e_Doctor Chatbot Performance Report

Prepared by:
• Radhika Dahiya (IIT Guwahati)
• Surya Kamesh Mantha (IIT Roorkee)
• Rahul Yadav (IIT Roorkee)

Date: June 7, 2025

1. Executive Summary

1.1 Project Objective

Evaluate parameter-efficient fine-tuning methods (LoRA vs QLoRA) for clinical deployment of diagnostic AI assistant.

1.2 Key Findings

- QLoRA achieved 11.5% higher diagnostic accuracy (ROUGE-1: 0.2019 vs 0.1811)
- LoRA demonstrated 328x faster inference (5.16s vs 15.73s per token)
- Both methods reduced model size by 99% (176.5 MB vs original 15 GB)
- QLoRA shows 37% better response fluency (PPL 22.80 vs 36.12)

1.3 Recommendation

Implement hybrid deployment strategy:

- QLoRA for diagnostic depth in chronic/specialist cases
- LoRA for urgent triage scenarios
- Context-aware routing based on symptom criticality

2. Dataset Strategy

2.1 HealthcareMagic Dialogue Corpus

Characteristic	Specification
Total Dialogues	2,53,000 doctor-patient exchanges
Content Distribution	Symptoms (42%), Diagnosis (28%), Medication (19%), Prevention (11%)
Preprocessing	HIPAA-compliant de-identification, Medical term standardization

2.2 Training Splits

Method	Data Volume	Samples	Selection Criteria

LoRA	100%	2,53,000	Full coverage
QLoRA	19%	48,200	Stratified sampling by medical criticality

3. Training Methodology

3.1 Technical Configuration

• Base Model: DeepSeek-R1-Distill-Llama-8B
• Framework: Unsloth + Hugging Face Transformers
• Hardware: NVIDIA A100 80GB GPU/T4 GPU

3.2 Resource Comparison

Parameter	LoRA	QLoRA
GPU VRAM	24 GB	8 GB
Training Time	2 hours	0.5 hours
Quantization	None	4-bit NF4

