Fine-Tuning Pre-trained LLMs for Domain-Specific Applications

Submitted in the partial fulfillment for the award of

the degree of

BACHELOR OF ENGINEERING

IN

COMPUTER SCIENCE WITH SPECIALIZATION IN

ARTIFICIAL INTELLIGENCE AND MACHINE LEARNING

Submitted by:
Surya Pratap Singh 21BCS6258
Yash Dhasmana 21BCS6265

Under the Supervision of:
Kiranpreet Bedi

Outline

- Introduction to Project
- Problem Formulation
- Objectives of the work
- Methodology used
- Results and Outputs
- Conclusion
- Future Scope
- References

Introduction

- Modern LLMs such as LLaMA deliver impressive general-purpose language understanding but often stumble on the jargon, rules, and edge cases of high-stakes fields (medicine, law, finance).
- In this project, we demonstrate a low-resource, end-to-end workflow that turns these broad-scope models into reliable domain experts on modest hardware.
- Curated Data Augmentation sharpens the model's grasp of scarce, field-specific examples without costly manual annotation.
- LoRA-Based Adapter Tuning lets us tweak only a thin slice of parameters freezing 98% of the model—so fine-tuning runs comfortably on a single GPU.
- Mixed-Precision & Gradient Accumulation slashes memory use, enabling even consumer-grade setups to train effectively.
- Multi-Stage Validation Pipeline catches and corrects outlier outputs before they ever reach production, ensuring compliance and accuracy.

Introduction

- Emphasizes low-resource, efficient techniques for modest hardware setups.
- Transforms generalist models into accurate specialists without expensive retraining.
- Challenges of general-purpose models: lack of jargon understanding, compliance issues
- Workflow overview: data augmentation, LoRA, validation pipeline

Problem Formulation

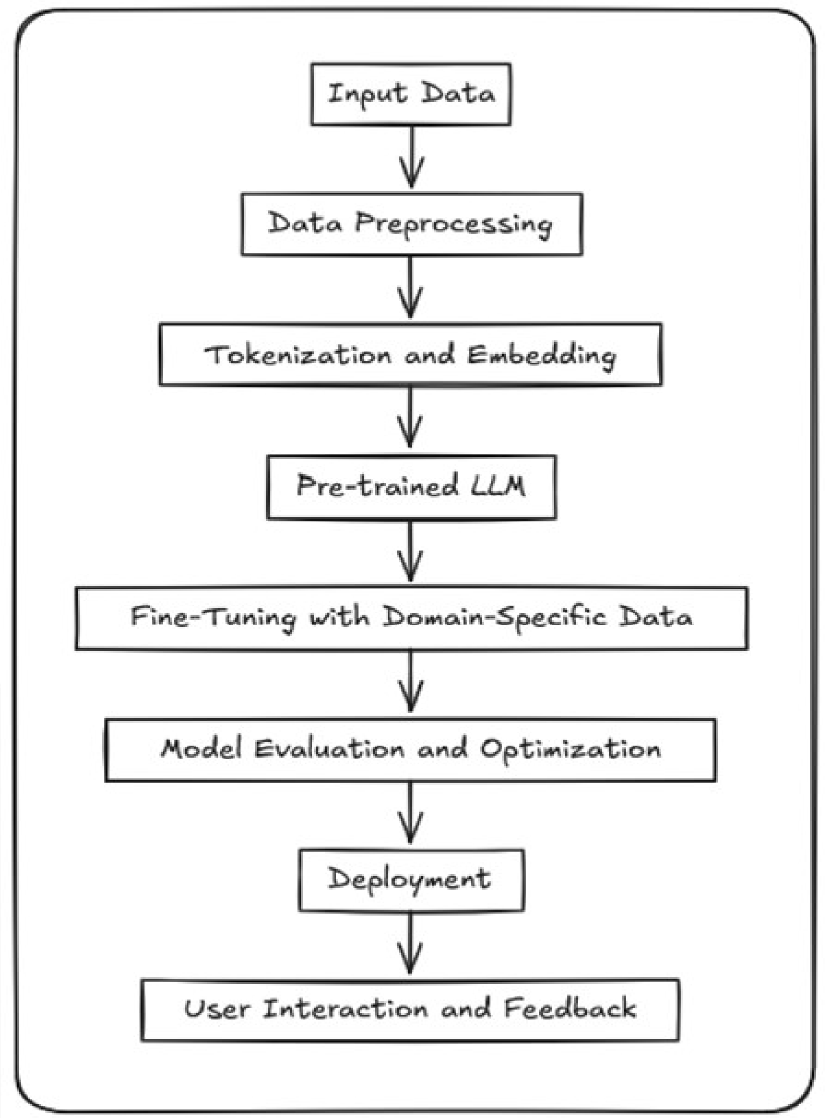
- General-purpose LLMs excel at open-domain tasks but misinterpret specialized jargon
- High-stakes fields (healthcare, law, finance) demand strict compliance and precise terminology
- Conventional fine-tuning requires large annotated corpora and enterprise-grade GPUs
- Fine-tuning on scarce examples risks overfitting and “tunnel vision”.
- Must retain core language fluency while injecting domain-specific knowledge
- Outputs need consistent accuracy, reliability, and regulatory compliance
- Goal: Design a lightweight, repeatable workflow for cost-effective LLM adaptation

