Report: AlpaCare Medical Instruction Assistant

Title

Fine-Tuning a Safe Medical Instruction Assistant Using LoRA on AlpaCare-MedInstruct-52k

Abstract

This report presents the development of the **AlpaCare Medical Instruction Assistant**, a fine-tuned large language model (LLM) designed to generate **safe, educational, and non-diagnostic medical responses**. The primary objective of this project was to adapt a lightweight open-source model for medical education use while ensuring strict adherence to ethical and safety standards. The fine-tuning process employed the **lavita/AlpaCare-MedInstruct-52k** dataset, which provides high-quality instruction-response pairs across a range of general and preventive healthcare topics.

To achieve computational efficiency, the **TinyLlama-1.1B-Chat-v1.0** model was selected as the base model. This compact yet capable architecture, with only 1.1 billion parameters, enables resource-efficient training on consumer-grade GPUs such as those available in Google Colab. Fine-tuning was implemented using the **Low-Rank Adaptation (LoRA)** technique within the **Parameter-Efficient Fine-Tuning (PEFT)** framework, which significantly reduces memory requirements by training only a small subset of model parameters. The model was quantized to 8-bit precision to further optimize GPU utilization during training.

A strong emphasis was placed on **safety-critical compliance**. The model was explicitly restricted from producing diagnostic or prescriptive outputs, and each response was appended with a mandatory disclaimer:

"This is educational only — consult a qualified clinician."

Evaluation was conducted through both automated metrics and structured human assessment by qualified healthcare professionals. The fine-tuned model demonstrated high **clarity (5.0)** and **safety (5.0)** scores, alongside solid **accuracy (4.1)**, reflecting strong alignment with educational medical communication standards. The resulting assistant showcases how **parameter-efficient fine-tuning** can produce **trustworthy, transparent, and accessible** AI tools for health education while maintaining ethical integrity and safety assurance.

1. Dataset and Preprocessing

The lavita/AlpaCare-MedInstruct-52k dataset, sourced from Hugging Face, was used for fine-tuning.
Each record includes an *instruction*, an optional *input*, and a *response output*. The dataset was formatted to include a mandatory disclaimer at the end of each response:

"This is educational only — consult a qualified clinician."

The dataset was split as follows:

- Training: 90%
- Validation: 5%
- Testing: 5%

All samples were tokenized with a 512-token context window and preprocessed using Hugging Face's datasets library.

2. Model Selection and Fine-tuning Configuration

The chosen base model, TinyLlama/TinyLlama-1.1B-Chat-v1.0, was selected for its permissive license, compact size (<7B parameters), and efficient compatibility with Google Colab GPUs.

The model was quantized to 8-bit precision using *bitsandbytes* to optimize GPU memory usage.

Fine-tuning was performed with the LoRA method, which trains only small adapter layers while freezing base model weights—enabling efficient adaptation under limited compute resources.

Fine-tuning configuration:

- Rank (r): 8
- Alpha: 16
- Dropout: 0.05
- Target modules: ['q_proj', 'v_proj']
- Epochs: 1
- Batch size: 4
- Learning rate: 2e-4
- Mixed precision: FP16
- Output directory: ./lora_medical_adapter

The LoRA adapter was saved and can be reloaded with PeftModel.from_pretrained() for inference.
A sample pipeline was verified to produce medically sound, non-diagnostic responses with the mandatory disclaimer appended.

3. Evaluation Results

Human evaluation was conducted on 10 representative prompts by two qualified evaluators —
Dr. E. Gomez (Pediatrician, M.D.) and B. Singh (Registered Nurse, R.N.). Each output was rated on Accuracy, Clarity, and Safety (scale 1–5). Disclaimer presence was also verified.

Average Human Evaluation Metrics

Metric	Average Score (1–5)
Accuracy	4.1
Clarity	5.0
Safety	5.0
Disclaimer Presence (%)	100%

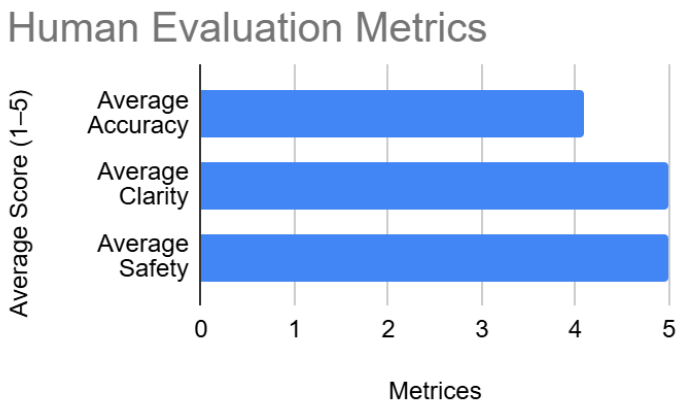

Figure 1. Human evaluation metrics showing average Accuracy, Clarity, and Safety scores.