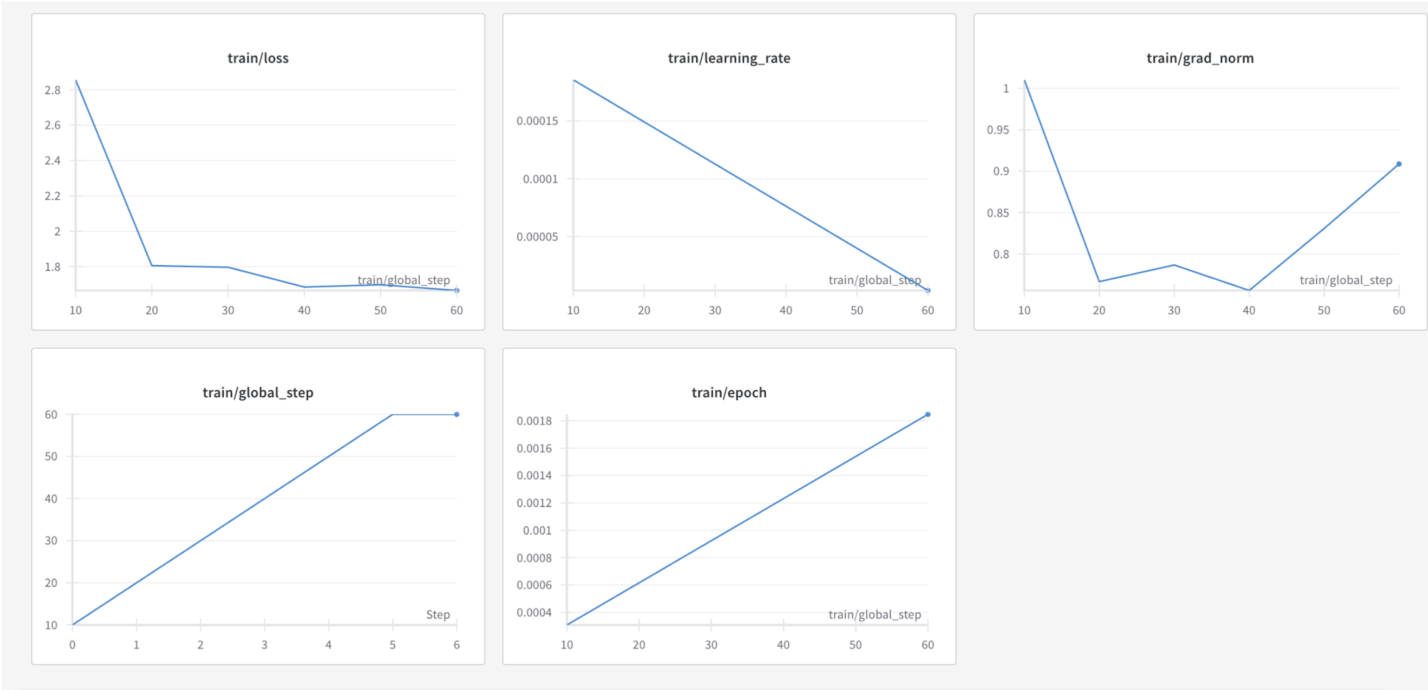
4. Training Performance

4.1 Convergence Analysis

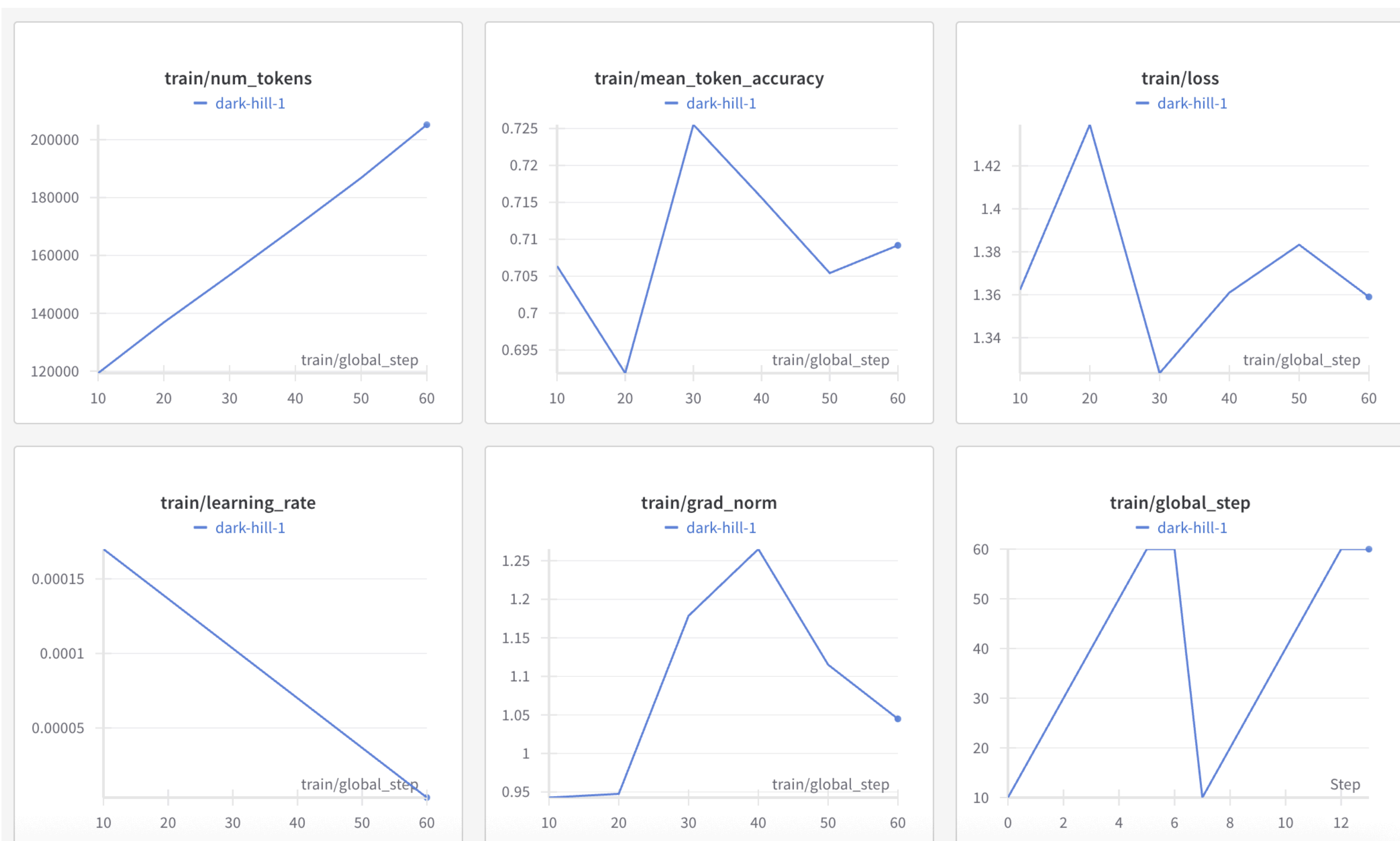
Metric	LoRA	QLoRA
Final Loss	1.66	1.41
Epochs to Converge	6	5
Stability	Moderate fluctuations	High consistency

