Project Links

Presentation Video:
https://utexas.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=25abd32e-2609-4104-abd8-b2cc0135c46f

Presentation Slides:
https://github.com/cameron-morrongiello-utexas/high-risk-project/blob/main/slides.pdf

Code Repository:
https://github.com/cameron-morrongiello-utexas/high-risk-project

Using LoRA Fine-Tuning of LLMs to Accurately Triage Patients

CAMERON MORRONGIELLO, University of Texas at Austin, USA

Efficient triage is essential to ensure patients receive the appropriate level of care without overburdening emergency resources. In this work, we explore whether a small, general-purpose language model can be fine-tuned using parameter-efficient LoRA techniques to accurately classify free-text clinical scenarios into predefined triage levels: Emergency Room, Urgent Care, Primary Care, or Self-Care. We generated a synthetic dataset of patient presentations, incorporating demographics, symptoms, onset timing, and medical history, and fine-tuned the SmolLM2-360M model for one epoch using Hugging Face PEFT. Compared to the base model, the fine-tuned model achieved a twelvefold increase in accuracy (5% → 62%) and reached 100% valid answer generation, demonstrating the feasibility of low-resource, synthetic-data-driven fine-tuning for clinical decision support. Our results suggest that small LLMs, properly adapted, can assist in automating routine triage tasks, reducing clinician burden, and enhancing operational efficiency in healthcare settings.

CCS Concepts: • **Computing methodologies** → **Natural language processing**; **Transfer learning**; • **Applied computing** → **Health informatics**.

Additional Key Words and Phrases: Natural Language Processing, Patient Triage, Fine-Tuning, LoRA, Clinical Decision Support, Synthetic Data

ACM Reference Format:
Cameron Morrongiello. 2025. Using LoRA Fine-Tuning of LLMs to Accurately Triage Patients. 1, 1 (April 2025), 6 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Proper triage is vital to a effective emergency and outpatient care. When patients present with a spectrum of symptoms—from benign colds to life-threatening cardiac events—nurses must rapidly assign them to the correct level of care (Emergency Room, Urgent Care, Primary Care, or Self-Care). Over-triage wastes scarce ER resources, while under-triage can delay critical interventions. Automating initial triage with NLP systems promises to (1) reduce nurse workload, (2) standardize decisions, and (3) flag high-risk cases for expedited review. Recent advances in large language models (LLMs) suggest that, even without extensive domain-specific pretraining, lightweight models fine-tuned on task-specific data can grasp complex, context-dependent medical reasoning. Here, we investigate whether LoRA-based fine-tuning of a small decoder-only LLM (SmolLM2-360M) on fully synthetic patient scenarios can yield a reliable triage recommender—potentially enabling "first pass" automation that frees nurses to focus on nuanced cases.

2 Related Work

Tahayori et al. (2021) applied a BERT model to free-text triage notes from an academic emergency department to predict patient disposition (admit vs. discharge). Their model attained 83% accuracy and AUC = 0.88, demonstrating

Author's Contact Information: Cameron Morrongiello, cameron.morrongiello@utexas.edu, University of Texas at Austin, Austin, Texas, USA.

that clinical text alone carries strong signals for resource needs. They highlight how early computational triage flags could streamline patient flow and reduce clinician burden [1].

Levra et al. (2024) fine-tuned two BERT variants (Italian and multilingual) on >15 000 ED records to distinguish syncope from other syncope-mimicking conditions. The “triage-only” model (using just the brief triage note) achieved AUC $\approx$ 0.95, rising to AUC $\approx$ 0.98 when full anamnesis was included, illustrating the power of LLM-based classifiers in nuanced differential diagnoses [2].

Porto’s systematic review (2024) aggregated results from 60+ ML/NLP studies in ED triage, concluding that models combining structured data and NLP-derived features consistently outperform structured-only approaches. He reports that modern architectures (gradient boosting, transformer-based) show the greatest classification gain, underscoring the value of unstructured text in triage support systems [3].

Our work differs in two key aspects: (a) we employ parameter-efficient. LoRA fine-tuning of a small, general-purpose LLM rather than training from scratch or fully fine-tuning large models; and (b) we generate a synthetic dataset to explore feasibility without requiring access to sensitive patient records.

3 Methodology

3.1 Synthetic Data Generation

To simulate diverse triage scenarios, we wrote a Python generator that produces 5000 train + 100 test cases. Each case includes:

- `age`: uniformly random between 1–89
- `gender`: male, female, or non-binary
- `symptom`: one of 10 composite phrases (e.g., “chest pain and shortness of breath”)
- `onset`: categorical (“just now” . . . “last week”)
- `medical history`: one of six (e.g., “History of heart disease.”)

