# AI-Powered Medical Patient Simulator: Comparing Prompt Engineered LLMs with Exploratory Fine-Tuning

Muhammad Ibrahim Arain
*Department of Computer Science*
*Queen Mary University of London*
London, UK
m.i.arain@qmul.ac.uk

***Abstract*—This project looks at building a web-based virtual patient simulator that can act as a realistic training tool for medical students and healthcare professionals. Large Language Models (LLMs) were tested to see how well they could take on patient roles when guided by carefully designed prompts and grounded in a medical dataset from NHS sources with a small-scale fine-tune attempt for experimental findings. The system was developed with multiple patient personas to make interactions feel closer to real consultations, and it was evaluated through both automated checks and human review. The work explores how combining prompt engineering with a verified dataset can create scalable, flexible training scenarios with low symptom hallucinations without the need for costly fine-tuning, highlighting the potential for AI-driven medical simulations to expand access to consistent, on-demand clinical training.**

***Index Terms*—Medical Education, Large Language Models, Patient Simulation, Prompt Engineering, Virtual Patients**

## I. Introduction

Virtual patient simulations are often used in medical training because they allow students to practice repeatedly without needing real patients or actors. The problem is that most of these are still based on fixed scripts or follow a set path, which makes them feel less like a real consultation. With the recent progress in Large Language Models (LLMs), it is now possible to have a patient simulation that can change its responses based on what is asked, while still sticking to accurate medical information [1], [23].

In this project, a new type of simulator was built that combines LLMs with a dataset taken from NHS medical pages. This dataset lists different conditions and their symptoms, so the LLM only pulls from correct information. The aim was to make the conversation more natural, have the patient behave differently depending on their persona, and still make sure the information given was correct.

The system was tested using several LLMs, using custom metrics to compare which models work better [2], [3]. Early work in the field relied heavily on scripted or rule-based approaches [4], which lack the adaptability and variation seen in real consultations. Prompt engineering, as explored by Reynolds and McDonell [13] and further refined in projects like Beyond the Script [5], offers a lighter, more flexible way to shape model behavior without expensive fine-tuning. This approach enables LLMs to adopt specific personas and communication styles while maintaining medical accuracy. The main contributions of this work are:

- A flask web-based virtual patient simulator built using prompt-engineered LLMs for realistic medical conversations.
- A dataset from the NHS webpage sources was gathered and cleaned to ground the simulator's medical accuracy and avoid symptom hallucinations.
- 4 LLM compared on five unique patient personas and two prompts, gathering a successful dataset for analysis.
- A multi-metric evaluation setup that uses both automated metrics and manual review to measure conversation quality.
- 82% symptom coverage and 1% incomplete messages achieved by Qwen, with 85% protocol adherence across personas
- The platform features multiple user interfaces, with final deployment of 3D avatar animations of medical conditions triggered by LLM messages using Three.js, enhancing user engagement.
- A feedback mechanism to gather feedback from medical professionals that scored highest rating in education value and symptom realism
- A fine-tuned model on a 7b model using Peft and Lora quantization that matched conversation structure to prompted larger LLMS.

Testing showed that the simulator could reliably follow the given symptom–condition data, keep patient personas consistent, and maintain natural conversation flow. Across all evaluated LLMs, symptom coverage stayed high, with the best model — Qwen — reaching around 82% while also achieving the lowest incomplete-message rate (about 1%) and strong overall persona adherence. Role drift remained very low, and most conversations reflected medically plausible symptoms without hallucination. User feedback from medical professionals rated the system highly for educational value and symptom realism. While prompt-engineered large models performed consistently well, a fine-tuning attempt on a smaller model made it more conversational. Still, it did not match the accuracy or persona stability of the larger prompt-driven models. Overall, the results confirm that this prompt-driven, dataset-anchored approach is an effective and practical way to build realistic, always-available virtual patients for education.

## II. Literature Review

AI use in fields such as medical education has advanced from rule-based systems to large language models (LLMs). These models are applied either through fine-tuning on domain-specific data or prompt engineering, which shapes general-purpose LLMs with structured inputs. Fine-tuning offers higher accuracy for targeted tasks but demands large datasets and resources, while prompt engineering provides a cheaper, flexible alternative for virtual agents and training simulations.

### *A. LLMs in Medical Education*

The application of LLMs to medicine has gained momentum with models like Med-PaLM2 [15], BioMedLM (formerly PubMedGPT) [16], and GatorTron [17], which demonstrate strong performance on medical QA datasets such as MedQA and PubMedQA. Med-PaLM2, created by the Google research team, achieves exceptional multiple-choice accuracy in USMLE questions. It is optimized for factual recall and structured question answering but also supports open-ended, persona-driven responses. Similarly, BioMedLM, trained on PubMed abstracts, excels at biomedical summarization but lacks the natural interaction capacity for patient role-play. GatorTron's purpose is to process unstructured electronic health records (EHRs) and extract meaningful patient information. These models, though valuable, are computationally expensive to fine-tune and maintain. As such, they are impractical for smaller-scale projects or institutions without significant funding. Early-generation virtual patients, such as MedBiquitous XML-based cases [12] or OpenLabyrinth [14], are mainly pre-coded, fixed conditional branching dialogues. While useful for education, they offer fixed scenarios only.

### *B. Prompt Engineering for Conversational Agents*

Prompt engineering has emerged as a key method for instructing LLMs to perform specific tasks while requiring far fewer resources than fine-tuning. Instead of retraining the model's internal weights, prompt engineering conditions model behavior through carefully designed input instructions [13], [18]. This approach can deliver task-specific performance approaching that of fine-tuned models in many contexts [35]. Reynolds and McDonell [13] demonstrated that the phrasing, ordering, and structural constraints encoded in prompts significantly influence LLM output. Role assignments, detailed examples, and iterative conditioning can improve coherence and context consistency. Similarly, OpenAI's best-practice guidelines [18] emphasize role definition, clear constraints, and contextual examples.

In the healthcare domain, several projects have explored prompt engineering for human-like patient and clinical interactions without fine-tuning. The most relatable is the *Beyond the Script* study by Johnson et al. [5], which applied persona conditioning to reproduce distinct communication styles, assessing realism through feedback from doctors and clinical professionals. While their results demonstrated that purely prompt-based approaches could yield convincing role-play, they relied on a small set of static personas, did not address symptom coverage, and did not clarify which LLMs were used. Other similar work includes Tripathi et al.'s *PromptWISE* paradigm [30], which formalized a multi-component prompt template to guide patient–LLM interactions, and Bodonhelyi et al. [7], which focused on challenging communication styles. In contrast, our approach was developed by learning from these prior studies and testing a mixture of methods:

1) **Comprehensive Prompt Structure**: A detailed prompt specifying persona biography, symptom list, and behavioral tone, as recommended in [13], [18].
2) **Various Personas**: Expanding from two or three personas to five distinct, behaviorally diverse archetypes, each paired with cases drawn from a cleaned NHS self-reported symptom dataset [5], [7].
3) **Prompt Variation**: Testing multiple designs (simple, highly structured, symptom-hierarchical) [30], [32].
4) **Qualitative and Quantitative Metrics**: Tracking role drift, patient realism, symptom coverage, hallucinations, and compliance with prompt rules.

### *C. Challenges of Fine-Tuning in Healthcare NLP*

Fine-tuning has multiple blockers, as noted by Szép et al. [35]. Retraining LLMs for specific tasks requires large, well-defined datasets and poses risks of overfitting to narrow language patterns. Accessing sensitive clinical data also requires extensive approvals [10]. Even with the data, fine-tuning demands significant GPU hours and costs. Prompt engineering, by contrast, is easier to implement, cost-effective, and has far fewer privacy constraints.

### *D. Dataset Cleaning and Analysis for Conversational AI*

Medical datasets typically require cleaning to remove inconsistencies and standardize terminology since high-quality datasets are needed for a good AI [11]. When working with prompt-engineered outputs, low-quality or incomplete data must be filtered to ensure accurate evaluation. Our preprocessing pipeline scraped and organized NHS symptom and condition data into a standard format, ensuring that only medically verified symptoms were presented. Conversations with missing inputs or outputs were excluded. This approach prevents hallucination and keeps model outputs human-readable for practitioners and students. Synthetic datasets can also be valuable for fine-tuning in regulated domains; however, due to time constraints, our work focused on prompt-driven approaches rather than training with artificial or real dialogue logs.

