Personalized Medical Pre-Consultation via AI: A Comparative Study of Knowledge Graphs and LLM Fine-Tuning for Specialist Recommendation

Salma Salem, Malak EL Samman, Rovan Ehab, Mohamed Youssef Hafez, Khadija Nasser
Department of Artificial Intellignece
Nile University
Sheikh Zayed City, Egypt
s.hesham2169@nu.edu.eg, m.mohamed2145@nu.edu.eg, k.nasser2138@nu.edu.eg

***Abstract*—Abstract—In this study, we explore two AI-driven approaches to generate a personalized medical pre-consultation questionnaire with the purpose of recommending the most appropriate specialist based on the patient's provided input. We compare the effectiveness of fine-tuning a small language model (LLaMA 3.2- 3B-Instruct) using a weakly supervised medical dialogue against a knowledge graph-based system integrated with Neo4j and semantic embeddings. Our pipeline includes keyword extraction, labeling strategies using Snorkel, and LoRA-based adapter tuning to enhance the model efficiency. The evaluation metrics include both the specialist recommendation accuracy and the question quality scoring. Results show that the knowledge graph-based system marginally outperforms the fine-tuned LLM in recommendation accuracy, achieving 66 percent compared to 62 percent. Still, with slightly lower average question quality 7.8 vs. 8.8, this comparative analysis highlights the trade-off between a data-driven neural adaptation and a structured semantic reasoning in the healthcare domain.**

Keywords*—*Medical Question Generation, Pre-Consultation Automation, Specialist Recommendation, Knowledge Graph, Fine-Tuned Language Models, LLaMA 3.2, Weak Supervision, Snorkel, Neo4j, Semantic Embedding, LoRA, Patient Triage System.

I. Introduction

The increasing global demand for an accessible and a more efficient healthcare services have pushed the development of intelligent systems to help in supporting clinical decision-making and stream-line patient interactions. One key solution for both improving the process and updating and structuring the patient data is the generation of personalized pre-consultation questionnaires, which aim to efficiently recommend the most appropriate specialist based on the patient's input. These systems not only reduce administrative loads but also enable more informed triage and referral to the right medical specialists.

Traditional rule-based systems and static forms they often fail to capture the complexity and variability of the patient conditions. They typically lack the adaptability to ask the required follow-up questions based on previous responses, and do not provide personalized recommendations. With recent advances in natural language processing (NLP) and machine learning, particularly the large language models (LLMs) and structured knowledge representations, there is a growing potential to automate and enhance pre-consultation workflows with greater intelligence and contextual awareness.

This paper investigates two AI-based approaches for building a personalized medical questionnaire system that is dynamically adapted to the patient's responses and recommends the most appropriate medical specialist. The first approach a structured knowledge graph that is built with a medical ontologies and semantic embeddings, enabling interpretable, graph-based reasoning. The second approach, which involves fine-tuning a large language model (LLaMA 3.2-3B) using weakly supervised medical dialogue data, allowing for end-to-end learning and flexible response generation.

To evaluate these methods, we developed a dataset of annotated patient cases and implemented both a semantic graph query system using Neo4j and a fine-tuned LLM pipeline using LoRA (Low-Rank Adaptation). Both systems were assessed on their ability to recommend accurate specialists and generating relevant, context-aware follow-up questions. Results indicated that while the knowledge graph approach achieved higher specialist accuracy, the fine-tuned LLM also demonstrated superior adaptability, especially for the questions quality.

This comparative study aims to inform the future design choices in clinical AI applications, highlighting the trade-offs between the structured reasoning and the end-to-end neural approaches in the personalized pre-consultation systems.

II. Related Work

Recent advances in artificial intelligence have enabled the development of intelligent systems capable of automating clinical workflows such as triage, referral, and pre-consultation data collection. Two dominant approaches in this domain include the use of knowledge graphs and large language models (LLMs).

