TINY EXPERT? ARCHITECTURAL OPTIMIZATION FOR RESOURCE-CONSTRAINED DOMAIN TASKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent advances in large language models have led to increased adoption across specialized domains, but their effectiveness on tasks with limited training data remains unclear. We investigate this question through bias detection in medical curriculum text, comparing models ranging from DistilBERT (67M parameters) to Llama-3.2 (1.2B parameters) using both sequence classification and causal language modeling approaches. Our findings challenge conventional assumptions about model scaling: while the instruction-tuned Llama achieved the strongest screening performance (AUC: 0.7904, F2: 0.5760), architectural choices proved more critical than model size. DistilBERT demonstrated competitive performance through targeted architectural choices, achieving the second-highest AUC (0.8857) despite its smaller size. These results suggest that for specialized classification tasks with limited training data, architectural alignment and instruction tuning may be more crucial than increased model capacity. Our work provides practical insights for deploying language models in domain-specific applications where expert annotation is expensive and dataset size is necessarily limited.

1 INTRODUCTION

Recent advances in large language models have led to their increased adoption across specialized domains, based on their theoretical advantages in transfer learning and multi-task capabilities Brown et al. (2020); Raffel et al. (2020); Bommasani et al. (2021). However, the effectiveness of these models on domain-specific tasks with limited training data remains unclear. We investigate this question through bias detection in medical curriculum text Salavati et al. (2024), where expert annotation is expensive and dataset size is necessarily limited.

2 METHODS

In this study, we investigated the relative importance of model architecture versus model size for bias detection in medical curriculum text using the BRICC dataset Salavati et al. (2024); Butts et al. (2024), which contains 1,530 annotated text segments with a 1:4 ratio of gender-biased to non-biased samples. Our experimental setup examined two primary axes of variation: model architecture (sequence classification versus causal language modeling) and model size (from DistilBERT at 67M parameters through Llama-3.2 at 1.2B parameters). We compared BERT-base (109M parameters) Devlin et al. (2019), DistilBERT (67M parameters), and Llama-3.2 (1.2B parameters) Touvron et al. (2023b;a) in both sequence classification and causal language modeling configurations. The BERT variants utilized bidirectional attention in their base form, while Llama employed unidirectional attention. Each model variant was trained under standardized conditions using AdamW optimization with a learning rate of 2e-5 and batch size of 32 for eight epochs, which were chosen due to their strong performance in Salavati et al.'s model.

3 RESULTS

Analysis of model performance metrics reveals critical insights when evaluating these systems as screening tools for expert review of potential bias. In this context, F2 and AUC metrics are particularly relevant, as they better capture a model's ability to minimize false negatives while main-

Table 1: Performance Metrics for Different Models

Model	Accuracy	Precision	Recall	F1	F2	AUC
BERT (Base)	0.5784	0.1747	0.3883	0.1589	0.2302	0.4803
BERT (Fine-tuned)	0.7876	0.3393	0.4043	0.3689	0.3893	0.7018
DistilBERT (Base)	0.3301	0.1141	0.7340	0.1975	0.3517	0.4680
DistilBERT (Fine-tuned)	0.8497	1.0000	0.0213	0.0417	0.0265	0.8857
Llama 3.2 1B (Base)	0.4003	0.1339	0.6383	0.2021	0.3375	0.4919
Llama 3.2 1B (Fine-tuned)	0.1855	0.1535	0.9521	0.2643	0.4664	0.5056
Llama 3.2 1B Instruct (Base)	0.8105	0.0000	0.0000	0.0000	0.0000	0.4983
Llama 3.2 1B Instruct (Fine-tuned)	0.8922	0.8375	0.5345	0.6522	0.5760	0.7904

taining reasonable precision. As shown in Table 1, the fine-tuned instruction-tuned Llama 3.2 1B model achieved the strongest overall screening performance, with an AUC of 0.7904 and F2 score of 0.5760. This represents a substantial improvement over the base BERT model's metrics (AUC: 0.4803, F2: 0.2302). DistilBERT demonstrated intriguing performance characteristics after fine-tuning, achieving the second-highest AUC (0.8857) but a notably low F2 score (0.0265), suggesting potential challenges in practical deployment as a screening tool. This disparity between AUC and F2 metrics indicates that while the model effectively ranks examples, its specific classification threshold may need adjustment for screening purposes. The standard fine-tuned BERT model achieved more balanced screening performance (AUC: 0.7018, F2: 0.3893), while the base Llama variants showed limited effectiveness (AUC: 0.50, F2: 0.3375-0.4664). These results suggest that instruction tuning may be particularly valuable for developing effective screening systems in specialized domains with limited training data.