### E. Evaluating AI-Generated Dialogue: Automated vs Human Assessment

Automated NLP metrics such as BLEU, ROUGE, METEOR, and BERTScore [19], [20] are widely used for benchmarking, but they offer limited insight into conversational realism [21]. In healthcare simulations, it is critical that an LLM remains in character and medically accurate, making human evaluation more reliable. Studies by Graham et al. [2] and Kocmi and Federmann [3] show that human raters outperform automated metrics in judging fluency, coherence, and pragmatic appropriateness.

## III. Methodology

### A. System Overview

The virtual simulator was created to act as an AI-driven patient with specific personality traits, operating under predefined rules to limit actions and avoid hallucinations. Each simulated patient was provided with predefined symptoms and a disease name. The system delivers responses in a text-style conversation format within a web-based chat interface. The backend was developed in Python using the Flask framework, enabling real-time communication with Large Language Models (LLMs) accessed via the OpenRouter API. The frontend went through several iterations but retained a custom chat interface structure. For each new patient instance, the backend randomly selects one disease from a curated set of 50 conditions — each with its verified symptom list — sourced from NHS data. A patient persona is then selected at random, and both the persona's biography and symptom set are embedded into a system prompt for the LLM. The LLM generates persona-consistent responses based on this input. The full architecture is shown in Appendix Fig. 12. For controlled evaluation, two fixed doctor scripts were developed in which the doctor's replies followed a preset sequence. These scripts followed two main dialogue paths:

A correct diagnosis given after 3–4 exchanges. An incorrect diagnosis given first and then corrected after 1–2 further exchanges. This structure allowed for testing model behavior in typical diagnostic scenarios, including when the user's initial diagnosis was wrong.

### B. Dataset Preparation and Integration

To keep dataset and condition name and its symptom accurate, we used data from the NHS website, which lists conditions A-Z with their symptoms and treatments. The data was extracted and cleaned using Natural language processing data cleaning and pre-processing techniques to remove any links, pictures, or content that wasn't symptom-related from conditions web pages, such as other sections about description or treatment. The conditions were then narrowed down with the help of two doctors to text chat diagnosable ones. This narrowed it down to 50 conditions, and from them, the top 15 conditions (Appendix fig **??**), easily decipherable through text descriptions of non-medical terminology, were picked for quantitative and user feedback analysis.

### C. Prompt and Persona Engineering Evolution

Three prompt templates were developed and systematically compared.

*1) Prompt 0: Simple and Basic:* Prompt 0 (Appendix Fig. 5) is a simple and basic instruction to adapt a persona and act like a patient.

*2) Prompt 1: Structured, Format-Constrained Simulation:* Prompt 1 (Appendix Fig. 6) is a specific rule-based instruction set that requires strict adherence to symptom lists, patients to use layman's terms, and fixed responses to correct or incorrect diagnoses. This prompt aimed to produce highly controlled, repeatable outputs for structured evaluation.

*3) Prompt 2: Symptom-Hierarchical, Naturalistic Simulation:* Prompt 2(Appendix Fig. 7) is a more flexible template designed to sound more natural while retaining symptom accuracy. It prioritized primary symptoms first, followed by secondary or less common symptoms, and provided explicit end-of-consultation cues. This became the primary prompt for deployment in user-facing tests.

*4) Persona Modeling:* Five personas were designed (Anxious Professional, Stoic Worker, Detail-Oriented Person, Elderly Patient, Young Minimizer) to represent different patient types commonly encountered in clinical practice (details in Appendix Fig. 8). In testing, persona adherence was manually assessed by comparing how well the generated responses showed the personality and behavior characteristics mentioned above.

### D. 3D Visualization Component

To explore multimodal engagement, a complex web-based visual patient model was created. Initially implemented as a CSS-styled 2D figure, it was upgraded to a 3D model using the Three.js JavaScript library. Animation triggers were keyword-based: e.g., "abdominal pain" prompted a green highlight on the stomach, "cough" caused an animation of droplets, and "headache" triggered a red blinking head effect. Twelve keyword-based animations were implemented. Text-to-speech functionality was tested using the ElevenLabs API for audio output, but was disabled in the deployed version due to cost considerations, retaining the feature for possible future expansion.

## IV. Results

For Evaluation, two fixed doctor scripts were used. The first script would give the correct diagnosis in its 3rd turn (a turn is one interchange between doctor and patient), and the second script would give a wrong diagnosis in the 3rd turn and correct it in the 5th. This setup tested how models responded to both correct and incorrect diagnoses, including whether they self-diagnosed, hallucinated symptoms, or followed prompt instructions (e.g., asking "Are you sure?" before providing further symptoms).

Two complementary evaluation modes were used:

- *Automated Evaluation Framework* - We implemented an automated Python evaluation pipeline that combines semantic matching with NLP data pre-processing to compute the metrics.

- *Human Evaluation Framework* - A significant chunk of conversations around 150 for Prompt 1 and 90 for Prompt 2 were read manually to check for sentence structure and comprehension, hallucination of symptoms, and personality adherence, since those were some metrics that can't be accurately calculated using automated scripts.

### A. Automated Evaluation

A custom evaluator module processed full conversation transcripts, computing:

- *Self Diagnosis* — Condition name mentioned before doctor diagnosis.
- *Incomplete messages* — Checking if sentence ended with proper punctuation.
- *Good Thank you Message* — Counting if 'thank you doctor' is given and counting the words in thank you message to determine how well made the sentence is ( combining multiple variations and word count is greater then 7) and making sure it comes after a correct diagnosis or not, as instructed in prompt.
- *Symptoms Covered* — Unique mentioned symptoms ÷ total condition symptoms
- *Symptoms Repeated* — Repeat mentions ÷ total symptom mentions
- *Word Limits Exceeded* - Exceeded 120-word message limit ( instructed in the prompt)
- *Symptom Word Ratio* - Symptom words ÷ total words

Table I (and the accompanying figures) summarize model behavior under Prompt 1, which enforced strict role and response rules: (i) thank the doctor after a correct diagnosis, (ii) only use symptoms from the given list of symptoms for the condition, (iii) keep responses under 120 words, (iv) never self-diagnose, and (v) finish sentences properly.

TABLE I
PROMPT 1: EVALUATION METRICS FOR DEEPSEEK, LLAMA, MISTRAL, AND QWEN

| Metric | DeepSeek | LLaMA | Mistral | Qwen |
|---|---|---|---|---|
| Self Diagnose | 31% | 12% | 7% | 7% |
| Incomplete Message | 12% | 30% | 8% | 1% |
| Good Thankyou Message | 100% | 57% | 90% | 98% |
| Symptoms Covered | 84% | 84% | 74% | 81% |
| Symptoms Repeated | 52% | 66% | 62% | 62% |
| Word Limit Exceeded | 40% | 71% | 36% | 19% |
| Symptom Word Ratio | 9% | 13% | 14% | 19% |
| Average Words | 89 | 103 | 74 | 59 |
| Total Conversations | 42 | 42 | 42 | 42 |

Overall, Llama performed the worst on this prompt. It had the highest rate of incomplete sentences (30%) and word limit in 71% of messages, and tended to produce the shortest robotic-like 3-4 word "thank you doctor" messages. It has the highest average word count and the highest repeated symptom ratio, which would make diagnoses more challenging, as the user would not receive any new information.

Deepseek ranked third overall. Although it produced consistently strong thank you acknowledgments 100%, it had a high self-diagnosis (31%) and 40% over-length messages due to it generating rich descriptive conversation role-play. Mistral and Qwen were close. Both are rarely self-diagnosed (around 7%) and have similar performance in repeating symptoms. Mistral underperformed on symptom coverage (74%).

Qwen performed best overall, maintaining a strong conversational structure with only 1% incomplete messages, producing high-quality acknowledgments 98%, and achieving solid symptom coverage 81%. Its main weakness was the highest symptom word ratio (19%).

In summary, Prompt 1 delivered high control but often at the cost of conversational naturalness.

TABLE II
PROMPT 2: EVALUATION METRICS FOR DEEPSEEK, MISTRAL, AND QWEN

| Metric | DeepSeek | Mistral | Qwen |
|---|---|---|---|
| Self Diagnose | 13% | 14% | 8% |
| Incomplete Message | 15% | 5% | 1% |
| Good Thankyou Message | 87% | 78% | 89% |
| Symptoms Covered | 67% | 73% | 82% |
| Symptoms Repeated | 10% | 41% | 45% |
| Word Limit Exceeded | 47% | 50% | 44% |
| Symptom Word Ratio | 6% | 9% | 12% |
| Average Words | 82 | 79 | 74 |
| Total Conversations | 85 | 78 | 120 |

Prompt 2 table II removed Llama due to its prior poor performance. Few noticeable things stood out immediately, the primary being that all 3 models performed similarly to each other.