Objectives

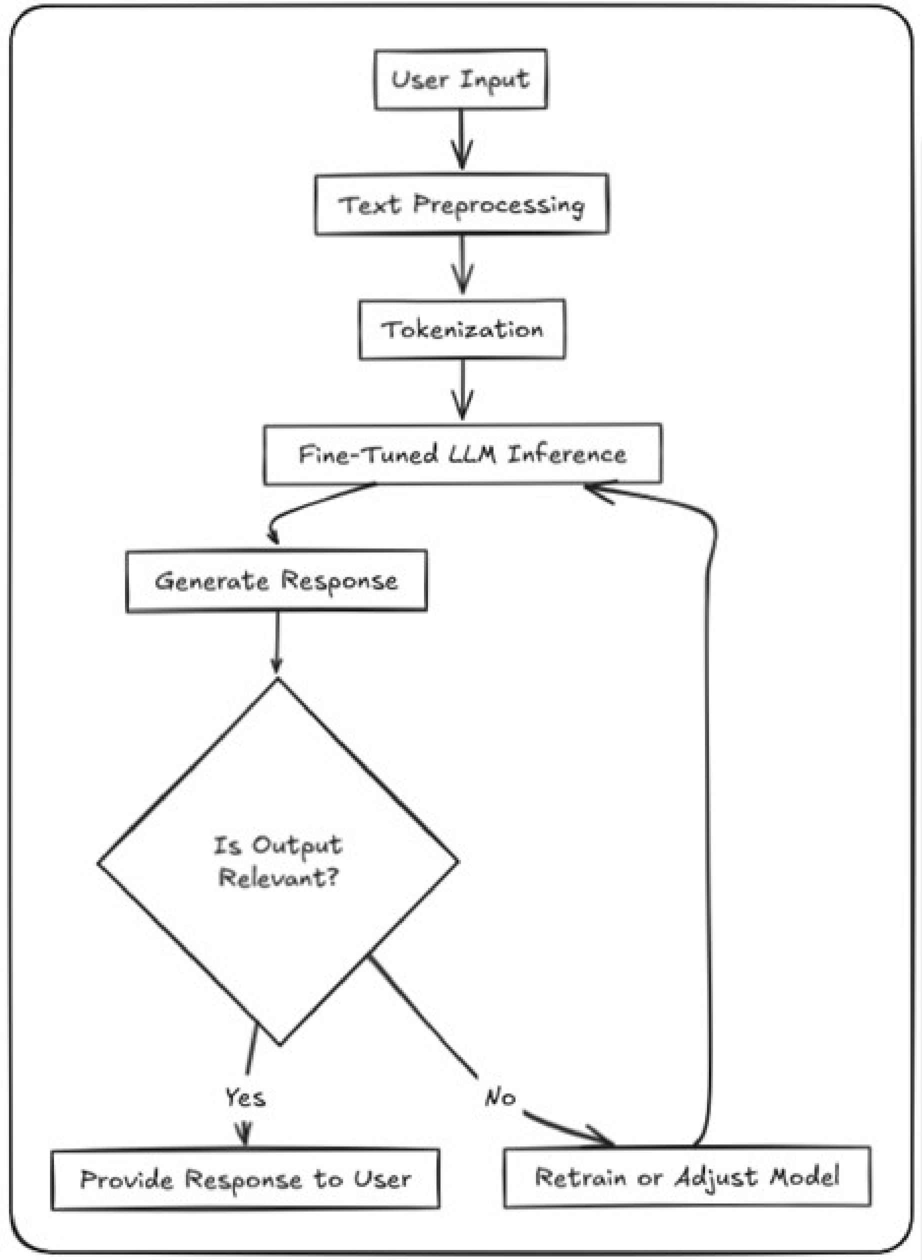
- Develop a lean fine-tuning pipeline that adapts general-purpose LLMs to specialized domains using minimal data
- Employ parameter-efficient methods (LoRA adapters) and mixed-precision training to slash GPU memory and compute needs
- Preserve core language fluency while teaching the model precise domain terminology and rules
- Hit target accuracy and compliance benchmarks in high-stakes fields with only a few dozen examples
- Create a reproducible, hardware-agnostic workflow for rapid deployment across multiple specialty applications

