Fine-Tuned Small LLMs vs. Zero-Shot Generative AI in Text Classification

Jason Gillette

University of Texas at San Antonio

October 21st, 2024

Introduction

- Paper: *FINE-TUNED SMALL LLMS (STILL) SIGNIFICANTLY OUTPERFORM ZERO-SHOT GENERATIVE AI MODELS IN TEXT CLASSIFICATION*
- Authors: Martin Juan José Bucher (Stanford), Marco Martini (University of Zurich)
- **Research Focus**:
  - Smaller fine-tuned models vs. zero-shot generative models (e.g., ChatGPT, Claude Opus)

Problem Statement

- **Generative AI's Promise**: No need for task-specific fine-tuning and labeled data.
- **Research Question**:
  - Can generative AI models like ChatGPT outperform fine-tuned small LLMs?
- **Importance**: Understanding trade-offs between generative and fine-tuned models.

Background on Fine-Tuned LLMs

- **Pre-training and Fine-tuning**:
  - Smaller models RoBERTa Base, RoBERTa Large, DeBERTa V3, Electra Large, and XLNet are pre-trained on large corpora.
  - Fine-tuning adapts models to specific tasks like sentiment analysis.
- **Previous State-of-the-Art**:
  - Fine-tuned LLMs outperform traditional methods like BoW and pre-transformer architectures.
  - Some generative models have outperformed fine-tuned models, yet empirical evidence is not conclusive.

Zero-Shot Generative Models

- **Generative AI Models**:
  - GPT-3.5, GPT-4, Claude Opus, and BART prompt-based models without further training.
- **Zero-Shot Approach**:
  - No labeled training data required.
  - *Q: The sentiment of <text> is... A:*
  - Generates outputs based on pre-trained *general* knowledge.
- **Appeal**: Simplicity and scalability.

Methodology – Overview

- **Task Setup**:
  - Four text classification tasks: sentiment analysis, stance classification (approval/disapproval), emotion detection, multi-class text classification.
  - Datasets: News, tweets, speeches, political texts (English & German).
- **Model Comparison**:
  - Fine-tuned small LLMs vs. zero-shot generative models.
- **Metrics**: Accuracy, Precision, Recall, F1-scores.

Case Study 1: Sentiment Analysis

- **Task**: Classify positive/negative sentiment in The New York Times articles.
- **Results**:
  - Fine-tuned models (RoBERTa, DeBERTa): ~90% accuracy.
  - Zero-shot models (ChatGPT, Claude): ~82-87% accuracy.
- **Key Insight**: Fine-tuning captures sentiment nuances better.

Case Study 1: Sentiment Analysis Results

Table 1: Results for Sentiment Analysis (US Economy)

Model Name	Accuracy	Prec. (wgt.)	Recall (wgt.)	F1 (macro)	F1 (wgt.)
MAJ-VOT	0.73 (±0.00)	0.53 (±0.00)	0.73 (±0.00)	0.42 (±0.00)	0.61 (±0.00)
ROB-BASE	0.89 (±0.00)	0.89 (±0.01)	0.89 (±0.00)	0.86 (±0.01)	0.89 (±0.01)
ROB-LRG	**0.92** (±0.01)	**0.92** (±0.01)	**0.92** (±0.01)	**0.90** (±0.01)	**0.92** (±0.01)
DEB-V3	0.92 (±0.02)	0.92 (±0.01)	0.92 (±0.02)	0.90 (±0.02)	0.92 (±0.01)
ELE-LRG	0.90 (±0.01)	0.90 (±0.01)	0.90 (±0.01)	0.88 (±0.02)	0.90 (±0.01)
XLNET-LRG	0.81 (±0.01)	0.85 (±0.01)	0.81 (±0.01)	0.78 (±0.01)	0.82 (±0.01)
BART-LRG	0.85 (±0.00)	0.84 (±0.00)	0.85 (±0.00)	0.80 (±0.00)	0.84 (±0.00)
GPT-3.5	0.82 (±0.00)	0.84 (±0.00)	0.82 (±0.00)	0.79 (±0.00)	0.83 (±0.00)
GPT-4	0.87 (±0.00)	0.87 (±0.00)	0.87 (±0.00)	0.84 (±0.00)	0.87 (±0.00)
CLD-OPUS	0.86 (±0.00)	0.87 (±0.00)	0.86 (±0.00)	0.83 (±0.00)	0.87 (±0.00)

Note: Results for fine-tuned models on unseen test set with $N = 200$. Results for BART, GPTs, and Claude on full data. Fine-tuned models use gradient accumulation with 8 steps and batch size 4, except DEB-V3 (batch size 2).