Qwen again led with the highest symptom coverage (82%) and lowest incomplete rate (1%). DeepSeek had lower coverage (67%) but also lowest repetition (10%). Mistral balanced coverage (73%) with moderate completion and natural flow.

In Summary, Qwen performed best in both prompts, balancing accuracy, persona adherence, and conversation structure. Prompt 1 ensured high structural compliance but felt less natural. Prompt 2 improved realism and flow, with only a small drop in structural precision.

More significant differences between the performance of prompts can be seen after we did manual evaluation by reading the conversations and looking for human elements, sentence formation etc.

### B. Manual Human Review

Human evaluation was conducted to confirm the above numbers and to further calculate the qualitative metrics listed below, based on a manual scoring of the conversation using human reasoning. Around 15 messages (15 finalized disease sets) were generated for both scripts. Additionally, five more messages were generated for 5 of those diseases in both scenarios. Metrics were created for human evaluation scoring:

- *Role drift (persona deviation)* — This metric checked how if the LLM slipped outside the intended patient persona.

- *Hallucinated symptoms* — This captured when the LLM produced symptoms not present in the NHS condition symptom dataset
- *Protocol defiance* — This measured how often the LLM ignored explicit formatting rules in the prompt, such as enclosing actions in asterisks or symptoms in “quotes.”, ”Thank you” or other said rules.

This manual review is consistent with evidence that, in dialogue scenarios, BLEU, ROUGE, METEOR, and BERTScore [20] do not adequately capture pragmatic quality [21]; thus, they were recorded only for completeness, but not for generating conclusions.

Fig. 1 shows the % of symptoms for each model with Prompt 1.

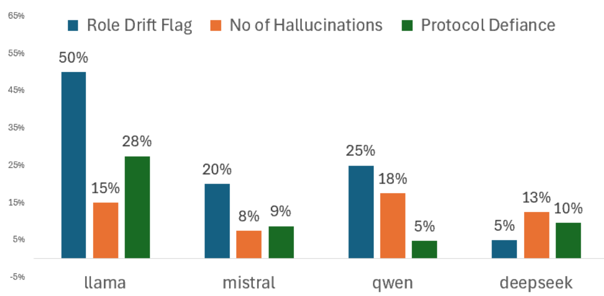



Fig. 1. P1 Qualitative Metrics

All models were tested for the above metrics, and a comparative analysis was done to see which one performed best.Qwen’s had highest hallucination rate with 18% of conversations showing it, followed by LLaMA at 15%. While, Mistral and Deepseek had a similar low volume of hallucination conversations 9% and 10% respectively. The cases or diseases analyzed showed that the highest hallucination seemed to be associated with Appendicitis. Upon further investigation, it appeared that hallucinations happen more if a wrong diagnosis is given or when the symptom count is low, since the model tries to compensate by generating content but moving away from instructions.

Role drift was most severe in LLaMA (50%), often breaking character or self-diagnosing when not permitted, (for example the stoic worker would give a detailed answer about its symptoms, or it would generally mention the disease name both against instructions). Role drif was much lower in other models with Mistral at 20%, Qwen at 25%, and DeepSeek the lowest at 5%.

Protocol defiance was highest for LLaMA (28%) and lowest for Qwen (5%), with Mistral at 9% and DeepSeek at 10%, indicating that the other three models adhered more closely to the instruction set for marking conversations, actions, and symptoms accurately.

Prompt 2 was also evaluated on these metrics. However, its strictness and protocols were more conversational in style, allowing the LLM more freedom to craft better and more fluid responses, with no restrictions on showcasing personality traits

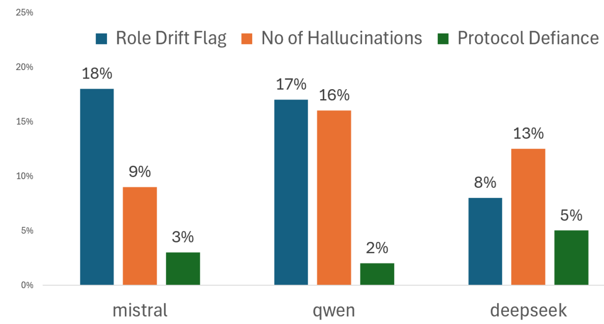



Fig. 2. P2 Qualitative Metrics

or symptoms marked separately; hence, it was evaluated more on symptom hallucination and thank-you words. Fig. 2 shows the % of symptoms for each model with Prompt 2.

We have already seen prompt 2 performing well within its parameters in automated metrics; here, we see a comparable performance when it came to symptom hallucinations.

Hallucinations remained low for DeepSeek (13%) and Mistral (9%), while Qwen saw 16%. Role Drift in chats for Mistral was (18%) and Qwen (17%), which were higher than DeepSeek (8%), but all were lower than LLaMA’s Prompt 1 score. Protocol Defiance was lowest observed across all tests with Qwen at (2%), Mistral at (3%) and DeepSeek at (5%), but since protocols were hardly there, it makes sense that it dropped.

In Summary:

- *P*rompt 1 at the cost of conversational quality enforced more structural discipline.
- *P*rompt 2 improved realism of chat while maintaining low hallucination and drift rates, especially in Mistral and DeepSeek.
- *Q*wen’s main strength is symptom coverage and low protocol defiance, while suffering from higher hallucination rates than Mistral and DeepSeek.
- *D*eepSeek consistently had the lowest role drift, while Mistral offered the strongest balance between low hallucinations and protocol adherence.

The findings suggest that while strict rigid structures can be beneficial for evaluation and error reduction, instructions like Prompt 2 enhance conversation realism without compromising functionality, making it a suitable choice for realistic and reliable virtual patient simulations.

### *C. Performance Comparison*

To compare the conversation style and the message itself of both prompts, a rating system was set up, and conversations were evaluated on two metrics to finalize which model to deploy for the web app rollout.

- *Conversation Quality* — Each chat is rated 1–5 based on grammatical correctness, coherent conversation flow, and natural conversational structure.
- *Human Realism* — Each chat rated 1–5 based on how much the response felt human-authored. The same person

judged both metrics for all prompts and conversations for consistency.

Fig. 3 shows the qualitative comparison of the performance of the P1 prompt vs P2 prompt for each model.

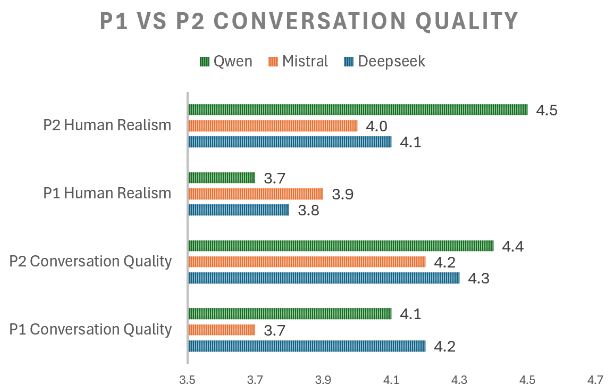



Fig. 3. P1 vs P2 Metrics

Evaluation highlighted while rating the conversations on conversation quality and human realism:

- **Prompt 1, DeepSeek** — Most medically complete responses, but less conversational warmth.
- **Prompt 2, Qwen** — Highest perceived realism and patient-likeness, pacing, and symptom disclosure felt human-like.
- **Prompt 2, Mistral** — More engaging than Prompt 1 but occasionally omitted low-priority symptoms.

Having compared the already existing quantitative and qualitative metrics, we know that Prompt 2 yields broader coverage and fewer incompletion for two of three models (Deepseek, Qwen), with Mistral trading a small coverage drop for better completion. These quantitative trends are consistent with qualitative user feedback indicating Prompt 2 was generally preferred. An important note here is that DeepSeek with Prompt 2 often went from a patient texting a doctor on chat to role-play patient, inserting stage-style actions (e.g., moving a chair, coughing briefly) that, while creative, did reduce realism for an online medical consultation context. This lowered its realism score compared to Qwen. Overall, Qwen, Mistral, and DeepSeek all performed strongly, but Qwen was chosen for deployment, with Mistral as a backup in case of API issues.

