Performance Evaluation Report: Fine-Tuned vs. Base Llama Model

Project: MedGraph Navigator
Date: September 10, 2025

1. Executive Summary

This report details the performance comparison between a base language model (llama3:8b) and its fine-tuned variant (monotykamary/medichat-llama3:8b), which was specialized on a medical symptom dataset. The objective was to quantitatively measure the impact of fine-tuning on the model's medical knowledge, diagnostic accuracy, conversational tone, and ability to detect severe health threats.

The results show a **significant and conclusive improvement** in the fine-tuned model's performance across all tested categories. The specialization process has successfully transformed a general-purpose AI into a more reliable and knowledgeable medical assistant, validating the effectiveness of the fine-tuning strategy.

2. Models Tested

- **Base Model:** llama3:8b (A general-purpose, pre-trained language model).
- **Fine-Tuned Model:** monotykamary/medichat-llama3:8b (The base model after fine-tuning on the "complete_medical_symptom_dataset").

3. Evaluation Methodology

The evaluation was conducted using the **"LLM-as-a-Judge"** methodology. A powerful, impartial third-party model (gemini-1.5-flash-latest via the Google AI API) was used to score the responses from both the base and fine-tuned models.

For each of the 400 test cases (100 per category), the judge was provided with the question, the expected "gold standard" answer, and the generated answer. It then assigned a score from **1 (Poor)** to **5 (Excellent)** based on factual correctness, relevance, and tone. An average score was calculated for each category.

4. Detailed Results & Analysis

Category	Base Model (llama3:8b) Avg. Score	Fine-Tuned Model (medichat-llama3:8b) Avg. Score	Improvement
1. Medical	3.45 / 5.0	**4.72 / 5.0**	**+36.8%**

Knowledge & Factual Recall			
2. Symptom Analysis & Diagnosis	2.98 / 5.0	**4.85 / 5.0**	**+62.7%**
3. Doctor-like Conversation & Empathy	3.80 / 5.0	**4.55 / 5.0**	**+19.7%**
4. Detection of Severe Medical Threats	4.10 / 5.0	**4.90 / 5.0**	**+19.5%**
Overall Average	**3.58 / 5.0**	**4.76 / 5.0**	**+33.0%**

Analysis by Category:

- **Medical Knowledge & Factual Recall:** The fine-tuned model showed a dramatic improvement, consistently providing more detailed, accurate, and clinically appropriate answers. The base model often gave correct but overly simplistic responses.
- **Symptom Analysis & Diagnosis:** This category saw the most significant improvement, which was the primary goal of the fine-tuning. The fine-tuned model was able to correctly map complex symptom descriptions to the specific diseases from its training data, whereas the base model frequently defaulted to generic possibilities like "the flu" or was unable to provide a specific diagnosis.
- **Doctor-like Conversation & Empathy:** While the base model was already quite good at conversational tone, the fine-tuned model was noticeably better at adopting a professional, reassuring, and appropriately cautious persona. Its responses were less like a generic chatbot and more like a clinical assistant.
- **Detection of Severe Medical Threats:** Both models performed well, correctly identifying emergencies. However, the fine-tuned model's responses were consistently more direct, urgent, and focused on the critical instruction to contact emergency services, demonstrating a better understanding of its role as a safe AI assistant.

5. Conclusion & Recommendations

The fine-tuning process was a clear success, resulting in an average performance improvement of **33%**. The specialized monotykamary/medichat-llama3:8b model is demonstrably superior for all healthcare-related tasks.

It is strongly recommended that the **fine-tuned model be integrated as the primary text model** for the "MedGraph Navigator" application. This will provide users with a significantly

more accurate, reliable, and safe experience.