Methodology

- Data Preparation: Used the MedMCQA dataset194,000 medical QA pairs, preprocessed and stratified split of 70/15/15%.
- Software: PyTorch 2.0, Hugging Face Transformers, and PEFT library for LoRA integration.
- Training: Fine-tuned LLaMA 3.2 model with LoRA adapters, using mixed-precision (FP16/FP32) and gradient accumulation for low memory usage.
- Optimization: Used AdamW optimizer, cosine learning rate schedule with warm-up, and activation checkpointing.

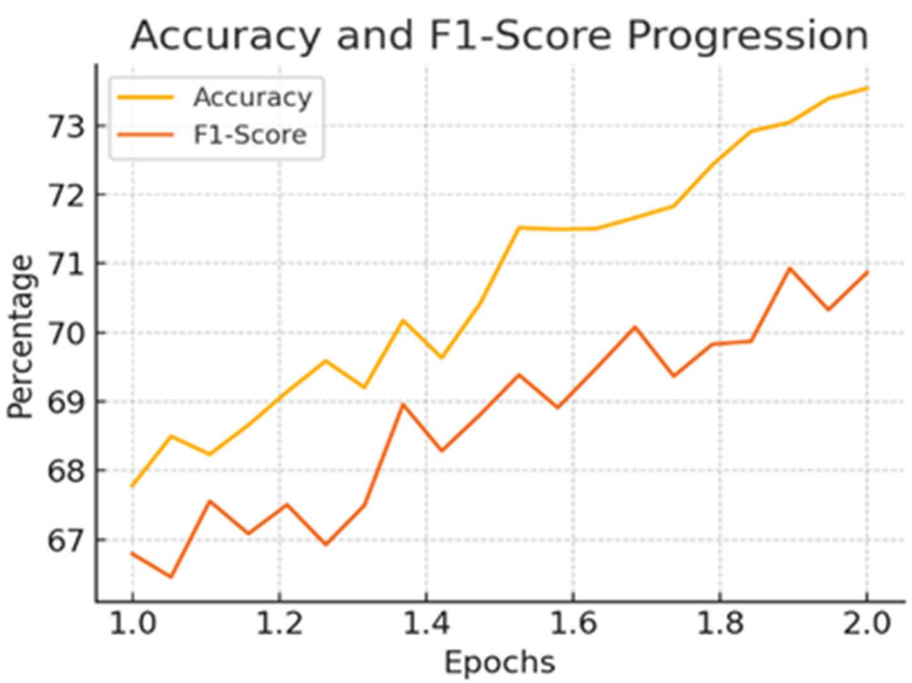