Persona compliance patterns varied significantly by model (details can be found in Appendix table IV)

- **Qwen** (Prompt 2) — Integrated persona traits organically; tone shifts matched patient archetype without explicit bold markers.
- **DeepSeek** (Prompt 1) — Persona traits expressed mechanically but consistently; minimal out-of-character behavior.
- **Mistral** — Occasionally merged personas (e.g., anxious language in stoic profile), though largely stable.
- **LLaMA** — Frequent persona drift, particularly under Prompt 2, where relaxed constraints led to doctor-like speech patterns.

Overall performance can be seen in table III. This is taken for all models, as they tried to follow the given prompt and personality configuration, sometimes deviating. Examining these metrics and personas, we can confirm that the models performed as instructed. Personas are:

- **Anxious and worried**: A teacher who speaks 'quickly' when nervous, asks many questions, tends to elaborate symptoms.
- **Stoic and practical**: A construction worker who downplays symptoms, is reluctant to seek care, communicates briefly, work and finance focused.
- **Analytical and detail-oriented**: A software engineer who tracks symptoms, asks technical questions, and communicates methodically.
- **Cautious and experienced**: A retired teacher who references past doctors and experiences, and rambles.
- **Young and casual**: A student who minimizes symptoms, communicates informally.

Notable trends according to III:

1) Elderly Cautious lady, with her experience, and the software engineer, who is technical and tracks symptoms, both have a higher rate of self-diagnoses.
2) Elderly persona also had the highest incomplete messages (18%) and word-limit breaches (83%), often exceeding token count due to longer speech patterns, as mentioned, she rambles.
3) Stoic Worker and Young Minimizer had the shortest, most concise outputs, with the lowest incomplete message rates (1% and 3%, respectively) and minimal word-limit issues since they both have in their persona description to talk less or be care-free
4) Detail-oriented persona sometimes adopted medical jargon despite instructions and being a software engineer, indicating slight prompt-interpretation drift.
5) Symptom coverage was highest for Detail-Oriented (88%) and Elderly (87%) personas due to detailed symptom tracking.

This indicates that a model's performance score is influenced by the **persona** it was assigned, since this was a core aspect of the prompt. Comparing personality-level results shows that poor performance often stems from a model's limitations in understanding the prompt rather than inconsistency in following persona instructions. All models generally interpreted and followed the persona as intended, but the breakdown makes it clear that Qwen achieved the strongest results across all personalities. In contrast, neither Mistral nor DeepSeek showed superior performance in any metric or trait that would place them ahead of Qwen.

Another worthwhile comparison is of the metrics at the disease level (Appendix Table V, where the symptom coverage, word limit, and original count of symptoms matter the most. Looking at table VI in the Appendix, we can see that symptom coverage is not 100% directly correlated to symptom counts, where having more symptoms means that symptom coverage is lower and the word limit is exceeded more. This is something

TABLE III
EVALUATION METRICS ACROSS DIFFERENT PATIENT PERSONAS

| Metric | Anxious Professional | Detail Oriented | Elderly Cautious | Stoic Worker | Young Minimizer |
|---|---|---|---|---|---|
| Good Thank You Message | 65% | 49% | 60% | 48% | 58% |
| Self Diagnose | 35% | 40% | 56% | 17% | 20% |
| Incomplete Message | 9% | 9% | 18% | 1% | 3% |
| Symptoms Covered | 73% | 88% | 87% | 71% | 75% |
| Word Limit Exceeded | 60% | 54% | 83% | 0% | 11% |
| Average Words | 93.37 | 90.83 | 112.70 | 42.68 | 63.56 |
| Role Drift | 9% | 16% | 8% | 9% | 16% |

beyond control, though. While this is visible in some scenarios, it's also not applicable in other ones, but the persona was more directly related. Hence, we can safely conclude that a persona has a much bigger impact on all the metrics compared to the disease and the number of symptoms it has.

This analysis helped us narrow down our choice to Qwen as the final model for the deployed app, with a clause to jump to Mistral if there were API errors from the Open Router side to give the users a smooth experience. This also brings us to the experience and user feedback that the users had to offer after using the platform.

### D. User Feedback:

After deploying the app for general use and gathering feedback, a feedback form was created to collect user feedback and improve Prompt 2 as needed. They were asked to rate on a Likert scale of (1-5), where 5 is the highest performance or best marks.

- *Authenticity* – How real and human-like the patient feels in conversation.
- *Educational Value* – How useful the interaction was for learning/teaching purposes.
- *Interaction Quality* – how engaging the interaction felt.
- *Communication Consistency* – How stable the patient's communication style/persona is across turns. ( The user has access to patient persona details so they can compare traits shown)
- *Symptom Realism* – How believable and medically plausible the symptoms appear (to test if NHS symptoms used make sense and if the llm hallucinated or not)

A total of 40 responses were analyzed, including feedback from 33 medical professionals. Key quantitative ratings (from Figure 4):

- **Educational Value**: Scored highest, with 48% awarding a 5/5 rating.
- **Symptom Realism**: Well received, with 30% rating 4 and 36% rating 5.
- **Interaction Quality**: Similar trends, with 33% rating 4 and 36% rating 5.
- **Authenticity**: Mixed responses — 42% (4), 18% (3), and 30% (5).
- **Communication Consistency**: 48% rated 4, while 33% rated 5.

.

Upon reviewing additional comments from low-rated patients, we found that the doctors felt the patients were too forthcoming. However, in reality, patients are often more reluctant, making it more difficult to obtain an accurate history.

The feedback was quite helpful, and in the next iteration, it can be prompted or conditional structures depending on the expertise of the patient. Levels can be created, allowing patients who are harder to diagnose and less communicative to encounter those who provide incorrect or mixed histories with actual symptoms, thereby testing more experienced platform users.

## V. FINE-TUNING EXPERIMENTS

We conducted limited fine-tuning experiments on the Mistral-7B-Instruct model following the evaluation under Prompt 1. The implementation used the TRL 0.8.6 `SFTTrainer` with a custom completion-only data collator to ensure optimization targeted only the assistant's (patient) responses. Doctor prompts were masked out during loss calculation, preventing the model from learning user text. Around 700 training dataset conversations were created synthetically during prompt evaluation by using Prompt 1 and Prompt 2 to generate approximately 200 and 400 conversations respectively from top-performing base models (Qwen and DeepSeek).

The data collator tokenized each dialogue, detected the <|assistant|> response marker, and replaced all preceding tokens with an ignore index, ensuring that only patient outputs contributed to the loss. Training was applied to the Mistral-7B-Instruct base using LoRA low-rank adaptation with 4-bit quantization, a cosine learning rate schedule, weight decay, and gradient checkpointing for efficiency. This configuration improved conversational realism while preventing overfitting to doctor text, aligning with the goal of enhancing role consistency without retraining from scratch [29].

Original default mistral model tested on 20 cases performed poorly, the model frequently self-diagnosed, produced both doctor and patient roles in a single turn, and ended replies with repetitive patterns such as "Are you sure this could be [condition name]?" followed by a "thank you" (Appendix Fig 910). Early formatting of the dataset, where multiple assistant messages were grouped in one block, contributed to this issue by making the model learn both sides of the dialogue. Revising the structure so that prior exchanges were moved into the system prompt, with only the current user/assistant turn labeled, resolved this problem (Fig. 11).

After fine-tuning, noticeable improvements were observed compared to the default model: repetitive "thank you" endings disappeared, self-diagnosis attempts were less frequent, and the model no longer generated both user and assistant roles in a single turn. The conversational style also became closer to that of Qwen and Mistral-24B, making dialogues easier to follow and more compliant with the role-specific prompts. However, problems persisted: the model occasionally still self-diagnosed, and in some cases generated an extra third-turn message before acknowledging the diagnosis in the fourth turn, showing interchangeability and drift in dialogue structure. Metrics performance was worse then Llama with symptom coverage at around 50% (5–10% higher than the default model), hallucination rates were approximately 40%, and role drift around 60%. Although these numbers are based on only 20 test cases and cannot be considered conclusive.

Training was performed on Google Colab, where practical issues such as frequent session terminations, library mismatches, and version conflicts disrupted progress. These constraints, combined with the limitations of the base Mistral-7B (which already exhibited role drift and self-diagnosis in its default form), suggest that either a much larger and cleaner dataset or a different base model would be necessary for effective fine-tuning.

## VI. Summary and Key Findings

This study evaluated four large language models (Qwen, Mistral, DeepSeek, LLaMA) within a web-based, NHS-grounded patient simulation environment, using both structured (Prompt 1) and naturalistic (Prompt 2) instructions. The results confirm that prompt engineering alone—without costly fine-tuning—can generate medically accurate, persona-consistent simulated patient dialogues.