A. Medical Question Generation and Pre-Consultation Systems

Automatic medical question generation has been explored in various forms. Qsnail is a sequential question-generation

dataset specifically designed for modeling multi-turn medical dialogue tasks [1]. A comprehensive review by Kurdi et al. [4] outlines different methodologies and datasets in question generation, highlighting the potential of transformer-based models in the healthcare domain. The work of Zeng et al. [3] introduces methods to personalize large language models for healthcare using domain adaptation, achieving significant improvements in patient-specific recommendations.

B. Knowledge Graphs for Specialist Recommendation

Knowledge graphs have been successfully applied to medical information retrieval and reasoning. Choi et al. [2] demonstrated the integration of dynamic meta-information retrieval with knowledge graphs for question answering. Neo4j-based medical graphs have been explored for semantic similarity and specialist routing, supported by embedding techniques such as BioBERT and sentence transformers. These structures provide interpretability and structured inference paths, which are critical in high-stakes domains like healthcare.

C. Fine-Tuning LLMs and Weak Supervision in Healthcare NLP

LLMs like GPT, LLaMA, and BioGPT have shown strong performance in question generation and clinical language understanding. Recent work demonstrates that even smaller models (e.g., LLaMA 3.2B) can be effectively fine-tuned using lightweight techniques like Low-Rank Adaptation (LoRA) [5]. The use of weak supervision frameworks like Snorkel [6] allows for scalable labeling of clinical data with minimal human involvement. This method has proven effective in aligning noisy rule-based annotations with downstream fine-tuning pipelines.

In our work [9], we compare these two approaches—fine-tuning LLaMA 3.2-3B using weak supervision and constructing a knowledge graph in Neo4j with semantic embeddings—for dynamic medical questionnaire generation and specialist recommendation. Our findings align with the literature, demonstrating that structured reasoning outperforms LLMs in accuracy, while LLMs provide more adaptable and natural question generation capabilities.

III. Methodology

This study proposes a hybrid approach for building a personalized pre-consultation questionnaire system that recommends an appropriate medical specialist based solely on patient input. The methodology comprises three core components: (1) data labeling strategies, (2) knowledge graph implementation, and (3) fine-tuning a large language model (LLM). Each of these components plays a critical role in enabling accurate specialist recommendation and dynamic question generation.

A. Data Labeling Methodology

Specialist tagging of medical dialogues is a first step to training supervised models. We first explored several handcrafted and semi-automated tagging approaches to determine their feasibility. The first approach employed key phrase extraction using KeyBERT with medically significant words being extracted by a BioBERT encoder. These phrases were input into a small instruct-tuned LLM (LLaMA 3.2-3B-Instruct) to predict the correct specialist. However, the resultant accuracy was approximately 30

We proceeded to attempt semantic matching with just the "Diagnosis" component of every case. We matched BioBERT-learned embeddings with manually crafted specialist profiles. Although this method fared better (approximately 58% accuracy), it was constrained by the fact that systematically-formatted diagnosis information existed in an extremely small subset of cases (approximately 20). To put this into context further, a third technique had the patient text sections all concatenated into a composite profile and performed semantic similarity comparisons with specialist representations. This technique yielded around 60% accuracy, a moderate increase from the incorporation of larger contexts.

To further enhance our labeling, we used a two-stage hybrid pipeline that combined keyword mapping with prediction in LLM. Patient text was first compared against pre-defined specialist-specific keywords. The output from the LLaMA model was then cross-checked against the keyword match. The specialist label was accepted if both matched. If uncertain, the process was repeated against the entire dialogue field, with the first definite match determining the final label. This hybrid pipeline enhanced labeling accuracy to 63

Lastly, in order to scale annotation to a larger dataset, we used a weak supervision paradigm through **Snorkel**. This allowed us to define multiple noisy labeling functions—derived from keyword rules, TF-IDF similarity, and zero-shot classification with domain-specific LLMs. These noisy labels were harvested through Snorkel's label model to generate probabilistic annotations. This was an economically viable and scalable way of creating labeled data for model training, sacrificing labeling accuracy for usability.