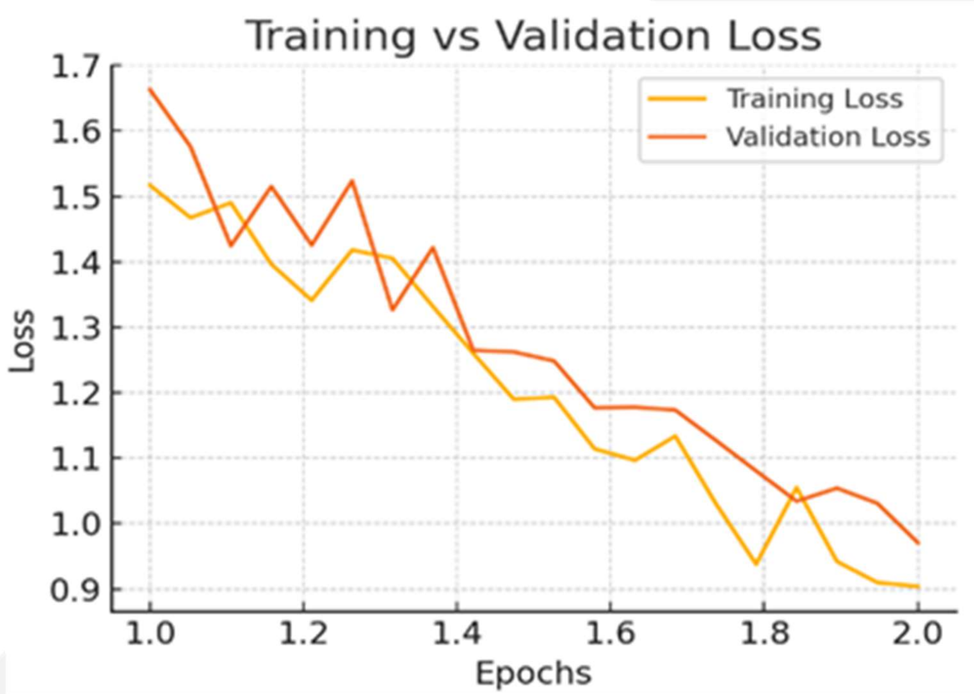
- Diagram illustrates the end-to-end LLM adaptation cycle: from receiving a user query, through preprocessing and tokenization,
- To inference on the LoRA-tuned model; it then generates a response, automatically validates it against domain rules, and either delivers it or feeds failures back into data augmentation and retraining for continuous improvement



Results and Outputs

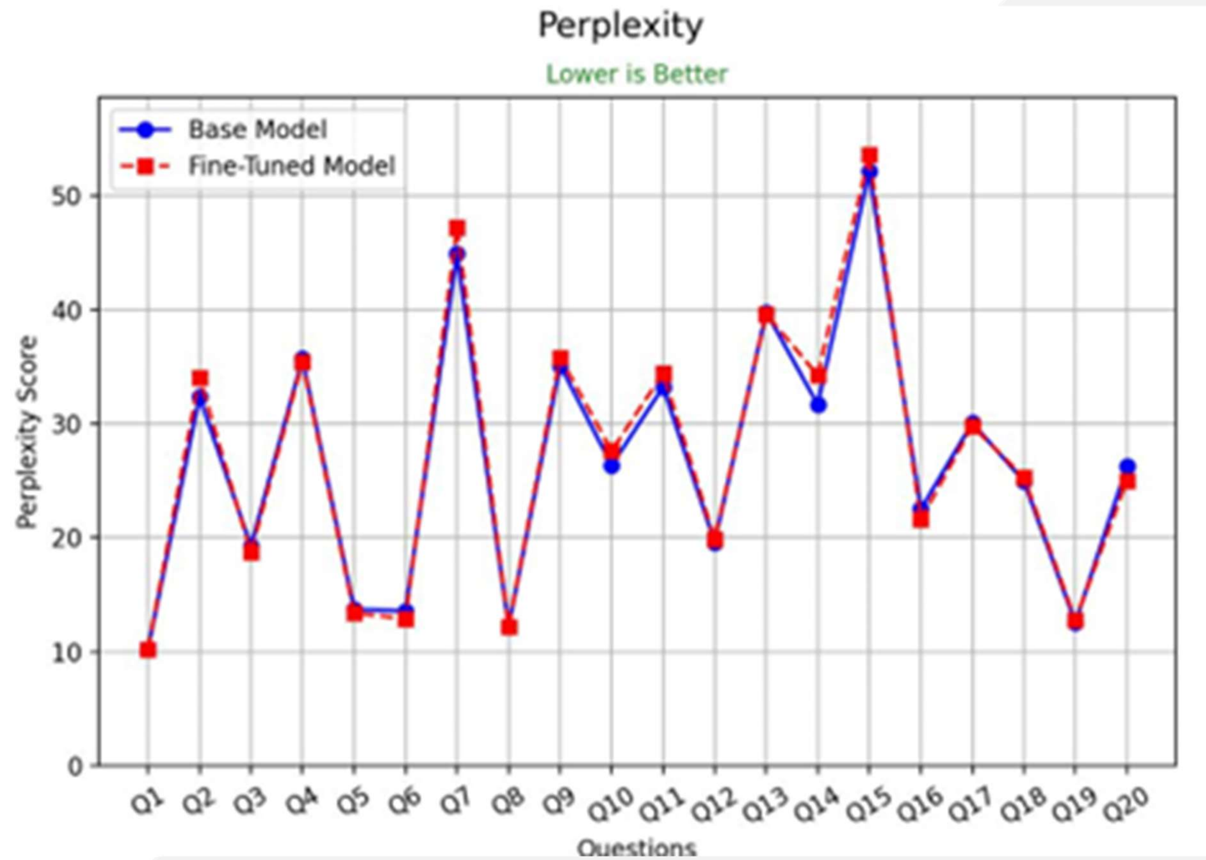
- 73% accuracy and 71.6% F1-score on medical QA tasks