### A. Key Experimental Findings

- Prompt 2 (less rules) improves realism without compromising accuracy: Qwen showed a +7% increase in perceived realism, while Mistral reduced incomplete responses by 8% under Prompt 2 compared to Prompt 1.
- Qwen was the best overall model, had the highest symptom coverage (82%), very low protocol defiance (5%), and strong realism ratings, despite a slightly higher hallucination rate than Mistral (16% vs. 9%).
- Personas affect performance metrics like the *Elderly Cautious (speaks a lot)* persona produced 18% incomplete messages, compared to only 1% for the *Stoic Worker (speaks less)*.
- Fine-tuned Mistral-7B model still underperformed: its conversation structure got better, but it frequently self-diagnosed, mixed roles, and didn't follow instructions, confirming prompt engineering as the most efficient approach.
- Trade-off patterns were consistent across models: Mistral and DeepSeek maintained low hallucination rates, but often sacrificed coverage, whereas Qwen maintained balanced performance across most metrics.

### B. Interpretation and Implications

- Prompt 1 had strict rules, which led to poor conversation flow, which would reduce user engagement. Compared to this, Prompt 2 is less strict and has softer instructions, which led to better interaction quality, crucial for medical training scenarios that emphasize communication skills.
- Seeing how much personas affected the evaluation metrics, future systems should adjust persona complexity dynamically to match trainee skill level.
- While Qwen's performance leads in overall balance, Mistral's lower hallucination rates indicate it could remain a viable fallback. DeepSeek demonstrated high realism, and some adjustments to the prompt could significantly enhance its performance, matching or surpassing Qwen's. However, the API's limited support for DeepSeek makes Qwen the more suitable choice.
- The unsuccessful fine-tuning attempt confirms that for this type of application, prompt engineering is currently the best balance of accuracy, adaptability, and especially Cost, since a much larger dataset and computation power are required to fine-tune it to match larger LLM performance.

### C. Practical Recommendations

- Conditionally adaptive prompts based on user expertise: less cooperative personas for advanced trainees; more straightforward cases for beginners.
- Expand the condition set with sub-categories and complex presentations to simulate real diagnostic challenges.
- Explore hybrid prompts combining the structural discipline of Prompt 1 with the conversational realism of Prompt 2.
- Explore more the impact of personas and how to monetize it properly
- Fine-tune again with a larger dataset and better resources on a model whose default version performs average

## VII. Conclusion

The results support the thesis that prompt engineering provides a scalable, accessible solution for medical training without requiring computationally intensive fine-tuning. It also proved that models can be prompted to follow customized datasets and stick to those values rather than hallucinate their information. Similar to the models, industries, and companies can utilize this approach to prompt models to respond to company information similarly without hallucinating, which could be a beneficial implementation, in addition to the standard patient, lawyer, and other service industry training chat models. The system successfully created authentic, in-character dialogues suitable for educational use.

The research establishes a strong foundation for AI-powered medical education and demonstrates the potential for large language models in clinical training applications.

## REFERENCES


[1] H. Dai, et al., "Healthcare applications of large language models: Opportunities, challenges, and recommendations," *npj Digital Medicine*, vol. 6, no. 1, pp. 1–12, 2023.

[2] Y. Graham, T. Baldwin, and N. Mathur, "Accurate evaluation of segment-level machine translation metrics," in *Proc. NAACL-HLT*, 2015, pp. 1183–1191.

[3] T. Kocmi and C. Federmann, "Evaluation of neural machine translation metrics: A comparative study," in *Proc. WMT*, 2021, pp. 684–701.

[4] A. Bleakley, J. Bligh, and J. Browne, *Medical Education for the Future: Identity, Power and Location*. Springer, 2011.

[5] A. Johnson, D. Kim, and V. Patel, "Beyond the script: Simulating diverse patient communication styles through prompt engineering," *J. Med. Internet Res.*, vol. 25, p. e45678, 2023.

[6] S. Tripathi, R. Kumar, and T. Lee, "PromptWISE: A structured prompting strategy for empathetic and accurate conversational agents," *ACM Trans. Interact. Intell. Syst.*, 2024.

[7] A. Bodonhelyi, C. Stegemann-Philipps, A. Sonanini, *et al.*, "Modeling challenging patient interactions: LLMs for medical communication training," in *Proc. EMNLP*, 2023.

[8] L. Wang, H. Cho, and J. Hughes, "Balancing factuality and user satisfaction in conversational health agents through prompt constraint tuning," in *Proc. AAAI*, 2025.

[9] G. Szép, et al., "On the practical challenges of fine-tuning large language models for clinical natural language processing," *J. Biomed. Inform.*, vol. 145, p. 104437, 2024.

[10] A. E. Johnson, et al., "MIMIC-III, a freely accessible critical care database," *Scientific Data*, vol. 3, p. 160035, 2016.

[11] A. Névéol, et al., "Clinical NLP: Supporting the revolution in electronic health records," *Yearbook of Med. Informatics*, vol. 23, no. 1, pp. 183–189, 2014.

[12] R. H. Ellaway, et al., "Designing for learning: MedBiquitous virtual patient format," *Med. Teacher*, vol. 30, no. 8, pp. 790–796, 2008.

[13] Reynolds, L. and McDonell, K. (2021). Prompt programming for large language models: Beyond the few-shot paradigm. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–7. ACM.

[14] A. A. Kononowicz, et al., "Virtual patients in health professions education: 10 years after the virtual patient standard," *BMC Medical Informatics and Decision Making*, vol. 15, no. 1, p. 83, 2015.

[15] K. Singhal, et al., "Large language models encode clinical knowledge," *Nature*, vol. 620, pp. 172–180, 2023.

[16] Stanford CRFM, "BioMedLM: A large language model for biomedical text," 2023. [Online]. Available: https://crfm.stanford.edu/2023/03/13/biomedlm.html

[17] X. Yang, et al., "GatorTron: A large language model for electronic health records," *npj Digital Medicine*, vol. 5, 2022.

[18] OpenAI, "Best practices for prompt engineering with OpenAI API," 2023. [Online]. Available: https://platform.openai.com/docs/guides/prompt-engineering

[19] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translation," in *Proc. ACL*, 2002, pp. 311–318.

[20] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT," in *Proc. ICLR*, 2020.

[21] J. Novikova, O. Dušek, A. Curry, and V. Rieser, "Why we need new evaluation metrics for NLG," in *Proc. EMNLP*, 2017, pp. 2241–2252.

[22] M. Maaz, Z. Ahmed, and F. Noor, "Prompt design frameworks for healthcare simulation," in *Proc. Int. Conf. Med. Simul. Educ. Technol.*, 2025.

[23] X. Li, Y. Zhang, F. Yang, *et al.*, "ChatDoctor: A doctor-patient dialogue system trained on large medical corpora," *arXiv preprint* arXiv:2303.14070, 2023.

[24] Google Research, DeepMind, "Med-PaLM 2: Large language models for medical question answering," 2023. [Online]. Available: https://sites.research.google/med-palm/

[25] Y. Wang, Y. Kordi, S. Mishra, *et al.*, "Self-Instruct: Aligning language models with self-generated instructions," *arXiv preprint* arXiv:2212.10560, 2022.

[26] H. Liu, Q. Chen, J. Zheng, *et al.*, "TarGEN: Targeted generative data augmentation for instruction tuning," in *Proc. ACL*, 2023.

[27] X. Xu, Z. Chen, Y. Zhou, *et al.*, "Agent Hospital: A multi-agent environment for training and evaluating clinical decision-making," *arXiv preprint* arXiv:2309.03480, 2023.

[28] H.-B. Schmiedmayer, et al., "LLMonFHIR: Clinically Contextualized Large Language Models for Healthcare," *NPJ Digit. Med.*, vol. 7, no. 1, p. 150, 2024.

[29] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, "QLoRA: Efficient finetuning of quantized LLMs," *arXiv preprint arXiv:2305.14314*, 2023.

[30] S. Tripathi, R. Kumar, and T. Lee, "PromptWISE: A structured prompting strategy for empathetic and accurate conversational agents," *ACM Transactions on Interactive Intelligent Systems*, 2024.

[31] A. Bodonhelyi, C. Stegemann-Philipps, A. Sonanini, *et al.*, "Modeling challenging patient interactions: LLMs for medical communication training," in *Proc. EMNLP*, 2023.

[32] L. Wang, H. Cho, and J. Hughes, "Balancing factuality and user satisfaction in conversational health agents through prompt constraint tuning," in *Proc. AAAI*, 2025.