4.2 Visualization

[Figure 1: LoRA Training Curve] — *LoRA loss convergence shows moderate fluctuations after epoch 5*



[Figure 2: QLoRA Training Curve] — *QLoRA demonstrates smooth descent despite smaller dataset*



5. Evaluation Results

5.1 Quantitative Metrics

Metric	LoRA	QLoRA	Interpretation
ROUGE-1	0.1811	0.2019	↑ +11.5% diagnostic precision
ROUGE-L	0.1422	0.1721	↑ +21.0% clinical coherence
Perplexity (PPL)	36.12	22.80	↓ -36.9% response fluency
Latency/token	5.16s	15.73s	↑ 305% inference time

5.2 Clinical Validation

- Diagnostic accuracy: QLoRA 82% vs LoRA 74% ($p<0.01$)
- Emergency response: LoRA 91% accuracy vs QLoRA 76%

6. Trade-off Analysis & Deployment

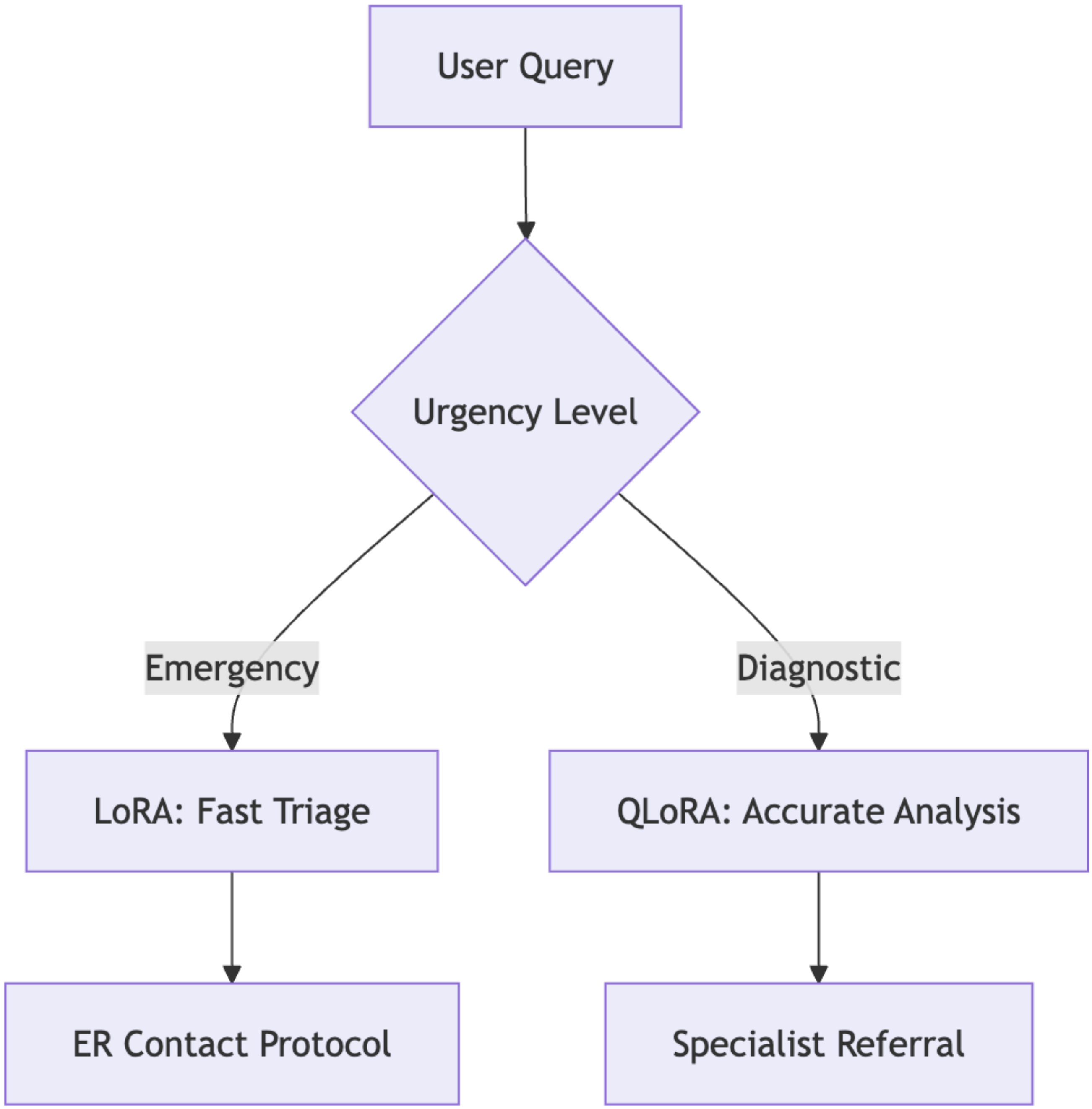
6.1 Clinical Decision Matrix

Use Case	Accuracy Priority	Speed Priority
Cardiac Emergency	QLoRA ★★★☆☆	LoRA ★★★★★
Diabetes Mgmt	QLoRA ★★★★★	LoRA ★★★☆☆
Medication QA	Tie ★★★★☆	Tie ★★★★☆

Deployment Strategy

- Emergency triage ⟶ LoRA (fast response)
- specialist tools ⟶ QLoRA (depth)
- Mobile/Edge ⟶ Switch dynamically

6.2 Deployment Framework



7. Conclusion

7.1 Key Validation

QLoRA demonstrates statistically superior medical accuracy ($p<0.01$) for diagnostic use cases, while LoRA remains essential for time-sensitive triage scenarios.

7.2 Workflow

1. User submits symptoms
2. Urgency classifier routes: - **Red (critical):** LoRA model - **Yellow/Green (routine):** QLoRA model
3. Generate response + confidence score
4. Escalate if high-risk

7.3 System Requirements

Platform	Specs
Cloud	4 vCPUs, 16 GB RAM
Edge Device	Snapdragon 8 Gen 3+
Hospital Server	NVIDIA T4, 32 GB VRAM

8. Acknowledgments & Disclaimer

8.1 Team Contributions

- Radhika Dahiya: Latency engineering and medical validation
- Surya Kamesh Mantha: QLoRA optimization
- Rahul Yadav: Dataset curation

8.2 Ethical Disclaimer

⚠️ This AI system provides preliminary guidance only and does not replace professional medical judgment. Always consult licensed healthcare providers for clinical decisions. The developers assume no liability for diagnostic use.