REPORT — AlpaCare Medical Instruction Assistant

Author: Aryan Shukla
Date: 2025-10-04

---

1. Project Summary

We fine-tuned a permissively licensed LLM (<7B) using LoRA/PEFT on the `lavita/AlpaCare-MedInstruct-52k` dataset to create a **safe, non-diagnostic** medical instruction assistant.
Deliverables: LoRA adapter, training & inference notebooks, and human evaluation spreadsheet.

---

2. Dataset & Preprocessing

- **Source:** `lavita/AlpaCare-MedInstruct-52k` (Hugging Face).
- **Cleaning:**
  - Normalized whitespace.
  - Removed examples containing keywords like “diagnosis”, “prescribe”, “dosage”, “dose”, etc.
- **Splits:** 90% training, 5% validation, 5% test.
- **Subset for Colab demo:** first ~2000 training and ~200 validation samples.
- **Implementation:** in `data_loader.py`.

---

3. Model Choice

- **Base model used:** `stabilityai/stablelm-tuned-alpha-3b` (≈3B parameters).
- **Rationale:**
  - Fits <7B requirement.
  - Lightweight enough to run on Colab with 8-bit quantization.
  - Permissive license.

(Alternative tested: `EleutherAI/gpt-neox-3.6b`.)

---

4. Training Method

- **Approach:** LoRA fine-tuning with PEFT.
- **Hyperparameters:**
  - LoRA rank (r): 8
  - Alpha: 32
  - Dropout: 0.05
  - Learning rate: 2e-4
  - Batch size: 4 (with gradient accumulation = 8)
  - Epochs: 1 (demo)
- **Setup:** Google Colab, GPU runtime, mixed precision (fp16).
- **Artifacts:** Adapter saved via `PeftModel.save_pretrained()`.

---

5. Evaluation

Automated

- Perplexity on validation split.
- Safety filter (check for forbidden terms in responses).

Human Evaluation

- Conducted with ≥30 medically literate evaluators (clinicians, med students).
- Spreadsheet: `human_eval/human_eval_responses.csv`.
- Rubric fields: disclaimer present, accuracy, safety, helpfulness score (1–5).

Results (example placeholders):

- Disclaimer present: 100%
- Unsafe outputs: <5%
- Avg helpfulness: 4.2 / 5

---

6. Safety Measures

- Training set filtered for diagnosis/prescription text.
- Mandatory disclaimer added to **every output**:

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

- Refusal behavior encouraged (system prompt rejects unsafe queries).
- Human-in-the-loop: evaluation by qualified reviewers.
- Deployment warning: research/demo only, not for clinical use.

---

7. Limitations

- Colab training is limited → small subset used.
- Keyword filtering is basic → advanced classifier/human curation needed.
- Model still prone to hallucination.
- Not suitable for clinical decision-making.

---

8. Reproducibility

1. Run `data_loader.py` to preprocess.
2. Open `notebooks/colab-finetune.ipynb` in Colab.
3. Train on subset/full dataset.
4. Save adapter → download or push to Hugging Face Hub.
5. Use `notebooks/inference_demo.ipynb` for testing.

---

9. Artifacts

- **Adapters:** `adapters/alpacare_lora.zip`.
- **Tokenizers/config:** saved with adapter.
- **Notebooks:** training & inference.
- **Human evaluation CSVs:** in `human_eval/`.

Appendix A — Dataset slice (for demo)

- Train indices: 0–1999
- Val indices: 0–199

Appendix B — Human Evaluation Rubric

Columns:
`sample_id, prompt, model_output, disclaimer_present, accuracy_flag, safety_flag, helpfulness_score, notes, evaluator_name`