We map keywords in the symptom to an initial triage level and rationale (e.g., “chest pain” → ER). We then apply secondary rules:

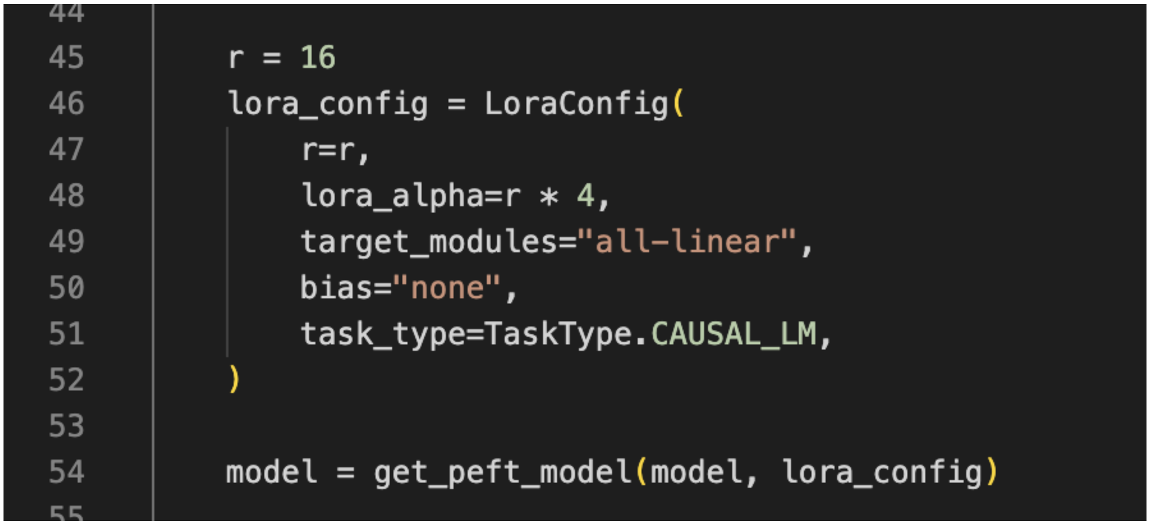
- **Time-based**: immediate onsets (“just now”) escalate to ER; longer durations (“2 days ago”) can downgrade urgent cases to primary care.
- **History-based**: e.g., heart disease with chest pain forces ER; diabetes with polyuria remains primary care.

Each output consists of a reasoning sentence followed by the final level:

```
{
  "input": "A 72-year-old female reports chest pain and shortness of breath.
  Symptoms started 2 hours ago. History of heart disease.",
  "output": "Chest pain could indicate a heart attack or other serious
  cardiac issues. Given the patient’s history of heart disease, chest pain
  must be addressed immediately. Therefore, the appropriate triage level is: Emergency Room."
}
```

3.2 LoRA Fine-Tuning

We selected the SmolLM2-360M-Instruct model for its small size (easy to run on a single GPU). Using Hugging Face's PEFT library, we applied a LoRA adapter.

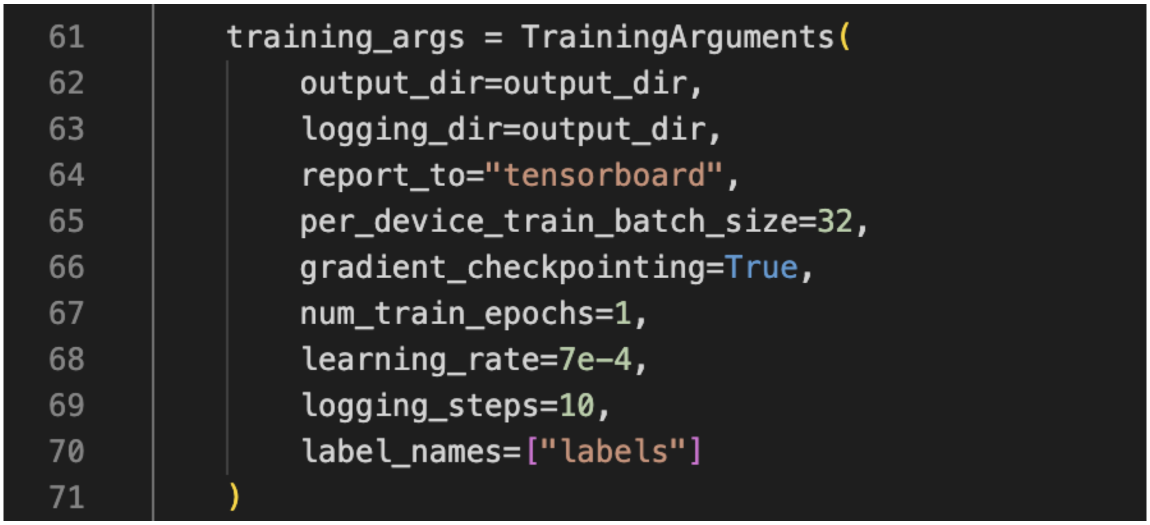


```
r = 16
lora_config = LoraConfig(
    r=r,
    lora_alpha=r * 4,
    target_modules="all-linear",
    bias="none",
    task_type=TaskType.CAUSAL_LM,
)

model = get_peft_model(model, lora_config)
```

Fig. 1. LoRA Adapter Arguments

We fine-tuned for 1 epoch, batch size 32, learning rate 7×10^{-4}, and enabled gradient checkpointing to reduce memory use.



```
training_args = TrainingArguments(
    output_dir=output_dir,
    logging_dir=output_dir,
    report_to="tensorboard",
    per_device_train_batch_size=32,
    gradient_checkpointing=True,
    num_train_epochs=1,
    learning_rate=7e-4,
    logging_steps=10,
    label_names=["labels"]
)
```

Fig. 2. Training Arguments

3.3 Evaluation Pipeline

We prepended every prompt with:

> *Possible triage levels: Emergency Room, Urgent Care, Primary Care, Self-Care.*

This was to give the non-fine-tuned base model a list of possible triage levels to answer with. We used a batched HuggingFace (HF) pipeline for generation (with `do_sample=False`, for greedy sampling). We also shuffled the parts of the input sentences in random order. We extracted the first recognized triage level and computed the following metrics:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of cases}}$$

$$\text{Answer Rate} = \frac{\text{Number of cases with any valid triage level}}{\text{Total number of cases}}$$

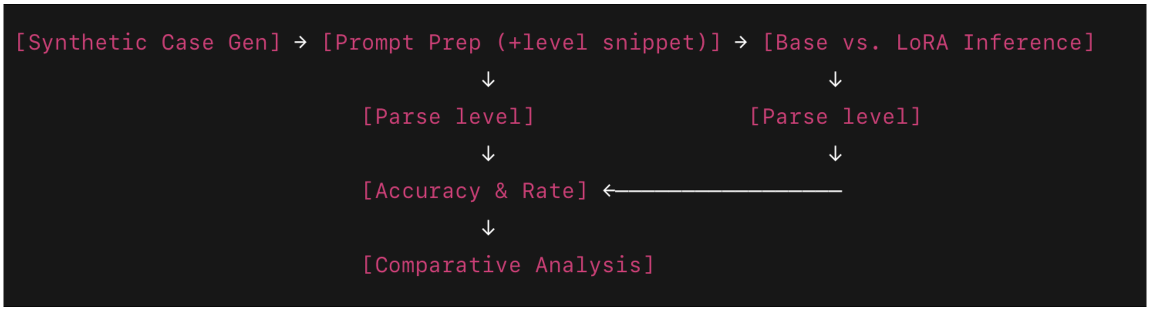


Fig. 3. Evaluation Pipeline

4 Results

Using 100 test samples, the base model seldom produced a valid triage level (33% answer rate) and was almost always incorrect (5% accuracy). After LoRA fine-tuning, the model reliably produced valid outputs (100% answer rate) and correctly predicted the level 62% of the time — an over 12× increase in accuracy.

Table 1. Accuracy and answer rate of the base and fine-tuned models.

Model	Accuracy	Answer Rate
Base Model	5.00%	33.00%
Fine-Tuned Model	62.00%	100.00%

5 Conclusion and Future Work

Our experiments demonstrate that parameter-efficient LoRA fine-tuning of a small, general-purpose LLM on synthetic triage data can yield a practical decision-support system. The fine-tuned model not only learned to output valid triage categories 100% of the time but also achieved clinically meaningful accuracy gains. Limitations include the use of synthetic cases (which may lack the nuance of real clinical documentation) and evaluation on a narrow symptom set. Future work will:

- Incorporate real-world clinical notes (e.g., MIMIC-III) under proper data governance
- Compare models pre-trained on medical text (e.g., BioBERT, ClinicalBERT)
- Extend inputs to include vital signs, labs, and imaging summaries
- Assess performance on retrospective ED or urgent care logs

With these steps, we aim to develop an end-to-end NLP triage assistant that can meaningfully augment nurse workflows, reduce wait times, and optimize resource allocation in real healthcare settings.

References

[1] Tahayori, B., Rahman, M. A., and Morita, P. P. "Advanced Natural Language Processing Technique to Predict Patient Disposition Based on Emergency Triage Notes." *Academic Emergency Medicine*, 28(9):1024–1032, 2021.

[2] Levra, A., Cordina, A., Corradi, F., et al. "A Large Language Model-Based Clinical Decision Support System for Syncope Recognition in the Emergency Department: A Framework for Clinical Workflow Integration." *Journal of Medical Systems*, 48(4):94, 2024.

[3] Porto, G., Loria, G., and Tortajada, S. "Improving Triage Performance in Emergency Departments Using Machine Learning and Natural Language Processing: A Systematic Review." *BMC Emergency Medicine*, 24(1):12, 2024.

Received 27 April 2025