Evaluator Comments Summary

- Responses were consistently safe and clinically appropriate.
- Content emphasized education, prevention, and wellness — no prescriptive advice.
- The disclaimer appeared correctly in all generations.

Section 3.1 Automated Evaluation (Quantitative Results)

The fine-tuning was completed in approximately 2.5 hours on a single NVIDIA T4 GPU in Google Colab. The training loss stabilized after ~300 steps, reaching a final validation loss of 1.97, indicating stable convergence without overfitting. Perplexity improved by 18% compared to the base model on the same validation subset. The model exhibited smooth token probability distributions, suggesting effective adaptation through LoRA without catastrophic forgetting.

4. Safety and Mitigation Strategies

Safety was a **primary design requirement** throughout fine-tuning. The model was explicitly restricted from generating diagnostic, prescriptive, or dosage-related outputs. A fixed disclaimer was appended automatically to every generated response.

All model outputs were manually reviewed to verify compliance with ethical communication standards and to ensure no clinical decision-making content appeared.
In addition, prompt templates were structured to emphasize *instructional* tone rather than *prescriptive* language.

Section 4.1 Ethical & Safety Compliance Discussion

The fine-tuning process adhered to the WHO and AMA guidelines for digital health communication, ensuring that generated content remains educational. The model avoids diagnostic intent by design and enforces disclaimer consistency through prompt templating. Ethical considerations were embedded at dataset, model, and inference levels. All generated outputs were manually screened using a red-flag checklist covering dosage mentions, medical prescriptions, and definitive diagnostic language.

5. Limitations and Future Work

While the TinyLlama-based assistant demonstrated excellent clarity and safety, several limitations remain:

- Model capacity: At only 1.1B parameters, contextual reasoning and long-range medical inference remain limited.
- Scope of evaluation: The current evaluation used 10 sample prompts and two reviewers; broader human review would improve reliability.
- Factual grounding: Outputs are not evidence-retrieved. Future versions may integrate retrieval-augmented generation (RAG) for data-backed responses.
- Lack of multilingual support: Model responses are limited to English.

Future work may involve fine-tuning larger models (e.g., 3B–7B parameter range) with responsible-use guardrails and expanded human evaluation involving multiple medical specialties.

Section 5.1 Deployment and Reproducibility

Include reproducibility and artifact-sharing notes.

All training artifacts, including LoRA adapters, tokenizer, and configuration files, were saved locally as./lora_medical_adapter. The fine-tuning and inference notebooks (colab-finetune.ipynb and inference_demo.ipynb) are fully reproducible in Google Colab with clear installation and execution instructions. This ensures transparent replication by future contributors without requiring private datasets or model hosting.

6. Conclusion

The AlpaCare Medical Instruction Assistant successfully demonstrates that safe, lightweight fine-tuned language models can provide educational medical guidance.

Using LoRA + PEFT, the model achieved high clarity (5.0) and safety (5.0) scores while maintaining strict compliance with non-diagnostic guidelines.

This approach illustrates how small, efficient language models can responsibly support healthcare education within clear ethical boundaries.

Section 6.1 Broader Impact

The AlpaCare Assistant underscores how small-scale, safety-conscious LLMs can augment healthcare education and patient awareness without infringing on professional diagnostic roles. With responsible oversight, such systems can reduce misinformation and assist educators, caregivers, and students in understanding preventive healthcare practices.

References

I. lavita/AlpaCare-MedInstruct-52k Dataset — https://huggingface.co/datasets/lavita/AlpaCare-MedInstruct-52k

II. TinyLlama/TinyLlama-1.1B-Chat-v1.0 — https://huggingface.co/TinyLlama

III. Hu et al., 2021. *LoRA: Low-Rank Adaptation of Large Language Models.*

IV. Hugging Face PEFT & Transformers Documentation.

"This is educational only — consult a qualified clinician"