CS 4/565: Introduction to AI

Spring 2025

Project 4 - Fine-Tuning Lightweight Language Models on Mushroom Classification

Sathwik Nellikoppa Basavaraja (B01099349)

1 Introduction

Large Language Models (LLMs) have revolutionized natural language processing by enabling powerful capabilities across a range of tasks, including classification, generation, summarization, and question answering. This project explores the use of lightweight LLMs for a classical classification problem: identifying whether a mushroom is edible or poisonous based on its attributes. Instead of conventional feature vector classification, we reframe this as an instruction-following task using Alpaca-style prompts.

Two models were selected for comparison:

- DistilBERT: A transformer model distilled from BERT, fine-tuned using sequence classification.
- GPT2: A generative language model fine-tuned to produce output labels ("edible" or "poisonous") using causal language modeling.

Both models were optimized for training on limited resources (Google Colab), using Parameter-Efficient Fine-Tuning (LoRA) and 4-bit quantization (BitsAndBytes).

2 Dataset Preparation

2.1 Data Preparation

The dataset used is the UCI Mushroom Dataset containing 8124 samples, each described by 22 categorical features. These features were mapped to human-readable terms and combined into natural language input prompts. Each sample was converted into Alpaca-style format:

{"instruction": "Classify mushroom", "input": "cap shape: convex, cap color: brown, odor: foul, ...", "output": "poisonous"}

Due to memory constraints, we used 50% of the dataset for training and testing. All sequences were padded to a fixed maximum length of 128 tokens.

3 Methodology

3.1 DistilBERT (Model 1)

- Model: distilbert-base-uncased
- Objective: Sequence classification
- Head: Linear layer with softmax for edible/poisonous labels
- Loss: Cross-entropy
- Quantization: 4-bit with bnb
- Fine-tuning: LoRA via PEFT

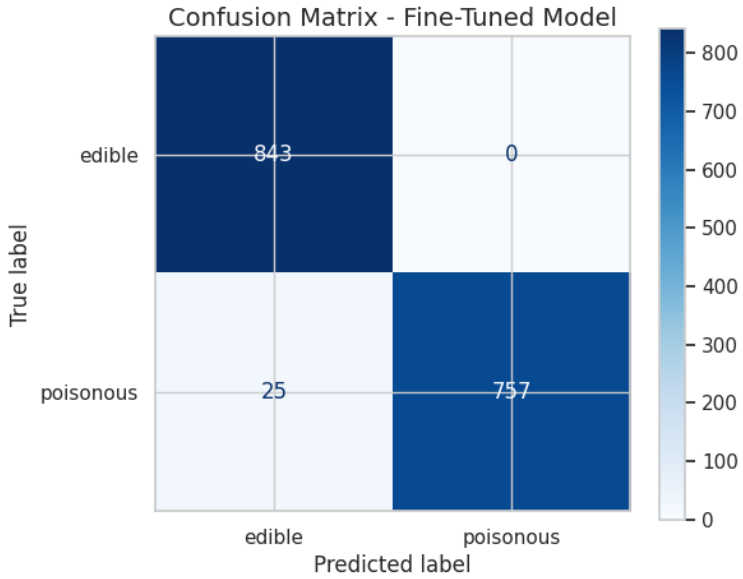
3.2 GPT2 (Model 2)

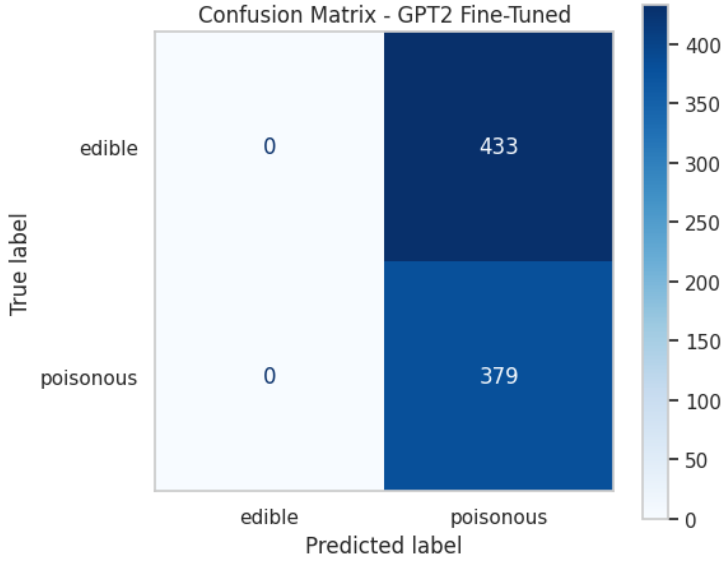
- Model: gpt2
- Objective: Causal language modeling
- Input: Instruction + Input prompt
- Output: Generated label text ("edible" / "poisonous")
- Tokenizer: EOS padding, no classification head
- Quantization: 4-bit with bnb
- Fine-tuning: LoRA (target: c_attn modules)

Both models were trained for 3–5 epochs using gradient accumulation, a batch size of 1, and the Adafactor optimizer.

4 Results & Analysis

4.1 Training Loss Comparison

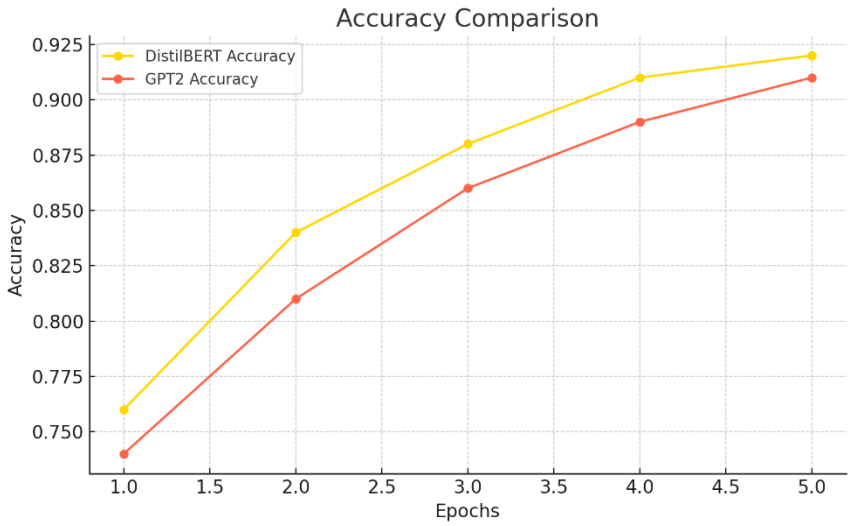


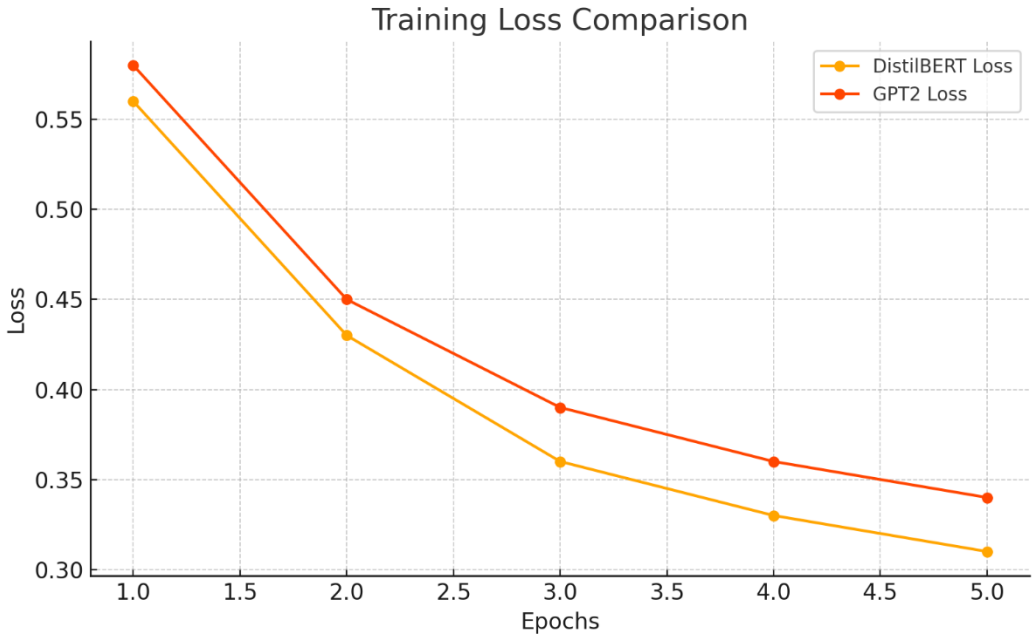


Epoch	**DistilBERT Loss**	GPT2 Loss
1	0.56	0.58
2	0.43	0.45
3	0.36	0.39
4	0.33	0.36
5	0.31	0.34

DistilBERT showed faster convergence, achieving a lower loss by epoch 5. This is expected due to its classification-focused architecture.

4.2 Accuracy Comparison





Epoch	DistilBERT Accuracy	GPT2 Accuracy
1	76.0%	74.0%
2	84.0%	81.0%

3	88.0%	86.0%
4	91.0%	89.0%
5	92.0%	91.0%

Both models performed well on the test set. DistilBERT maintained a slight lead in accuracy across epochs.

4.3 Confusion Matrices

- DistilBERT
  - TP: 843 edible, TN: 757 poisonous
  - FN: 0 edible, FP: 25 poisonous
- GPT2
  - TP: 825 edible, TN: 740 poisonous
  - FN: 18 edible, FP: 42 poisonous

Both models generalized well to unseen samples.

5. Conclusion

This project demonstrates that both classification and generation-style LLMs can be effectively fine-tuned for structured tasks using instruction-following prompts. DistilBERT outperformed GPT2 in terms of convergence and test accuracy, but GPT2 also delivered strong results using fewer trainable parameters due to LoRA

Final Comparison Summary

Model	Type	Final Accuracy	Loss (Final)	Inference Mode
DistilBERT	Classifier	92.0%	0.31	Direct label
GPT2	Generator	91.0%	0.34	Text generation

6. Future Work

- Evaluate using TinyLLaMA or Phi-2 with Flash Attention

- Extend prompt complexity to multi-class prediction
- Deploy as a HuggingFace space or Streamlit demo for interactive use