Roberta Large closest zero-shot model by ~5%

Case Study 2: Stance Classification

- **Task**: Classifying support/opposition in tweets about SCOTUS nomination.
- **Results**:
  - Fine-tuned models (DeBERTa, RoBERTa): ~94% accuracy.
  - Zero-shot models: Perform slightly better than baseline (50-60% accuracy).
- **Key Insight**: Zero-shot models struggle with nuanced stance classification.

Case Study 2: Stance Classification Results

Table 2: Results for Stance Classification (Nomination Approval)

Model Name	Accuracy	Prec. (wgt.)	Recall (wgt.)	F1 (macro)	F1 (wgt.)
MAJ-VOT	0.50 (±0.00)	0.25 (±0.00)	0.50 (±0.00)	0.33 (±0.00)	0.33 (±0.00)
ROB-BASE	0.86 (±0.01)	0.86 (±0.01)	0.86 (±0.01)	0.86 (±0.01)	0.86 (±0.01)
ROB-LRG	0.92 (±0.01)	0.93 (±0.01)	0.92 (±0.01)	0.92 (±0.01)	0.92 (±0.01)
DEB-V3	**0.94 (±0.01)**	**0.94 (±0.01)**	**0.94 (±0.01)**	**0.93 (±0.01)**	**0.94 (±0.01)**
ELE-LRG	0.74 (±0.01)	0.66 (±0.02)	0.74 (±0.01)	0.67 (±0.02)	0.69 (±0.02)
XLNET-LRG	0.83 (±0.01)	0.83 (±0.01)	0.83 (±0.01)	0.83 (±0.01)	0.83 (±0.01)
BART-LRG	0.53 (±0.00)	0.59 (±0.00)	0.53 (±0.00)	0.44 (±0.00)	0.44 (±0.00)
GPT-3.5	0.53 (±0.00)	0.58 (±0.00)	0.53 (±0.00)	0.48 (±0.00)	0.47 (±0.00)
GPT-4	0.58 (±0.00)	0.68 (±0.00)	0.58 (±0.00)	0.51 (±0.00)	0.51 (±0.00)
CLD-OPUS	0.61 (±0.00)	0.68 (±0.00)	0.61 (±0.00)	0.57 (±0.00)	0.57 (±0.00)

Note: Results for fine-tuned models on unseen test set with $N = 200$. Results for BART, GPTs, and Claude on full data. Fine-tuned models use gradient accumulation with 8 steps and batch size 4, except DEB-V3 (batch size 2).

DeBERTa performs over twice as well as zero-shot model.

Case Study 3: Emotion Detection

- **Task**: Detecting anger in German political texts.
- **Results**:
  - Fine-tuned models: ~88-89% accuracy.
  - Zero-shot models: Perform poorly (~15-20% accuracy).
- **Translation Experiment**: Minimal difference between German and translated English performance.
- **Key Insight**: Zero-shot models struggle with specialized tasks.

Case Study 3: Emotion Detection Results

Table 3: Results for Emotion Detection (Anger)

Model Name	Accuracy	Prec. (wgt.)	Recall (wgt.)	F1 (macro)	F1 (wgt.)
MAJ-VOT	0.71 (±0.00)	0.51 (±0.00)	0.71 (±0.00)	0.42 (±0.00)	0.59 (±0.00)
ROB-BASE	0.87 (±0.01)	0.88 (±0.01)	0.87 (±0.01)	0.82 (±0.01)	0.88 (±0.01)
ROB-LRG	0.88 (±0.01)	0.88 (±0.00)	0.88 (±0.01)	0.83 (±0.00)	0.88 (±0.00)
DEB-V3	0.88 (±0.01)	0.88 (±0.00)	0.88 (±0.01)	0.83 (±0.01)	0.88 (±0.00)
ELE-LRG	0.88 (±0.00)	0.88 (±0.02)	0.88 (±0.00)	0.84 (±0.00)	0.88 (±0.00)
XLNET-LRG	**0.89** (±0.00)	**0.89** (±0.00)	**0.89** (±0.00)	**0.85** (±0.00)	**0.89** (±0.00)
ELE-BS-GER	0.88 (±0.01)	0.88 (±0.01)	0.88 (±0.01)	0.83 (±0.02)	0.88 (±0.01)
BART-LRG	0.26 (±0.00)	0.36 (±0.00)	0.26 (±0.00)	0.24 (±0.00)	0.29 (±0.00)
GPT-3.5	0.15 (±0.00)	0.23 (±0.00)	0.15 (±0.00)	0.15 (±0.00)	0.16 (±0.00)
GPT-4	0.20 (±0.00)	0.18 (±0.00)	0.20 (±0.00)	0.18 (±0.00)	0.13 (±0.00)
CLD-OPUS	0.15 (±0.00)	0.16 (±0.00)	0.15 (±0.00)	0.14 (±0.00)	0.11 (±0.00)

Note: Results for fine-tuned models on unseen test set with $N = 200$. Results for BART, GPTs, and Claude on full data. Fine-tuned models use gradient accumulation with 8 steps and batch size 4, except DEB-V3 (batch size 2).

XLNET-Large performs 3x to 8x better than zero-shot prompting.

Case Study 4: Multi-Class Stance Classification

- **Task**: Predicting party positions on EU integration.
- **Results**:
  - Fine-tuned models: ~92% accuracy.
  - Zero-shot models struggle with multi-class classification.
- **Key Insight**: Fine-tuned models handle complex tasks better.

Case Study 4: Multi-Class Stance Classification Results

Table 4: Results for Multi-Class Stance Classification (EU Positions)

Model Name	Accuracy	Prec. (wgt.)	Recall (wgt.)	F1 (macro)	F1 (wgt.)
MAJ-VOT	0.83 (±0.00)	0.68 (±0.00)	0.83 (±0.00)	0.30 (±0.00)	0.75 (±0.00)
ROB-BASE	0.84 (±0.00)	0.87 (±0.01)	0.84 (±0.00)	0.70 (±0.02)	0.85 (±0.00)
ROB-LRG	0.88 (±0.01)	0.88 (±0.01)	0.88 (±0.01)	0.72 (±0.03)	0.87 (±0.01)
DEB-V3	**0.92 (±0.01)**	**0.91 (±0.01)**	**0.92 (±0.01)**	**0.82 (±0.02)**	**0.91 (±0.01)**
ELE-LRG	0.88 (±0.01)	0.88 (±0.01)	0.88 (±0.01)	0.75 (±0.03)	0.87 (±0.01)
XLNET-LRG	0.87 (±0.01)	0.89 (±0.01)	0.87 (±0.01)	0.75 (±0.02)	0.88 (±0.01)
BART-LRG	0.82 (±0.00)	0.77 (±0.00)	0.82 (±0.00)	0.34 (±0.00)	0.75 (±0.00)
GPT-3.5	0.24 (±0.00)	0.65 (±0.00)	0.24 (±0.00)	0.17 (±0.00)	0.27 (±0.00)
GPT-4	0.38 (±0.00)	0.73 (±0.00)	0.38 (±0.00)	0.26 (±0.00)	0.45 (±0.00)
CLD-OPUS	0.26 (±0.00)	0.75 (±0.00)	0.26 (±0.00)	0.25 (±0.00)	0.29 (±0.00)

Note: Results for fine-tuned models on unseen test set with $N = 200$. Results for BART, GPTs, and Claude on full data. Fine-tuned models use gradient accumulation with 8 steps and batch size 4, except DEB-V3 (batch size 2).

Again, DeBERTa out performs zero-shot prompting.

Impact of Training Data Size

- **Ablation Study**: Effect of varying training set size on performance.
- **Findings**:
  - Performance improves with larger training data, plateaus after ~500 samples.
  - Fine-tuned models outperform zero-shot models after just 200 samples.
- **Conclusion**: Moderate amounts of training data improve fine-tuned models significantly.

Ablation Study Results

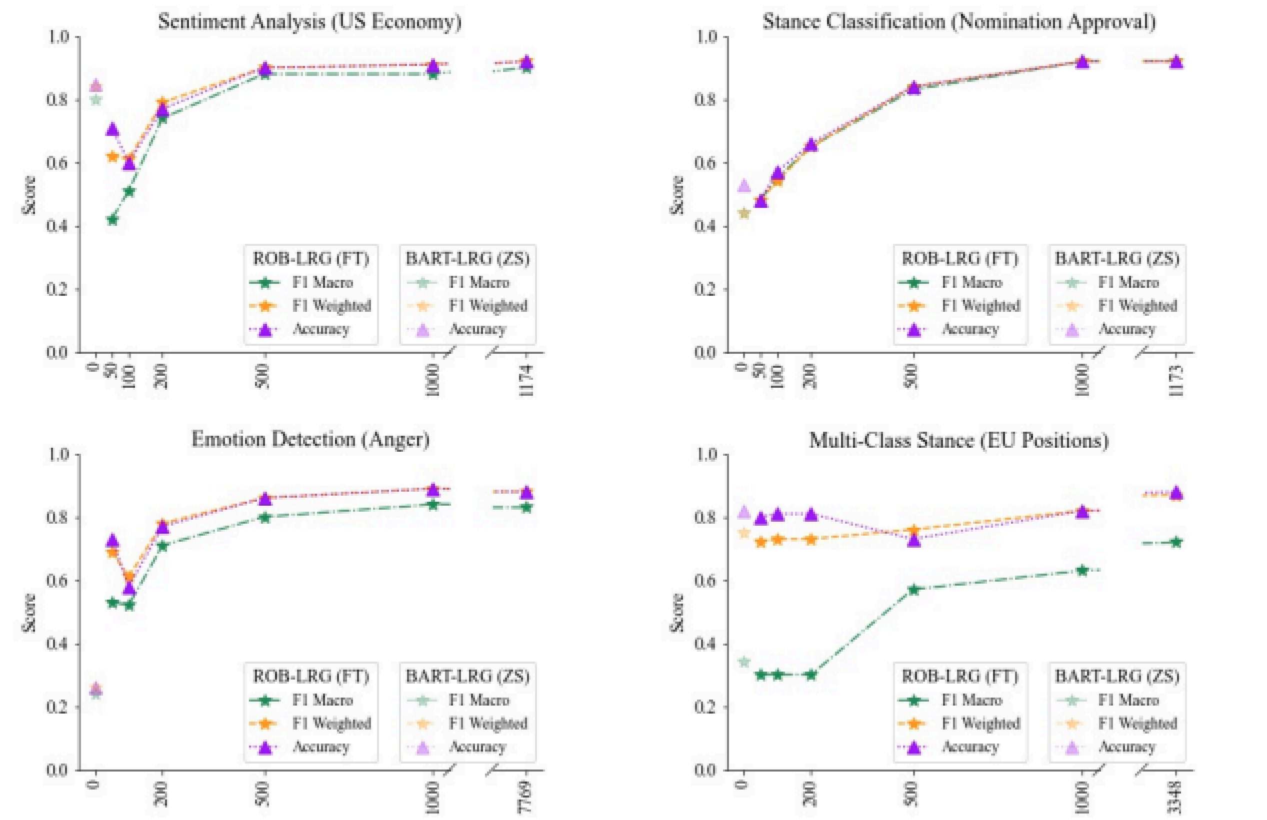

Figure 4: Effect of training set size on model performance: Results for ROB-LRG with varying number of training observations $N = \{50, 100, 200, 500, 1000\}$. The translucent markers above the 0-point denote the zero-shot results of BART. The rightmost points denote model performance if trained on the full dataset.

Why Fine-Tuning Prevails

- **Application-Specific Data**: Fine-tuned models gain task-specific knowledge.
- **Fine-Tuning Strengths**: Better at capturing nuanced distinctions.
- **Limitations of Zero-Shot Models**: Struggle with niche, specialized tasks.

Future Directions in Generative AI and Fine-Tuning

- **Few-Shot Learning**: Potential to bridge the gap between zero-shot and fine-tuned models.
- **Data Augmentation**: Techniques like back-translation and token perturbation can reduce the need for large labeled datasets.
- **Model Architecture**: Multi-modality and quality improvements in generative models.

Conclusion and Takeaways

- **Summary of Key Findings**:
  - Fine-tuning outperform zero-shot models in specialized tasks.
  - Zero-shot models are easy but struggle with domain-specific tasks.
- **Toolkit Availability**:
  - Accessible Jupyter Notebook for text classification fine-tuning.
  - Supports binary and non-binary tasks.
  - Supports class imbalances in data.
- **Final Remark**: Fine-tuned LLMs are still relevant.

Q&A

- How would few-shot prompting compare?
- Was zero-shot evaluation performed in a single prompt or per sample and how might this impact results?

Bucher, M. J. J., & Martini, M. (2024). Fine-tuned 'small' LLMs (still) significantly outperform zero-shot generative AI models in text classification. arXiv preprint arXiv:2406.08660.