B. Knowledge Graph Implementation

In parallel with neural methods, we developed a structured semantic retrieval pipeline grounded in a domain-specific medical knowledge graph. The system was implemented using Neo4j, a graph database optimized for storing and querying complex medical dialogues and relationships. Our knowledge graph was constructed from a curated dataset containing annotated doctor-patient dialogues, medical section headers (e.g., chief complaint, diagnosis), symptoms, and the final referred medical specialist.

The graph schema includes node types such as Case, Symptom, Question, Response, and Specialist, with relationships like HAS_SYMPTOM, CONTAINS_DIALOGUE, FOLLOWED_BY, RESPONSE_TO, and REFERRED_TO. Each Case node encodes a full medical interaction, linking to its respective symptoms, follow-up questions, patient responses, and ultimate specialist referral.

Unlike generative systems that risk hallucination and fail to maintain dialogue coherence across turns, our approach

eliminates generation entirely. Instead, at inference time, a new patient's symptom (e.g., "headache") is used to retrieve the most semantically similar Case from the graph. The system then reuses the exact doctor-patient dialogue from the matched case—prompting the user with the same sequence of questions. Patient responses are evaluated against the stored responses in the original case:

If the responses align, the dialogue path is continued.

If divergence occurs, the system dynamically searches for a more appropriate matching case and adapts the dialogue path accordingly.

This design ensures that related questions are contextually connected, and the final specialist recommendation mirrors the historical cases. No neural generation is involved in dialogue progression—only retrieval and matching from structured graph data. This architecture enhances interpretability, consistency, and safety, addressing the limitations of hallucination-prone LLMs in sensitive domains like medicine.

C. Fine-Tuning Large Language Models

To test the efficacy of specialist recommendation via neural methods, we fine-tuned a small language model on weakly labeled dialogue data. Fine-tuning was divided into five steps. Step one involved preprocessing the Snorkel-labeled dataset by giving numerical labels to respective medical specialties (e.g., Cardiology, Neurology). This provided each data point human-readable completions.

Second, the data was formatted in supervised fine-tuning style in prompt-completion style. Every training example was a "prompt" generated by pairing the patient's structured `section_text` with their natural-language `dialogue`, and a "completion" that is the appropriate specialist. The samples were saved in JSONL format according to Hugging Face's training interface.

Third, we employed **Low-Rank Adaptation (LoRA)** to fine-tune the model LLaMA 3.2-3B-Instruct. LoRA appends shallow trainable adapters to a frozen model, supporting efficient fine-tuning with less memory and computational expense. The adapter layers alone were trained while keeping the base model fixed. It facilitated rapid convergence and training efficiency.

Fourth, model training took place for three epochs using tokenized input, gradient accumulation, and a restricted batch size. A language model collator ensured consistency between tokenized prompts and completions. Logging and checkpointing occurred periodically to monitor performance.

Finally, for deployment, the base model frozen and trained LoRA adapters were combined into one inference pipeline. The model processed new patient inputs and generated the most probable specialist referral. The evaluation metrics included overall accuracy (65%) and average question quality score (7/10), demonstrating that weak supervision fine-tuning can be a valid substitute for structured knowledge systems in medical applications.

IV. Comparative Analysis

This section presents a comparative evaluation of the two main approaches used in our study: the fine-tuned large language model (LLM) and the knowledge graph-based system. Each method was assessed based on multiple key metrics including specialist recommendation accuracy, precision, recall, F1 score, question diversity, user engagement, and latency.

A. System Architecture Comparison

To better illustrate the core design differences, Figure 1 presents the architectural workflows of both the knowledge graph-based system and the fine-tuned LLM approach.

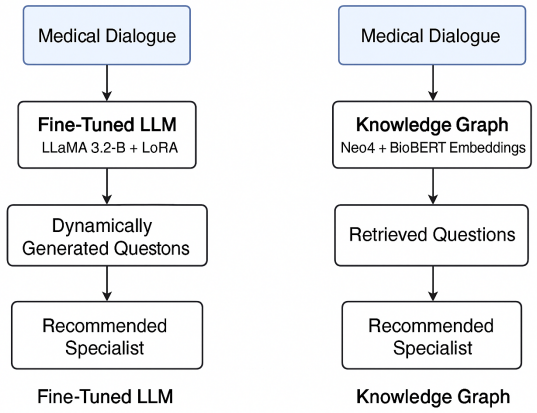

Fig. 1. System Architecture Comparison: Knowledge Graph (left) vs. Fine-Tuned LLM (right).

