Project Progress Report

By: Jeremy Flagg

Spring Semester 2025

1. Study Overview

This study focuses on:

- Training Merjek AI models on a GPU cluster.

2. Early Steps & Prompt Generation

The initial phase involved testing different LLM models for prompt generation and analyzing their outputs after database insertion.

3. Models Tested

Several models were tested for effectiveness and performance:

- Open-source models (e.g., LLama 3.1 8B, DeepSeek R1 1.5B, Mistral 7B v0.3)

4. GPU Cluster Specifications

Cluster Quota specifications:

- Max Jobs: 6
- Max Nodes: 3
- Max GPUs per Job: 4
- Max Runtime per Job: 48 hours

Training Progress:(1/24)

- **Initial meeting**

Training Progress: (2/7)

- Installation of Ollama and different open-source LLM models.
- Prompt generation and insertion into MySQL Workbench.

Training Progress:(2/14-2/28)

- Initial training/test practice, locally and in GPU Cluster,with Human Trafficking and Campus csv files.
- Migration to MongoDB Atlas/Compass

Training Progress: (3/7/25)

- Dataset: 2,000 documents (subset of 10K)
- Split: 80% train, 20% test
- Tested on 2 GPUs (1 node)
- Estimated training time: ~52 minutes for 1 epoch

Training Progress: (3/14/25)

- Created Merjek Github
- Meeting at library helping Md with MongoDB setup and prompt generation

- Continue generating ~8K prompts for the entire dataset of ~10K documents.
- Mistral 7B v0.3 is the model used for prompt generation. (LM Studio on my Windows setup)

- After generation, iterated through MongoDB collection to add prompts into arrays.

- Edit Slurm training script before executing within GPU cluster.

- Scaled training from 2,000 docs at 1 epoch to 10,000 docs at 3 epochs.

✅ Loaded 305835 valid prompts from the first 10,000 documents.

