MSc Project - Reflective Essay

Project Title:	AI-Powered Medical Patient Simulator: Comparing Prompt Engineered LLMs with Exploratory Fine-Tuning
Student Name:	Muhammad Ibrahim Arain
Student Number:	220991593
Supervisor Name:	Fandi Meng
Programme of Study:	Machine Learing for Visual Data Analytics

INTRODUCTION

For my MSc project, I developed an AI-powered virtual patient simulator for medical training. Instead of fixed conditional statements on binary trees, the simulator uses large language models (LLMs) with prompt engineering. The prompts are created through predefined patient personas and a condition and symptom taken from NHS website about symptoms and conditions. The aim was to simulate realistic patient interactions that could help medical students and healthcare workers practise diagnostic questioning in a safe, repeatable environment.

The topic was chosen because there are limitations to current virtiual patients and my medical student fellows have often discussed the lack of good diagnosis practice. Furthurmore LLM prompt engineering and fine tuning are currently industry desired learnings and hence both medical and industrial applicatiosn were there to test it out. The challenge was not simply getting the model to “talk” as a patient but ensuring it would remain accurate, avoid hallucinating symptoms, adhere to designed personas, and adapt to different consultation styles.

From the outset, I knew the project would require a balance between **medical accuracy**, **conversational realism**, and **technical feasibility**, and that each of these would compete for priority at different stages of development.

Analysis of Strengths and Weaknesses

One of the main strengths of the project was grounding it in authentic NHS data. This ensured medical accuracy and reduced symptom hallucination — a problem I had noted in many medical AI prototypes that rely solely on model-trained knowledge. Scraping and cleaning NHS web data into a structured JSON format created a dependable knowledge base across 50 selected conditions, vetted by medical professionals to be diagnosable through text conversation alone. Furthermore, if successful on making LLMs only rely on provided symptoms and not hallucinate, it could set a good precedent for industries to prompt engineer LLMs on their data and use it with more confidence that the LLM wont hallucinate the information like making a Q&A bot of a companys terminologies and KPIs.

Another key strength was the **iterative approach to prompt design**. Initially, I defaulted to fine-tuning as the route to “teaching” a model patient behaviour, but as I focused on generating dataset for fine tuning I discovered that prompt engineering, paired with high-quality data, could achieve similar control without heavy compute costs. This shift led to the development of multiple prompt types, with Prompt 1 delivering format-structured accuracy and Prompt 2 fostering naturalistic, persona-rich conversations. Testing them systematically on multiple models enabled a data-driven selection of Qwen for deployment.

The overall architecture also demonstrated flexibility. Using OpenRouter allowed for multi-model integration during testing without vendor lock-in, and the modular backend/frontend design meant updates to the conversation engine could be made without full redeployment.

However, several weaknesses emerged:

- **Overly rigid initial prompts** made conversations feel mechanical, with patients "announcing" symptoms and traits in a way that sounded scripted.
- In Prompt 2, while naturalness improved, looser constraints sometimes caused small but noticeable hallucinations.
- **Visualisation limits**, my 3D avatar system responded to symptom keywords with basic animations, but lacked nuanced body language, facial expressions, or continuous motion.
- Some personas, particularly the "Elderly Cautious" type, led to frequent word-limit breaches and verbosity-driven drift.
- Despite planned evaluation of reinforcement learning from conversation histories, time constraints kept experimentation here minimal.

One of the most time-consuming hurdles came during my fine-tuning attempts. Generating the dataset through the API was slow, and then every conversation had to be checked because some sessions were incomplete, either the API failed midway or went into pause mode, leaving gaps. I solved this by putting 15-30 sec delay after every 5 api calls and retrying if it failed after 30 sec again to keep the process running as automated. After dataset was completed, I had to clean it out, to remove api error and empty response conversations, before formatting everything into the **assistant/system message** style required for training. Running the training on Google Colab brought its own problems; the GPU switched twice during runs, which killed the process partway both times. Even when it completed, the first model learned to produce **both doctor and patient turns**, which broke role separation entirely. I tried again with a slightly different format, but the same thing happened, and at one point I was close to giving up. In the final attempt, I restructured the data so that past conversation history and turn-taking were embedded into the system prompt instead. Just when it seemed promising, the script started failing due to outdated library dependencies. After updating the environment and fixing the code, the training finally completed successfully — proving it could be done, but at a cost of significant time and effort compared to prompt engineering.

The fine tuning only improved the conversation style which was the though process going in. Mistral 7b default tended to produce suggestive question to the doctor about if they can solve the problem and Thankyou like a formal letter style, something none of the bigger LLMs did. The thought process I had to do this was if I could change it to speak more like bigger LLMs especially regarding these doctor question. Despite stylistic improvements, medical accuracy didn't im prove much compared to default 7b model. Symptom coverage was 5–10% lower, and hallucination rates were approximately 6% higher compared to baseline LLaMA. Thus, fine-tuning improved style and compliance but not clinical precision. Overall, these findings confirm that prompt engineering remains the more effective strategy for realistic, cost-efficient simulation within our resource constraints.

Presentation of Possibilities for Future Work

If I had more time and resources, I would:

1. **Expand the dataset** - add sub-conditions, co-morbidity patterns, and more symptom overlap to increase diagnostic complexity.
2. **Integrate dynamic difficulty scaling** - adjust patient cooperativeness, accuracy, or misdirection based on the trainee's proficiency.
3. **Hybrid prompt strategies** - develop a context-aware system that uses Prompt 1 discipline for initial fact-gathering, then transitions to Prompt 2 for open, human-like exchange.
4. **Richer multi-modal integration** - implement facial expression mapping, posture shifts, and realistic audio output; eventually link to AR/VR systems for immersive OSCE-style simulations.
5. **Clinical validation studies** - compare trainee diagnostic accuracy and communication skill development when using the simulator vs. standardised patient actors or scripted systems.
6. **EHR case realism** - adapt the simulator to draw from anonymised electronic health record data to create longitudinal patient profiles.
7. **Multi-user simulation** - allow trainee groups to collaboratively assess a patient, reflecting team-based healthcare practice.
8. **Automated performance feedback** - integrate symptom coverage analysis, missed cues, and conversational effectiveness scoring into post-session review.
9. **Using different models for fine-tuning** – Since we learned that fine tuning does have an impact, we could try on 1-2 other models with better datasets to see if we can achieve a higher LLM performance

Critical Analysis of the Relationship Between Theory and Practical Work

Theoretically, the project draws on **LLM behaviour shaping via prompt engineering**, conversational simulation design, and structured data integration. The process confirmed that while theory supports strict constraints for controlling output and reducing hallucinations, these constraints directly impacted engagement in a training context. In practice, this meant **user experience became inseparable from accuracy** something not always captured in purely technical AI development literature.

The architecture supported theory-to-practice alignment: roles were explicitly defined in system prompts, symptom hierarchies mirrored clinical questioning, and persona notes embedded behavioural variance. The inclusion of an NHS-grounded dataset provided a layer of truthfulness that theory suggests should constrain hallucination, and evaluation confirmed this benefit.

Evaluation itself blended **automation**, measuring symptom coverage, repetition, and adherence to rules, with **manual qualitative review**. The latter was essential: metrics alone could not capture tone-shifts, subtle persona drift, or whether a patient “felt” human. This supports existing research emphasising that conversational realism evaluation remains a partly human judgement problem.

User feedback at the end was also helpful, for example, very early feedback it was suggested that the LLM doesn’t always speak the most relevant symptom first, like for migraine instead of first symptom being headache it would say tiredness or fatigue, so this was a quick fix where symptoms were divided into Most important and then other symptoms, and an update in prompt to start with most important ones first solved the issue.

Awareness of Legal, Social, Ethical Issues, and Sustainability

From the start, I recognised patient safety and accuracy as non-negotiable. Every symptom generated was checked against the NHS dataset to ensure it was grounded in verified medical information. Any deviation was treated as a critical fault in testing.

Data privacy considerations were addressed by using only publicly available medical sources and avoiding real patient records. This eliminated the need for complex GDPR compliance measures but also limited access to contextualised, case-based conversation material.

Social and fairness concerns were also part of persona design. Personas were created with behavioural, not demographic, identifiers to avoid reinforcing cultural or identity-based stereotypes. Nonetheless, I am aware that subtle biases can creep into interactive AI and would require ongoing monitoring.

From a sustainability perspective, the choice of **prompt engineering over fine-tuning** not only kept computational costs low but reduced environmental impact — a long-term benefit if the system were adopted across institutions.

On the broader ethical horizon, tools like this must ensure equitable access, avoid training only on high-resource language/cultural contexts, and include transparency features so trainees know when and how the AI’s medical grounding is applied.

Conclusion

This project enabled me to blend my technical skills in AI, web development, and data handling with an application deeply rooted in educational value. The final simulator proved that **prompt engineering plus high-quality data grounding** can successfully replicate realistic, medically accurate patient interactions without the prohibitive costs of full model fine-tuning.

A central takeaway is the trade-off between **conversation control** and **realism**. The choice of Qwen as the deployed model was not purely about highest accuracy, it was about balancing accuracy, role consistency, and the human feel that engages learners.

The most important lesson was the **value of pivoting** when an original plan no longer serves the project goals. Shifting from fine-tuning to a prompt-driven approach strengthened the deliverable, made it feasible within time limits, and created a more adaptable foundation for future development.

With further advancement, especially in dynamic difficulty, richer multimodal cues, and validated training outcomes, I believe this simulator could become a significant, sustainable addition to the medical training toolkit, helping bridge the gap between textbook scenarios and the unpredictable reality of real patient interaction.