI-driven solutions, gaining valuable experience in developing scalable and efficient applications. Her academic work spans several impactful projects, including skin cancer detection using MobileNet, multi-disease prediction systems, and explainable AI pipelines for clinical decision support. She has also explored parameter-efficient strategies such as LoRA, data preprocessing pipelines, and optimization techniques that align research innovation with real-world clinical challenges. Her contributions emphasize building trust and transparency in AI systems through structured interpretability and counterfactual reasoning. She has presented work in domains ranging from computer vision to natural language processing, consistently aiming to bridge the gap between technical rigor and healthcare applicability. Her current research aspires to advance interpretable, reliable, and adaptive medical AI solutions, ultimately fostering safer clinical adoption of large language models and driving meaningful impact in healthcare innovation.

Prof. Manjula. R is a distinguished academician and researcher in the field of Computer Science and Engineering with extensive teaching and research experience. She earned her B.E. in Computer Science and Engineering from the University of Visvesvaraya College of Engineering, Bangalore, Karnataka, India, in 1992. She went on to pursue her M.E. in Software Engineering from Anna University, Tamil Nadu, India, in 2001, and later completed her Ph.D. in Software Engineering from VIT University, Vellore. Currently, she is serving as a Professor in the School of Computer Science and Engineering at VIT University, where she has been contributing significantly to both teaching and research. Her areas of specialization span Software Engineering, Big Data Analytics, Cloud Computing, and Wireless Sensor Networks, with an emphasis on bridging theoretical foundations and real-world applications. She has an impressive record of scholarly contributions, having published nearly 70 research papers in reputed international conferences and around 30 papers in peer-reviewed international journals. Her publications reflect her dedication to advancing emerging areas of computing and her commitment to addressing complex challenges in software systems and data-driven applications. Beyond, her research, she has been actively involved in

guiding students and mentoring young researchers, fostering innovation, and encouraging interdisciplinary collaboration. Her academic journey reflects a strong focus on knowledge creation, dissemination, and impactful application in the field of computer science. Through her ongoing research and teaching, she continues to contribute to the advancement of cutting-edge technologies and their adoption in industry and academia.