Training samples: 244668

Validation samples: 61167

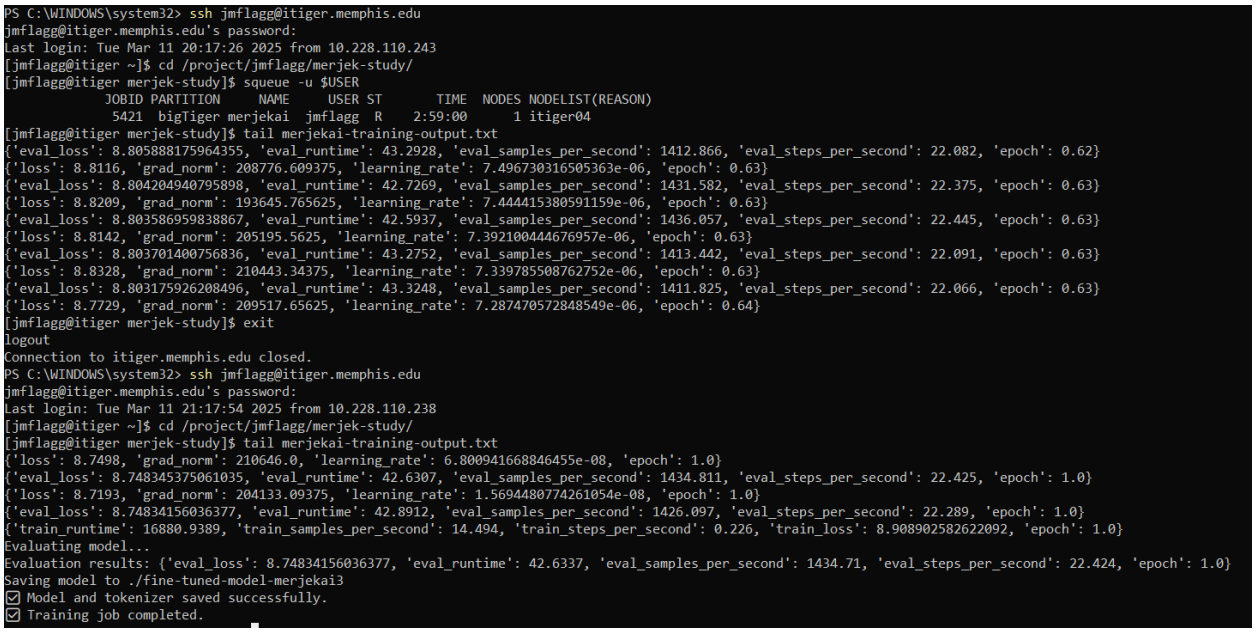
🖥️ Using device: cuda, Batch size: 16

GPU #: 4

Estimated train time for 1 epoch: 4 hours 41 minutes

Estimated train time for 3 epochs: 14 hours 4 minutes

View inside cluster after 1 epoch for 10K docs:



```
PS C:\WINDOWS\system32> ssh jmflagg@itiger.memphis.edu
jmflagg@itiger.memphis.edu's password:
Last login: Tue Mar 11 20:17:26 2025 from 10.228.110.243
[jmflagg@itiger ~]$ cd /project/jmflagg/merjek-study/
[jmflagg@itiger merjek-study]$ squeue -u $USER
             JOBID PARTITION     NAME     USER ST       TIME  NODES NODELIST(REASON)
              5421  bigTiger merjekai  jmflagg  R    2:59:00      1 itiger04
[jmflagg@itiger merjek-study]$ tail merjekai-training-output.txt
{'eval_loss': 8.805888175964355, 'eval_runtime': 43.2928, 'eval_samples_per_second': 1412.866, 'eval_steps_per_second': 22.082, 'epoch': 0.62}
{'loss': 8.8116, 'grad_norm': 208776.609375, 'learning_rate': 7.496730316505363e-06, 'epoch': 0.63}
{'eval_loss': 8.804204940795898, 'eval_runtime': 42.7269, 'eval_samples_per_second': 1431.582, 'eval_steps_per_second': 22.375, 'epoch': 0.63}
{'loss': 8.8209, 'grad_norm': 193645.765625, 'learning_rate': 7.444415380591159e-06, 'epoch': 0.63}
{'eval_loss': 8.803586959838867, 'eval_runtime': 42.5937, 'eval_samples_per_second': 1436.057, 'eval_steps_per_second': 22.445, 'epoch': 0.63}
{'loss': 8.8142, 'grad_norm': 205195.5625, 'learning_rate': 7.392100444676957e-06, 'epoch': 0.63}
{'eval_loss': 8.803701400756836, 'eval_runtime': 43.2752, 'eval_samples_per_second': 1413.442, 'eval_steps_per_second': 22.091, 'epoch': 0.63}
{'loss': 8.8328, 'grad_norm': 210443.34375, 'learning_rate': 7.339785508762752e-06, 'epoch': 0.63}
{'eval_loss': 8.803175926208496, 'eval_runtime': 43.3248, 'eval_samples_per_second': 1411.825, 'eval_steps_per_second': 22.066, 'epoch': 0.63}
{'loss': 8.7729, 'grad_norm': 209517.65625, 'learning_rate': 7.287470572848549e-06, 'epoch': 0.64}
[jmflagg@itiger merjek-study]$ exit
logout
Connection to itiger.memphis.edu closed.
PS C:\WINDOWS\system32> ssh jmflagg@itiger.memphis.edu
jmflagg@itiger.memphis.edu's password:
Last login: Tue Mar 11 21:17:54 2025 from 10.228.110.238
[jmflagg@itiger ~]$ cd /project/jmflagg/merjek-study/
[jmflagg@itiger merjek-study]$ tail merjekai-training-output.txt
{'loss': 8.7498, 'grad_norm': 210646.0, 'learning_rate': 6.800941668846455e-08, 'epoch': 1.0}
{'eval_loss': 8.748345375061035, 'eval_runtime': 42.6307, 'eval_samples_per_second': 1434.811, 'eval_steps_per_second': 22.425, 'epoch': 1.0}
{'loss': 8.7193, 'grad_norm': 204133.09375, 'learning_rate': 1.5694480774261054e-08, 'epoch': 1.0}
{'eval_loss': 8.74834156036377, 'eval_runtime': 42.8912, 'eval_samples_per_second': 1426.097, 'eval_steps_per_second': 22.289, 'epoch': 1.0}
{'train_runtime': 16880.9389, 'train_samples_per_second': 14.494, 'train_steps_per_second': 0.226, 'train_loss': 8.908902582622092, 'epoch': 1.0}
Evaluating model...
Evaluation results: {'eval_loss': 8.74834156036377, 'eval_runtime': 42.6337, 'eval_samples_per_second': 1434.71, 'eval_steps_per_second': 22.424, 'epoch': 1.0}
Saving model to ./fine-tuned-model-merjekai3
☑ Model and tokenizer saved successfully.
☑ Training job completed.
```

View inside cluster after 3 epochs for 10K:

```
[jmflagg@itiger merjek-study]$ head merjekai-training-output.txt
🚀 Starting merjekai.py...
🚀 Starting merjekai.py...
☑ Connected to MongoDB Atlas successfully.
☑ Loaded 305835 valid prompts from the first 10,000 documents.
Training samples: 244668
Validation samples: 61167
💻 Using device: cuda, Batch size: 16
Starting training...
{'loss': 9.2462, 'grad_norm': 176753.75, 'learning_rate': 1.998256168802866-05, 'epoch': 0.0}
{'eval_loss': 9.237972259521484, 'eval_runtime': 43.31, 'eval_samples_per_second': 1412.307, 'eval_steps_per_second': 22.073, 'epoch': 0.0}
[jmflagg@itiger merjek-study]$ tail merjekai-training-output.txt
{'loss': 8.2178, 'grad_norm': 242138.140625, 'learning_rate': 3.3132792745662224e-08, 'epoch': 3.0}
{'eval_loss': 8.314220428466797, 'eval_runtime': 43.37, 'eval_samples_per_second': 1410.351, 'eval_steps_per_second': 22.043, 'epoch': 3.0}
{'loss': 8.2782, 'grad_norm': 253271.640625, 'learning_rate': 1.5694480774261054e-08, 'epoch': 3.0}
{'eval_loss': 8.314230918884277, 'eval_runtime': 43.1083, 'eval_samples_per_second': 1418.914, 'eval_steps_per_second': 22.177, 'epoch': 3.0}
{'train_runtime': 50719.8742, 'train_samples_per_second': 14.472, 'train_steps_per_second': 0.226, 'train_loss': 8.52074397049928, 'epoch': 3.0}
Evaluating model...
Evaluation results: {'eval_loss': 8.31423282623291, 'eval_runtime': 42.679, 'eval_samples_per_second': 1433.188, 'eval_steps_per_second': 22.4, 'epoch': 3.0}
Saving model to ./fine-tuned-model-merjekai3
☑ Model and tokenizer saved successfully.
☑ Training job completed.
[jmflagg@itiger merjek-study]$
```

View inside MongoDB Compass: