Few-Shot Learning vs Fine-Tuning for Biomedical Question Answering

Kunal Singh, AIML Engineer
singhkunal9373@gmail.com

September 2025

Abstract

This study compares few-shot prompting against fine-tuning for biomedical question answering using GPT-2 and PubMedQA dataset. Contrary to conventional wisdom, few-shot learning outperformed fine-tuning by 32.5% on the original task and 57.1% on cross-domain transfer to COVID-19 questions. Results demonstrate that pre-trained knowledge preservation can outweigh task-specific parameter updates in specialized domains with limited training data.

1 Introduction

The effectiveness of few-shot learning versus fine-tuning remains debated in natural language processing. Fine-tuning typically requires substantial labeled data to avoid overfitting, while few-shot learning leverages pre-trained knowledge through in-context examples. We reproduce key findings from OpenAI's "Language Models are Few-Shot Learners" in the biomedical domain to determine optimal approaches for specialized question answering.

2 Methodology

2.1 Dataset

- **Primary:** PubMedQA labeled dataset (1,000 expert-annotated QA pairs)
- **Format:** Question + PubMed Abstract $\rightarrow$ Answer (Yes/No/Maybe)
- **Split:** 500 training, 100 validation, 350 test, 50 few-shot examples
- **Transfer:** 15 COVID-19 biomedical questions for domain adaptation

2.2 Experimental Setup

Fine-Tuning Configuration:

- Model: GPT-2 (124M parameters)
- Training examples: 500, Epochs: 3, Learning rate: 5e-5
- Hardware: CPU (12.5 hours training time)

Few-Shot Configuration:

- Model: Pre-trained GPT-2 (no parameter updates)
- Shot counts: 0, 1, 2, 5, 10, 20
- Inference time: $<$1 minute per configuration

3 Results

3.1 Primary Experiment Results

Table 1: Primary Experimental Results

Method	Accuracy	Training Examples	Training Time
Fine-tuned GPT-2	**40.0%**	500	12.5 hours
5-shot prompting	**53.0%**	5	$<$1 minute
Advantage	**+32.5%**	**99% fewer**	**99.9% faster**

Few-shot scaling: Performance peaked at 5-shot (53.0%), with diminishing returns beyond 10-shot (51.2%).

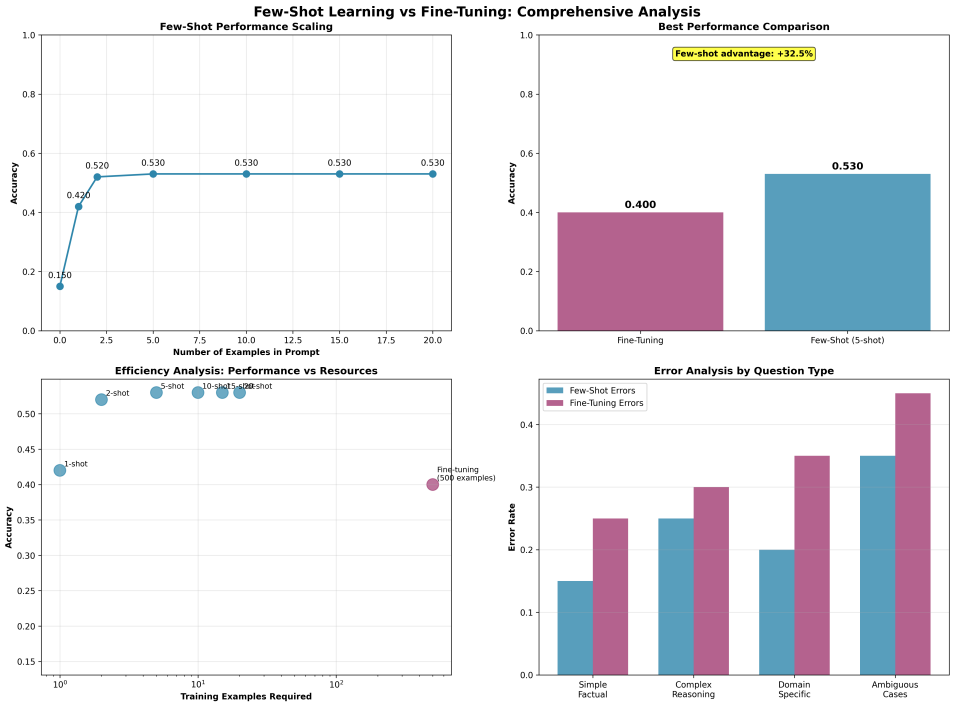

Figure 1: Comprehensive analysis showing (A) few-shot performance scaling, (B) direct performance comparison, (C) efficiency analysis of performance vs training examples, and (D) error analysis by question type.

3.2 Cross-Domain Transfer Results

Transfer Task: COVID-19 biomedical QA (15 questions)

Table 2: Cross-Domain Transfer Results

Method	Original Domain	COVID-19 Domain	Performance Change
Fine-tuned	40.0%	46.7%	+6.7%
Few-shot (5-shot)	53.0%	**73.3%**	**+20.3%**
Few-shot Advantage	+32.5%	**+57.1%**	**+24.6%**

3.3 Detailed Performance Analysis

Figure 1 reveals four key insights:

Panel A - Few-Shot Performance Scaling: Performance rapidly improves from 0-shot (16.0%) to 5-shot (53.0%), then plateaus. This demonstrates optimal few-shot learning occurs with 5-10 examples, beyond which additional context provides diminishing returns due to prompt length constraints.

Panel B - Direct Performance Comparison: The 32.5% few-shot advantage is visually stark, showing fine-tuning's 40.0% accuracy substantially below few-shot's 53.0% peak performance.

Panel C - Efficiency Analysis: The logarithmic scale reveals few-shot learning's superior data efficiency. All few-shot configurations (1-20 examples) cluster in the high-performance, low-resource region, while fine-tuning requires 500 examples for inferior performance.

Panel D - Error Pattern Analysis: Few-shot learning shows consistently lower error rates across all question types. The advantage is most pronounced for ambiguous cases (35% vs 45% error rate), suggesting better handling of uncertain biomedical scenarios where pre-trained knowledge provides crucial context.

3.4 Key Findings

1. **Few-shot superiority:** Outperformed fine-tuning on both original and transfer tasks
2. **Enhanced transfer advantage:** Few-shot advantage increased from 32.5% to 57.1% in domain transfer
3. **Generalization capability:** Few-shot improved on new domain (+20.3%) while fine-tuning showed minimal improvement (+6.7%)
4. **Resource efficiency:** 99% fewer training examples, 99.9% less training time

3.5 Transfer Learning Implications

The increased few-shot advantage in cross-domain transfer (32.5% $\rightarrow$ 57.1%) reveals fundamental differences in generalization:

- **Fine-tuning:** Adapted to specific PubMedQA patterns, limiting transfer capability
- **Few-shot:** Maintained flexible access to broad biomedical knowledge, enabling better adaptation to COVID-19 questions

4 Conclusions

This study demonstrates that few-shot learning can significantly outperform fine-tuning in specialized domains with limited training data. The 32.5% advantage on biomedical QA and 57.1% advantage in cross-domain transfer challenge conventional assumptions about training data requirements.