Virtual Patient Model Report

# Model Choice

We selected a LLaMA-based lightweight model (e.g., 7B parameter variants or quantized Gemma versions compatible with Ollama) as it provides a strong balance between performance and resource requirements. These models operate efficiently on modest hardware, aligning with our goal of ensuring usability even in resource-constrained environments.  
  
To enable efficient fine-tuning, we applied LoRA (Low-Rank Adaptation). Unlike full model retraining, LoRA modifies only a small subset of the model’s weights, significantly reducing memory and compute requirements while preserving model performance.

# Approach

Our methodology included the following steps:  
1. Formatting synthetic doctor-patient conversation scripts into structured prompts with patient persona labels (e.g., calm, anxious, rude) to guide response personalities.  
2. Loading the base model in 8-bit quantized mode to minimize memory usage.  
3. Applying LoRA fine-tuning on persona-conditioned dialogues to teach realistic patient responses.  
4. Training with a reasonable batch size and mixed precision to maximize hardware compatibility.  
5. Testing the model by generating persona-conditioned responses to evaluate personality consistency and naturalness.

# Challenges

Throughout development, we encountered several challenges:  
- Persona Consistency: Ensuring the model maintained its assigned persona across an entire conversation proved difficult. Careful labeling and hyperparameter tuning were required.  
- Resource Constraints: Limited memory and compute capacity necessitated the use of smaller models, 8-bit quantization, and LoRA fine-tuning.  
- Data Variety: Synthetic scripts lacked diversity, limiting the model’s ability to capture nuanced patient emotions and dialogue styles.  
- Balancing Realism and Flexibility: Although we aimed for natural interactions, generic training examples sometimes led to overly safe or repetitive responses.

# Improvements for the Future

To enhance the system further, we propose the following improvements:  
- Enrich Training Data: Incorporate diverse, high-quality doctor-patient dialogues and persona examples.  
- Enhanced Persona Conditioning: Utilize advanced prompt engineering or control tokens to improve persona fidelity and allow smooth persona switching.  
- Post-Training Enhancements: Explore methods such as Reinforcement Learning with Human Feedback (RLHF) to refine authenticity and helpfulness.  
- Multimodal Integration: Add voice tone and facial expression outputs for an immersive VR patient experience.  
- Optimized Inference Latency: Apply quantization and model distillation techniques to improve response speed, particularly on edge devices.

# Conclusion

By selecting efficient models and leveraging lightweight fine-tuning techniques like LoRA, we successfully developed a practical, persona-aware virtual patient. The system demonstrates realism and flexibility while remaining optimized for limited-resource environments. Future improvements will further enhance diversity, fidelity, and responsiveness, enabling more engaging VR medical training experiences.