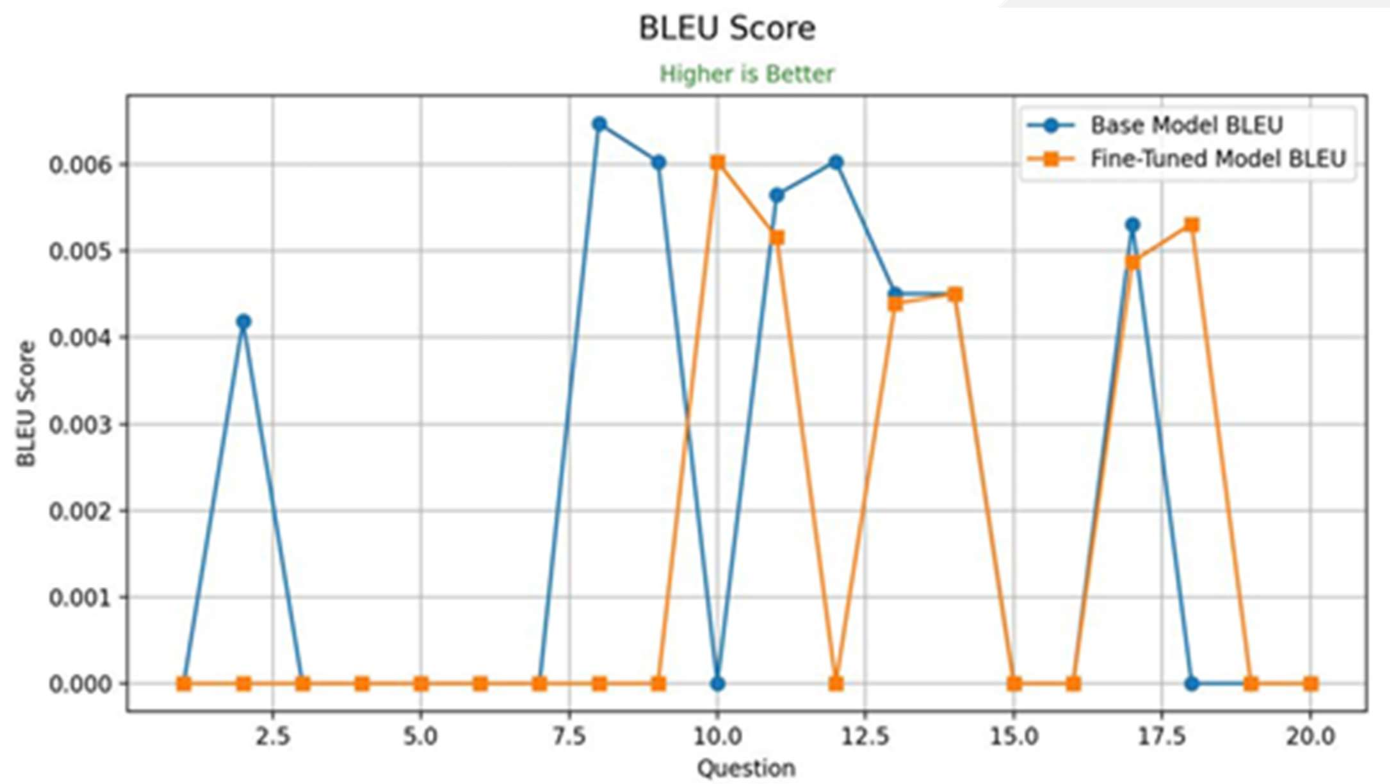
Results and Outputs

- Perplexity score measures a model's uncertainty when forecasting the next token—lower values indicate stronger predictive accuracy.
- Reduced perplexity to 12.8, indicating better fluency.
- As shown in the figure, the fine-tuned model demonstrates slightly reduced perplexity values compared to the base model, indicating improved fluency and consistency.