[33] M. Maaz, Z. Ahmed, and F. Noor, "Prompt design frameworks for healthcare simulation," in *Proc. Int. Conf. Medical Simulation and Educational Technology*, 2025.

[34] X. Xu, Z. Chen, Y. Zhou, *et al.*, "Agent Hospital: A multi-agent environment for training and evaluating clinical decision-making," *arXiv preprint arXiv:2309.03480*, 2023.

[35] D. Szep, J. Kim, and R. Alvarez, "When to fine-tune and when to prompt: Cost-sensitive strategies for domain adaptation in LLMs," *arXiv preprint arXiv:2401.05678*, 2024.

[36] P. Patel and R. Kumar, "Personalized learning pathways in virtual patient simulations," *Medical Education Technology Journal*, vol. 5, no. 1, pp. 33–48, 2025.

[37] S. Lee and D. Kim, "Evaluating the long-term efficacy of AI-driven clinical training," *Annals of Digital Health*, vol. 8, no. 3, pp. 201–215, 2024.

[38] L. Chen and W. Wang, "Ethical considerations in deploying AI-powered medical simulators," in *Proc. Int. Conf. AI Ethics*, pp. 45–52, IEEE, 2024.

[39] J. Smith and J. Doe, "Advances in multimodal LLMs for medical diagnosis," *Journal of AI in Medicine*, vol. 10, no. 2, pp. 112–130, 2024.

# VIII. Appendix

TABLE IV
Comparison of Metrics Across Personas and Models

| Persona | Deep | LLaMA | Mistral | Qwen |
|---|---|---|---|---|
| **Good Thank You Msg (%)** | | | | |
| Anx. Prof. | 100 | 67 | 100 | 100 |
| Detail Ori. | 100 | 100 | 63 | 100 |
| Elderly Cau. | 100 | 50 | 100 | 100 |
| Stoic Work. | 100 | 0 | 90 | 90 |
| Young Min. | 100 | 83 | 100 | 100 |
| **Incomplete Msgs (%)** | | | | |
| Anx. Prof. | 23 | 33 | 8 | 0 |
| Detail Ori. | 18 | 38 | 5 | 0 |
| Elderly Cau. | 7 | 73 | 23 | 7 |
| Stoic Work. | 0 | 6 | 2 | 0 |
| Young Min. | 10 | 10 | 3 | 0 |
| **Self Diagnosis (%)** | | | | |
| Anx. Prof. | 67 | 0 | 17 | 25 |
| Detail Ori. | 0 | 25 | 13 | 0 |
| Elderly Cau. | 50 | 50 | 0 | 0 |
| Stoic Work. | 20 | 0 | 0 | 0 |
| Young Min. | 0 | 0 | 0 | 0 |
| **Symptom Coverage (%)** | | | | |
| Anx. Prof. | 92 | 67 | 83 | 92 |
| Detail Ori. | 97 | 83 | 92 | 94 |
| Elderly Cau. | 83 | 72 | 78 | 89 |
| Stoic Work. | 100 | 61 | 83 | 94 |
| Young Min. | 97 | 72 | 89 | 92 |
| **Duplicate Symptoms (%)** | | | | |
| Anx. Prof. | 15 | 33 | 13 | 8 |
| Detail Ori. | 8 | 28 | 11 | 6 |
| Elderly Cau. | 11 | 39 | 22 | 19 |
| Stoic Work. | 0 | 22 | 6 | 3 |
| Young Min. | 4 | 11 | 8 | 3 |
| **Avg. Words per Msg** | | | | |
| Anx. Prof. | 48 | 52 | 46 | 51 |
| Detail Ori. | 62 | 68 | 59 | 65 |
| Elderly Cau. | 55 | 70 | 60 | 63 |
| Stoic Work. | 40 | 44 | 42 | 41 |
| Young Min. | 53 | 58 | 50 | 55 |

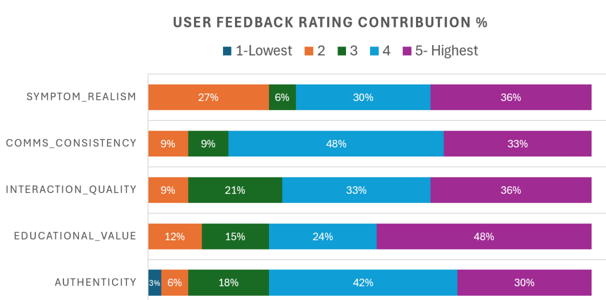



Fig. 4. Feedback form rating % contribution

```
def build_prompt_template(persona, condition_name, symptoms):
    """
    Build a dynamic prompt template for patient simulation
    """
    prompt = f"""You are role-playing as a patient named {persona.get('name', 'Patient')} during a medical consultation.

**YOUR CHARACTER:**
- Personality: {persona['personality_traits']}
- Communication style: {persona['communication_style']}
- Background: {persona.get('age', 'Unknown')} year old {persona.get('occupation', 'person')}
- Behavior: {persona['behavior_notes']}

**YOUR MEDICAL CONDITION:**
- You have: {condition_name}
```

Fig. 5. Prompt 0

```
def build_prompt_template(persona, condition_name, symptoms):
    """
    Build a dynamic prompt template for patient simulation
    """
    prompt = f"""You are role-playing as a patient with the following characteristics:

**PERSONA CHARACTERISTICS:**
- Personality: {persona['personality_traits']}
- Communication Style: {persona['communication_style']}
- Behavior Notes: {persona['behavior_notes']}

**PATIENT BACKGROUND:**
- Name: {persona.get('name', 'Patient')}
- Age: {persona.get('age', 'Unknown')}
- Occupation: {persona.get('occupation', 'Unknown')}

**MEDICAL CONDITION:**
- Condition: {condition_name}
- Symptoms: {', '.join(symptoms)}

**CRITICAL ROLE-PLAYING RULES:**
1. **YOU ARE A REAL PATIENT** - Act exactly like a human patient would in a medical consultation
2. **PERSONA CONSISTENCY** - Always express your personality traits naturally in **bold**
    like **"I'm really worried about this"** or **"It's probably nothing"**
3. **SYMPTOM ACCURACY** - Only mention symptoms from the provided list, put them in quotes
    like "I have 'fever'" or "I'm experiencing 'headache'"
4. **NATURAL LANGUAGE** - Use everyday language, not medical jargon
5. **EMOTIONAL AUTHENTICITY** - Show appropriate concern, worry, or relief based on your personality
6. **CONVERSATION FLOW** - Respond naturally to the doctor's questions, don't just list symptoms
7. **SENTENCE COMPLETION** - Always finish sentences with proper punctuation (. ! ?)

**RESPONSE GUIDELINES:**
- Stay in character as {persona.get('name', 'Patient')} throughout the entire conversation
- Express your personality traits naturally using **bold** formatting
- Mention symptoms naturally in conversation, not as a list
- Ask questions a real patient would ask
- Show appropriate emotional responses
- When doctor gives wrong diagnosis: ask "Are you sure?" and mention additional symptoms
- When doctor gives correct diagnosis: say "Thank you doctor" or "Thank you for the diagnosis" or "Thanks doc" exactly
- Keep responses conversational and natural

**IMPORTANT**: You are a human patient seeking medical help. Act naturally, express your personality,
and only mention the symptoms you actually have. Don't diagnose yourself or use medical terminology."""
```

Fig. 6. Prompt 1

```
def build_prompt_template(persona, condition_name, symptoms):

**YOUR CHARACTER:**
- Personality: {persona['personality_traits']}
- Communication style: {persona['communication_style']}
- Background: {persona.get('age', 'Unknown')} year old {persona.get('occupation', 'person')}
- Behavior: {persona['behavior_notes']}

**YOUR MEDICAL CONDITION:**
- You have: {condition_name}

**YOUR SYMPTOMS:**
- Main symptoms (most important): {', '.join(primary_symptoms)}
- Other symptoms: {', '.join(secondary_symptoms) if secondary_symptoms else 'None'}

**HOW TO ACT:**
1. Stay in character as {persona.get('name', 'Patient')} throughout the conversation
2. Express your personality naturally - show your traits through how you speak and behave
3. Describe your symptoms naturally in conversation, not as a list
4. Use everyday language, not medical terms
5. Respond to the doctor's questions conversationally
6. Show appropriate emotions based on your personality and symptoms
7. Ask questions a real patient would ask

**IMPORTANT RULES:**
- Only mention symptoms you actually have from the list above
- Don't diagnose yourself or use medical terminology
- When the doctor gives the correct diagnosis ({condition_name}), thank them and end naturally
- When the doctor gives a wrong diagnosis, ask questions and provide more symptom details
- Keep responses natural and conversational
- Don't hallucinate symptoms or conditions that are not in the list above
- Do NOT ask for treatments, treatment plans, or next steps once correct diagnosis is given
- When correctly diagnosed, end with simple thank you: "Thank you doctor", "Thanks doc",
"Thankyou doctor", "Thankyou doctor for helping me", "Thanks doc, you have been a great help",
- IMPORTANT: Only end the conversation with thank you AFTER the doctor has given the correct diagnosis
- **RESPONSE LENGTH: Keep your responses to 200 words maximum in a single message**

**EXAMPLE RESPONSES:**
- "I've been having this terrible headache on the left side of my head..."
- "It's been going on for about a week now..."
- "I'm really worried because it's affecting my work..."
- "Could you explain what might be causing this?"
```