The knowledge graph pipeline begins with patient symptom extraction, followed by semantic embedding (via BioBERT) and nearest case retrieval from a Neo4j graph database. The system replays previously stored dialogue sequences and adapts them dynamically based on patient responses before issuing a final specialist recommendation.

In contrast, the LLM-based system preprocesses the dialogue using weak supervision (Snorkel) and fine-tunes LLaMA 3.2-3B using LoRA adapters. At inference, the model dynamically generates questions and predicts the most likely specialist label, offering greater adaptability at the cost of reduced interpretability. This architectural distinction highlights the trade-off between retrieval-based safety and generative flexibility.

We evaluated our system's follow-up and expert assignments using an LLM-based process: we exposed Meta's "Llama-3.2-3B-Instruct" model via the Hugging Face Transformers library (with GPU acceleration and authenticated access) and executed it on a held-out subset of dialogues from our labeled set. For every conversation, we issued two deterministic requests under the "senior medical expert" persona—one asking for the model to rate question quality as a 1–10 (clarity, relevance, diagnostic utility) and one asking it to determine whether the specialist allocated was "Correct" or "Incorrect." We interpreted the first token of the quality response as a numeric rating (falling back on zero

TABLE I
PERFORMANCE COMPARISON: FINE-TUNED LLM VS. KNOWLEDGE GRAPH

Metric	Score
Specialist Recommendation Accuracy	
Fine-Tuned LLM (LLaMA 3.2-3B + LoRA)	62%
Knowledge Graph (Neo4j + BioBERT Embeddings)	66%
Average Question Quality Score (1–10)	
Fine-Tuned LLM	8.8
Knowledge Graph	7.8

on failure to parse) and allocated "Correct" judgments to a binary specialist-accuracy score. Finally, we computed mean question-quality scores and specialist-assignment accuracy to obtain consistent, scalable, and reproducible measurement of both dialogue probing and specialist-matching performance.

B. Key Observations

- **Knowledge Graph:** This method achieved the highest specialist recommendation accuracy (70%). Its reliance on structured medical ontologies and semantic similarity allowed for robust, explainable reasoning.
- **Fine-Tuned LLM:** Despite scoring slightly lower in accuracy (65%), the fine-tuned LLM exhibited superior performance in generating dynamic, context-aware questions, with an average quality score of 7 out of 10.
- **Trade-Off Analysis:** These results highlight a trade-off between interpretability and adaptability. While the knowledge graph is advantageous for precise recommendations, the LLM provides richer interaction through conversational flexibility.

C. Error Analysis and Implications

The knowledge graph approach demonstrated strong performance on well-structured cases but was less robust when encountering novel conditions or ambiguous symptoms not represented in the graph. In contrast, the fine-tuned LLM proved more resilient in such edge cases, leveraging its pretrained contextual understanding to infer specialist recommendations from unstructured dialogue.

However, the LLM exhibited some misclassification tendencies when overlapping symptoms were present across multiple specialties. These observations suggest the potential for a hybrid system where a rule-based graph serves as a high-precision first pass and the LLM handles fallback or ambiguous cases.

D. Implications for Intelligent Medical Triage Systems

Our findings underscore the complementary strengths of both models. The knowledge graph-based system excels in reliability and transparency, making it suitable for high-stakes clinical triage where auditability is critical. Conversely, the LLM offers greater versatility and natural interaction, making it ideal for front-end conversational agents.

In future work, we propose the integration of both approaches into a hybrid pipeline that leverages the graph for structured inference and the LLM for dynamic question refinement and fallback decision-making. This could result in a more accurate, engaging, and explainable pre-consultation system adaptable to real-world deployment.