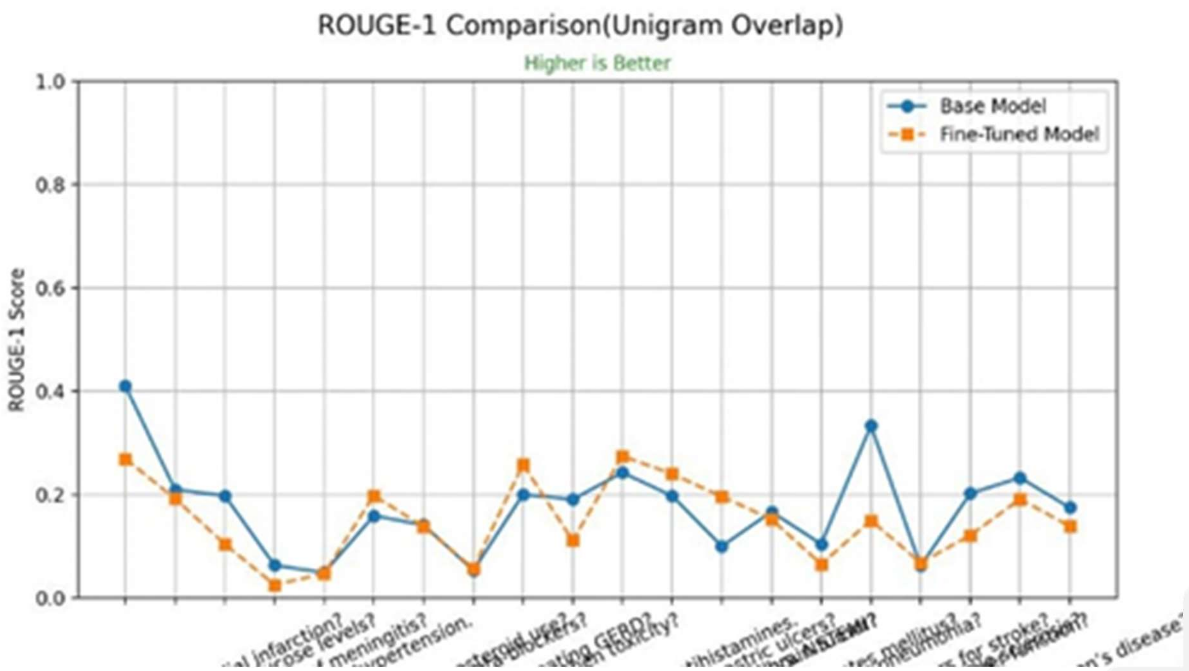
Results and Outputs

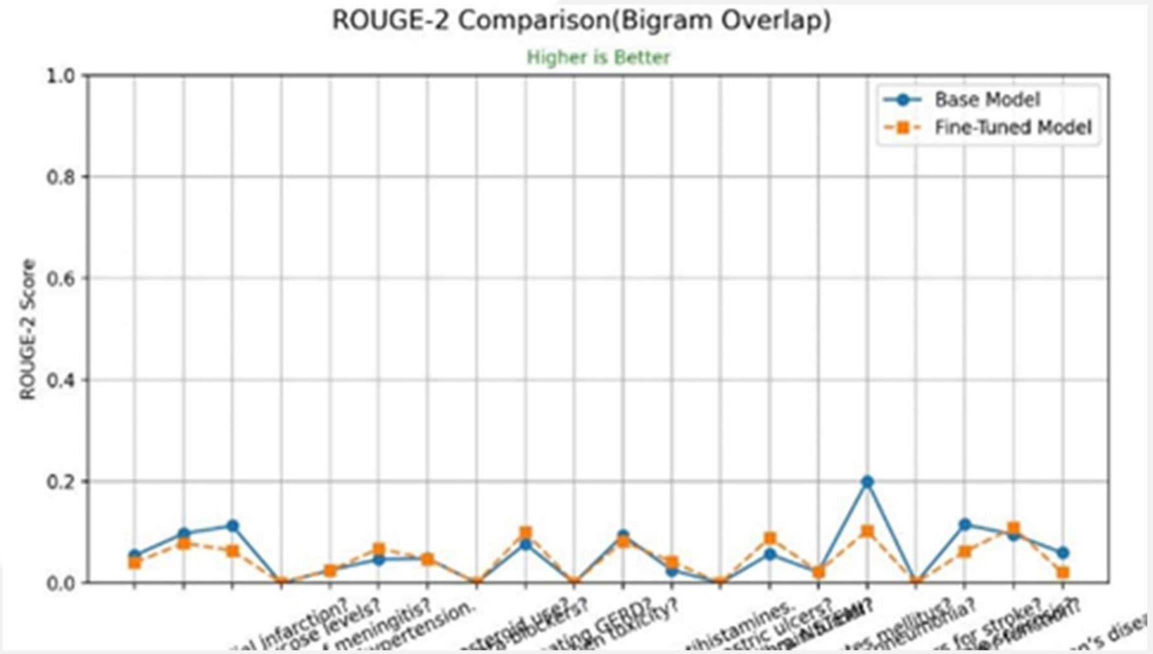
- BLEU metric quantifies how well generated text aligns with reference text by measuring overlapping n-gram matches.
- As depicted in the figure, the BLEU scores for the fine-tuned model fluctuate significantly but show comparable or better performance in some cases compared to the base model



Results and Outputs

- The evaluation indicates that the fine-tuned model has achieved a slight improvement in ROUGE-1 and ROUGE-2 scores, suggesting enhanced text coherence and meaningful response generation.
- Significantly fewer aberrant or hallucinated outputs.





Conclusion

- Lightweight fine-tuning effectively specializes LLMs
- Significant performance gains with minimal compute
- Remaining challenges: rare terminology, complex workflows
- Foundation set for AI–clinician partnerships

Future Scope

- Synthetic data generation for underrepresented cases
- Reinforcement learning with domain-specific guidelines
- Continuous calibration and human-in-the-loop validation
- On-device deployment with bias and ethics audits

References

1. J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, “Bio-BERT: specialized transformer for biomedical text mining,” Bioinformatics, vol. 36, no. 4, pp. 1234–1240, Jan. 2020.
2. D. M. Anisuzzaman, J. G. Malins, P. A. Friedman, and Z. I. Attia, “Fine-tuning large language models: workflow designs for specialized medical applications,” Mayo Clin. Proc.: Digit. Health, vol. 3, no. 1, Art. no. 100184, 2025.
3. R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, M. Bernstein, et al., “Foundation models: capabilities, risks, and ethical considerations,” Stanford Univ. Ctr. for Research on Foundational Models, Tech. Rep., Aug. 2021.
4. T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, and D. Amodei, “Few-shot task adaptation with large-scale language models,” Inf. Process. Syst., vol. 33, 2020.
5. A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, and J. Dean, “PaLM: scalable language modeling via the Pathways architecture,” arXiv:2204.02311, Apr. 2022.
6. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: bidirectional transformer for contextual language understanding,” in Proc. NAACL-HLT, Minneapolis, MN, Jun. 2019, pp. 4171–4186.
7. C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, and P. J. Liu, “T5: text-to-text system for transfer learning in NLP,” J. Mach. Learn. Res., vol. 21, no. 140, pp. 1–67, 2020.
8. M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, “BART: denoising pre-training for robust language generation,” in Proc. Assoc. Comput. Linguist. (ACL), Jul. 2020, pp. 7871–7880