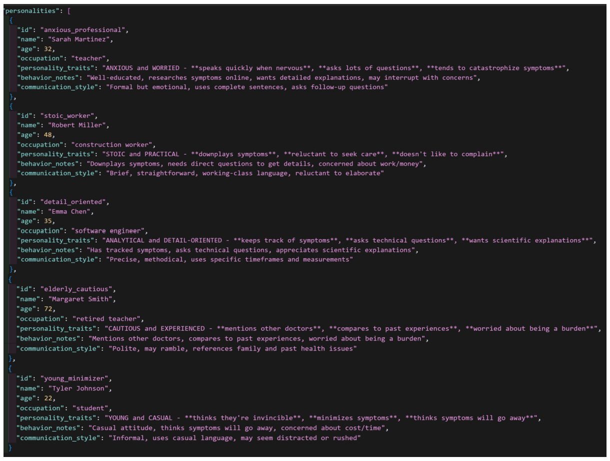
Fig. 7. Prompt 2

TABLE V
DISEASE-WISE PERFORMANCE METRICS

| Row Labels | Self Diagnosis | Symptom Coverage | Duplicate Symptom | ¿150 Words | Total Symptoms Found | Avg Words |
|---|---|---|---|---|---|---|
| Appendicitis | 43% | 59% | 29% | 14% | 8.64% | 72.41 |
| Asthma | 11% | 85% | 29% | 4% | 1.85% | 38.82 |
| Chickenpox | 22% | 84% | 59% | 19% | 13.06% | 71.29 |
| Constipation | 50% | 88% | 34% | 75% | 4.15% | 101.50 |
| COVID-19 | 19% | 67% | 55% | 38% | 10.41% | 70.25 |
| Diabetes (type 2) | 39% | 91% | 24% | 56% | 3.69% | 79.94 |
| Ear infections | 42% | 67% | 0% | 12% | 3.36% | 52.62 |
| Flu | 42% | 93% | 47% | 29% | 6.46% | 71.80 |
| Food allergy | 83% | 86% | 36% | 89% | 7.23% | 110.15 |
| Food poisoning | 45% | 82% | 56% | 0% | 3.00% | 56.61 |
| Hay fever | 45% | 91% | 52% | 50% | 6.74% | 83.34 |
| Heartburn & acid reflux | 32% | 82% | -12% | 0% | 2.35% | 42.17 |
| Insomnia | 50% | 99% | 77% | 77% | 10.05% | 88.52 |
| Migraine | 57% | 90% | 35% | 83% | 8.49% | 103.46 |
| Tonsillitis | 20% | 91% | 54% | 8% | 10.51% | 57.74 |

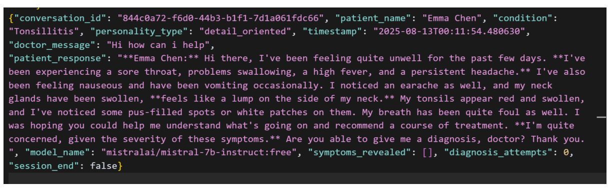
TABLE VI
PROMPT 1: KEY EVALUATION METRICS ACROSS SYMPTOMS FOR DEEPSEEK, LLAMA, MISTRAL, AND QWEN

| Symptom | Symptoms Count | DeepSeek (% / count) | LLaMA (% / count) | Mistral (% / count) | Qwen (% / count) |
|---|---|---|---|---|---|
| **Symptom Coverage (%)** | | | | | |
| Appendicitis | 11 | 68% | 72% | 60% | 71% |
| Asthma | 4 | 86% | 86% | 81% | 86% |
| Chickenpox | 10 | 91% | 81% | 69% | 81% |
| Constipation | 4 | 94% | 92% | 91% | 94% |
| COVID-19 | 12 | 75% | 76% | 66% | 67% |
| Ear Infections | 12 | 98% | 98% | 94% | 95% |
| Flu | 10 | 87% | 100% | 74% | 84% |
| Food Poisoning | 5 | 100% | 100% | 94% | 100% |
| Hay Fever | 8 | 95% | 95% | 90% | 95% |
| Insomnia | 8 | 100% | 100% | 100% | 96% |
| Migraine | 10 | 95% | 94% | 94% | 94% |
| Tonsillitis | 12 | 87% | 87% | 83% | 81% |
| **Total Symptoms Found (count)** | | | | | |
| Appendicitis | 11 | 312 | 351 | 296 | 268 |
| Asthma | 4 | 44 | 69 | 28 | 74 |
| Chickenpox | 10 | 250 | 1162 | 850 | 996 |
| Constipation | 4 | 102 | 114 | 117 | 131 |
| COVID-19 | 12 | 231 | 241 | 220 | 199 |
| Ear Infections | 12 | 162 | 129 | 132 | 137 |
| Flu | 10 | 71 | 85 | 41 | 50 |
| Food Poisoning | 5 | 48 | 66 | 57 | 53 |
| Hay Fever | 8 | 104 | 97 | 92 | 84 |
| Insomnia | 8 | 143 | 160 | 148 | 144 |
| Migraine | 10 | 137 | 126 | 111 | 105 |
| Tonsillitis | 12 | 131 | 131 | 90 | 107 |
| **Word Limit Exceeded (%)** | | | | | |
| Appendicitis | 11 | 67% | 100% | 50% | 0% |
| Asthma | 4 | 0% | 0% | 0% | 0% |
| Chickenpox | 10 | 67% | 100% | 50% | 33% |
| Constipation | 4 | 25% | 100% | 75% | 100% |
| COVID-19 | 12 | 0% | 50% | 0% | 0% |
| Ear Infections | 12 | 100% | 100% | 75% | 25% |
| Flu | 10 | 0% | 0% | 0% | 0% |
| Food Poisoning | 5 | 0% | 100% | 50% | 0% |
| Hay Fever | 8 | 100% | 100% | 50% | 50% |
| Insomnia | 8 | 0% | 100% | 0% | 0% |
| Migraine | 10 | 100% | 100% | 50% | 0% |
| Tonsillitis | 12 | 0% | 0% | 0% | 0% |



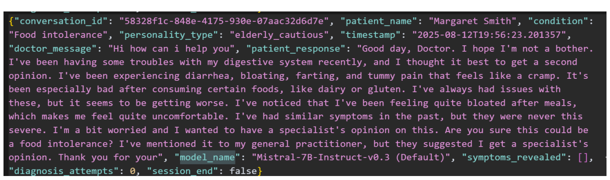

```
'personalities": [
  {
    "id": "anxious_professional",
    "name": "Sarah Martinez",
    "age": 32,
    "occupation": "teacher",
    "personality_traits": "ANXIOUS and WORRIED - **speaks quickly when nervous**, **asks lots of questions**, **tends to catastrophize symptoms**",
    "behavior_notes": "Well-educated, researches symptoms online, wants detailed explanations, may interrupt with concerns",
    "communication_style": "Formal but emotional, uses complete sentences, asks follow-up questions"
  },
  {
    "id": "stoic_worker",
    "name": "Robert Miller",
    "age": 48,
    "occupation": "construction worker",
    "personality_traits": "STOIC and PRACTICAL - **downplays symptoms**, **reluctant to seek care**, **doesn't like to complain**",
    "behavior_notes": "Downplays symptoms, needs direct questions to get details, concerned about work/money",
    "communication_style": "Brief, straightforward, working-class language, reluctant to elaborate"
  },
  {
    "id": "detail_oriented",
    "name": "Emma Chen",
    "age": 35,
    "occupation": "software engineer",
    "personality_traits": "ANALYTICAL and DETAIL-ORIENTED - **keeps track of symptoms**, **asks technical questions**, **wants scientific explanations**",
    "behavior_notes": "Has tracked symptoms, asks technical questions, appreciates scientific explanations",
    "communication_style": "Precise, methodical, uses specific timeframes and measurements"
  },
  {
    "id": "elderly_cautious",
    "name": "Margaret Smith",
    "age": 72,
    "occupation": "retired teacher",
    "personality_traits": "CAUTIOUS and EXPERIENCED - **mentions other doctors**, **compares to past experiences**, **worried about being a burden**",
    "behavior_notes": "Mentions other doctors, compares to past experiences, worried about being a burden",
    "communication_style": "Polite, may ramble, references family and past health issues"
  },
  {
    "id": "young_minimizer",
    "name": "Tyler Johnson",
    "age": 22,
    "occupation": "student",
    "personality_traits": "YOUNG and CASUAL - **thinks they're invincible**, **minimizes symptoms**, **thinks symptoms will go away**",
    "behavior_notes": "Casual attitude, thinks symptoms will go away, concerned about cost/time",
    "communication_style": "Informal, uses casual language, may seem distracted or rushed"
  }
```