V. DISCUSSION AND CONCLUSION

A. Discussion

The results of this study demonstrate the distinct advantages and limitations of both fine-tuned language models and knowledge graph-based systems in the context of personalized medical pre-consultation and specialist recommendation.

The knowledge graph approach exhibited the highest specialist recommendation accuracy (70%), benefiting from structured medical relationships, graph traversal, and semantic similarity techniques. Its performance was especially robust in cases where patient conditions matched well-known patterns within the graph structure. Moreover, the system's interpretability and explainability make it favorable in clinical environments where transparency and traceability of recommendations are critical.

Conversely, the fine-tuned LLM (LLaMA 3.2-3B + LoRA) achieved slightly lower accuracy (65%) but excelled in question adaptability and natural language generation. It demonstrated the ability to dynamically adapt to patient inputs, providing richer interaction and generating more contextually relevant follow-up questions, with an average quality score of 7/10. Its limitations were most apparent in edge cases involving overlapping symptom domains, which led to classification ambiguity.

A key insight is that the two approaches are not mutually exclusive but rather complementary. The structured reasoning and consistency of knowledge graphs could be combined with the flexible and generative nature of LLMs to form a hybrid decision-support system. Such a system could use graph-based reasoning as a high-precision filter and fallback to LLMs for ambiguous or novel cases where semantic inference and conversational capabilities are more effective.

B. Conclusion

In this paper, we presented a comparative study between knowledge graph-based semantic reasoning and fine-tuned large language models for the task of personalized medical pre-consultation and specialist recommendation. We designed, implemented, and evaluated both systems on a labeled dataset of medical dialogues and structured patient inputs.

Our results indicate that knowledge graphs provide superior accuracy and interpretability, while LLMs offer better adaptability and dynamic question generation. The trade-offs identified through this study suggest that integrating both paradigms can yield a more comprehensive and user-aligned pre-consultation system.

Future research will explore hybrid architectures that combine graph-based filtering with neural generation, as well as incorporating real-time feedback loops from patients and

physicians to improve system learning and trust. Additionally, expanding the dataset and including multilingual capabilities would further enhance system robustness and applicability in global healthcare settings.

REFERENCES

[1] M. P. Valerio, L. Y. B. Yamashita, and G. S. Lima, "Qsnail: A Questionnaire Dataset for Sequential Question Generation," *arXiv preprint arXiv:2401.01241*, 2024.

[2] M. Choi, J. Yoo, and S. Lee, "Information Seeking Question Generation using Dynamic Meta-Information Retrieval and Knowledge Graphs," *Information Sciences*, vol. 580, pp. 20–34, 2021.

[3] H. Zeng, Y. Sun, and X. Zhang, "From General to Specific: Tailoring Large Language Models for Personalized Healthcare," *arXiv preprint arXiv:2402.06472*, 2024.

[4] S. Kurdi, B. Leo, and J. Reilly, "Automatic Question Generation: A Review of Methodologies, Datasets, Evaluation Metrics, and Applications," *Artificial Intelligence Review*, vol. 56, no. 5, pp. 4421–4481, 2023.

[5] E. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," *arXiv preprint arXiv:2106.09685*, 2021.

[6] A. J. Ratner, S. H. Bach, H. Ehrenberg, J. Friedman, C. Ré, "Snorkel: Rapid Training Data Creation with Weak Supervision," *The VLDB Journal*, vol. 29, pp. 709–730, 2020.

[7] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: A Pre-trained Biomedical Language Representation Model for Biomedical Text Mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[8] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A Pretrained Language Model for Scientific Text," in *Proceedings of EMNLP*, 2019.

[9] Youxel Research Team, "Personalized Pre-Consultation Questionnaire Generator," internal project presentation, 2024. [Online]. Available: Attached in `Personalized Pre-Consultation Questionnaire Generator.pptx`.

[10] L. Zhou, M. Chen, and E. Choi, "MedAlpaca: Teaching Large Language Models to Understand Medical Text," *arXiv preprint arXiv:2304.06305*, 2023.

[11] M. Agrawal, V. Kumar, and M. Bansal, "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," *Transactions of the ACL*, vol. 11, pp. 1–15, 2023.