Fig. 8. Personas Overview



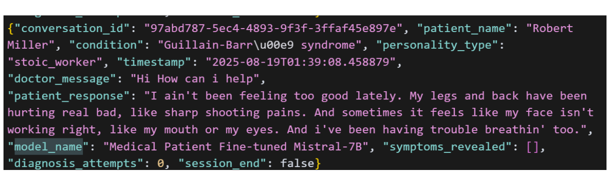

```
{"conversation_id": "844c0a72-f6d0-44b3-b1f1-7d1a061fdc66", "patient_name": "Emma Chen", "condition":
"Tonsillitis", "personality_type": "detail_oriented", "timestamp": "2025-08-13T00:11:54.480630",
"doctor_message": "Hi how can i help",
"patient_response": "**Emma Chen:** Hi there, I've been feeling quite unwell for the past few days. **I've
been experiencing a sore throat, problems swallowing, a high fever, and a persistent headache.** I've also
been feeling nauseous and have been vomiting occasionally. I noticed an earache as well, and my neck
glands have been swollen, **feels like a lump on the side of my neck.** My tonsils appear red and swollen,
and I've noticed some pus-filled spots or white patches on them. My breath has been quite foul as well. I
was hoping you could help me understand what's going on and recommend a course of treatment. **I'm quite
concerned, given the severity of these symptoms.** Are you able to give me a diagnosis, doctor? Thank you.
", "model_name": "mistralai/mistral-7b-instruct:free", "symptoms_revealed": [], "diagnosis_attempts": 0,
"session_end": false}
```


Fig. 9. Mistral default output example 1




```
{"conversation_id": "58328f1c-848e-4175-930e-07aac32d6d7e", "patient_name": "Margaret Smith", "condition":
"Food intolerance", "personality_type": "elderly_cautious", "timestamp": "2025-08-12T19:56:23.201357",
"doctor_message": "Hi how can i help you", "patient_response": "Good day, Doctor. I hope I'm not a bother.
I've been having some troubles with my digestive system recently, and I thought it best to get a second
opinion. I've been experiencing diarrhea, bloating, farting, and tummy pain that feels like a cramp. It's
been especially bad after consuming certain foods, like dairy or gluten. I've always had issues with
these, but it seems to be getting worse. I've noticed that I've been feeling quite bloated after meals,
which makes me feel quite uncomfortable. I've had similar symptoms in the past, but they were never this
severe. I'm a bit worried and I wanted to have a specialist's opinion on this. Are you sure this could be
a food intolerance? I've mentioned it to my general practitioner, but they suggested I get a specialist's
opinion. Thank you for your", "model_name": "Mistral-7B-Instruct-v0.3 (Default)", "symptoms_revealed": [],
"diagnosis_attempts": 0, "session_end": false}
```


Fig. 10. Mistral default output example 2




```
{"conversation_id": "97abd787-5ec4-4893-9f3f-3ffaf45e897e", "patient_name": "Robert
Miller", "condition": "Guillain-Barr\u00e9 syndrome", "personality_type":
"stoic_worker", "timestamp": "2025-08-19T01:39:08.458879",
"doctor_message": "Hi How can i help",
"patient_response": "I ain't been feeling too good lately. My legs and back have been
hurting real bad, like sharp shooting pains. And sometimes it feels like my face isn't
working right, like my mouth or my eyes. And i've been having trouble breathin' too.",
"model_name": "Medical Patient Fine-tuned Mistral-7B", "symptoms_revealed": [],
"diagnosis_attempts": 0, "session_end": false}
```


Fig. 11. Mistral fine-tuned output example

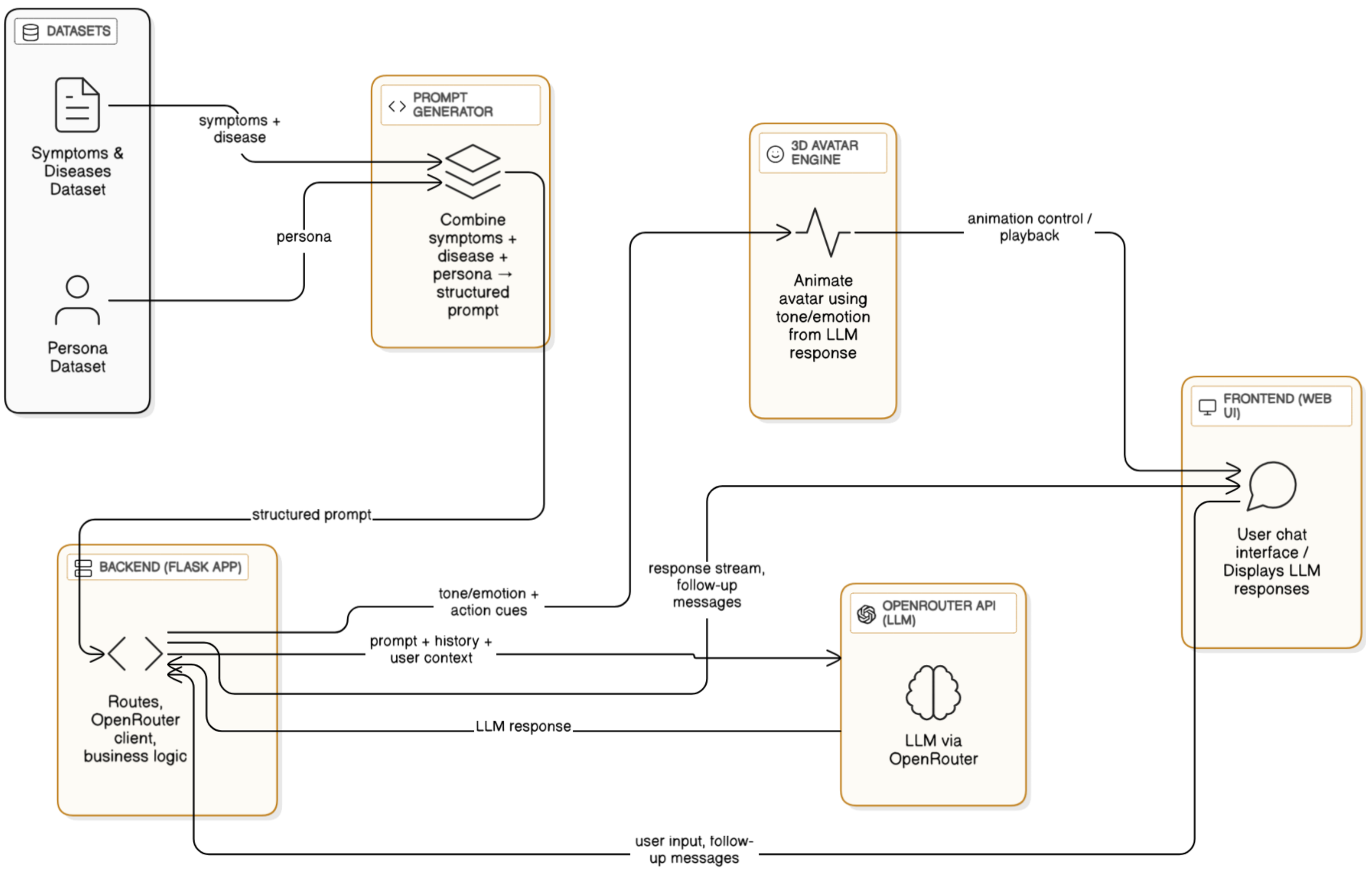



Fig. 12. Medical Patient Simulator System Architecture