4 Discussion

The substantial performance gap between our results and those reported by Salavati et al. highlights critical challenges in reproducing and extending work on specialized NLP tasks. A significant limitation arose from our inability to replicate their precise data filtering methodology, which selected examples containing explicit social identity markers in quoted text. Without access to these filtering criteria, our broader dataset likely included more ambiguous cases that complicated the classification task.

Additionally, the experimental process revealed inherent difficulties in diagnosing the root causes of poor model performance. When results deviated from expected benchmarks, it remained unclear whether the issues stemmed from fundamental differences in dataset composition, suboptimal training procedures, evaluation metric implementation discrepancies, or model architecture misalignment. This diagnostic uncertainty was compounded by limited documentation for newer model architectures, particularly regarding their adaptation to specialized classification tasks. While base model usage is often well-documented, guidance for configuring model heads, implementing appropriate preprocessing steps, and optimizing training hyperparameters remains sparse.

These challenges underscore the broader need for more comprehensive documentation and reproducibility protocols in specialized NLP tasks, especially when working with evolving model architectures on nuanced classification problems.

The comparative analysis between models yielded particularly intriguing insights about the relationship between model size and task performance. While conventional wisdom suggests that larger models should demonstrate superior transfer learning capabilities, our results indicate that architectural alignment with the task may be more crucial than parameter count. The strong performance of DistilBERT relative to larger models suggests that specialized classification tasks with limited training data may benefit more from targeted architectural choices than from increased model capacity.

These challenges underscore the broader need for more comprehensive documentation and reproducibility protocols in specialized NLP tasks, especially when working with evolving model architectures on nuanced classification problems.

REFERENCES

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Gavin Butts, Pegah Emdad, Jethro Lee, Shannon Song, Chiman Salavati, Willmar Sosa Diaz, Shiri Dori-Hacohen, and Fabricio Murai. Towards fairer health recommendations: finding informative unbiased samples via word sense disambiguation. *arXiv preprint arXiv:2409.07424*, 2024.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2019.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, et al. Cladder: A benchmark to assess causal reasoning capabilities of language models. *Advances in Neural Information Processing Systems*, 36, 2024.

Xiaoyu Liu, Paiheng Xu, Junda Wu, Jiaxin Yuan, Yifan Yang, Yuhang Zhou, Fuxiao Liu, Tianrui Guan, Haoliang Wang, Tong Yu, et al. Large language models and causal inference in collaboration: A comprehensive survey. *arXiv preprint arXiv:2403.09606*, 2024.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Chiman Salavati, Shannon Song, Willmar Sosa Diaz, Scott A Hale, Roberto E Montenegro, Fabricio Murai, and Shiri Dori-Hacohen. Reducing biases towards minoritized populations in medical curricular content via artificial intelligence for fairer health outcomes. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pp. 1269–1280, 2024.

V Sanh. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

Hugo Touvron, Louis Martin, Pierre Labatut, Tianqi Su, and Guillaume Lample. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023a.

Hugo Touvron, Louis Martin, Kevin Stone, Pierre-Emmanuel Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Shruti Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023b.

A APPENDIX: SUPPLEMENTARY METHODS

A.1 DATASET AND TASK DEFINITION

The BRICC dataset comprises medical curriculum text annotated for various types of bias. After initial challenges with considering the full dataset, we limited our scope to filter for examples of gender bias and a proportional number of text snippets classified as not having gender bias. Key characteristics include:

- 1,530 annotated text segments in filtered dataset
- 1:4 split; 306 examples are biased

A.2 Experimental Overview

Our investigation consists of two primary experimental tracks, each designed to test distinct hypotheses about model performance in specialized domains with limited training data. The first track examines architectural impacts on model performance, while the second investigates fine-tuning strategies for domain adaptation.

A.2.1 Architecture Comparisons

We compare sequence classification and causal language modeling approaches across different model sizes. This comparison aims to disentangle the effects of model architecture from model size in specialized domain tasks. Key experimental variables include:

- **Architecture Type**: Sequence classification vs. causal language modeling
- **Model Size**: BERT-base (109M parameters) vs. Llama-3.2 (1.2B parameters)
- **Attention Patterns**: Bidirectional (BERT) vs. unidirectional (Llama, BERT with `is_decoder=True`)

Each architectural variant was evaluated under identical training conditions to isolate architecture-specific effects. The sequence classification models implement task-specific heads optimized for binary classification, while the causal language modeling variants approach the task through next-token prediction.

A.3 Model Architectures

We investigated two primary architectural approaches using BERT-base and Llama-3.2 models: sequence classification and causal language modeling (CausalLM). While our experiments focused on these two models, we include architectural comparisons with DistilBERT, which established the baseline results on the BRICC dataset in prior work.

The sequence classification variants implement task-specific classification heads differently across model families. BERT applies linear layers to its encoder outputs, while Llama performs classification on the final token's representation, following the approach used by other causal models like GPT-2 Radford et al. (2019). For context, DistilBERT's architecture Sanh (2019), which achieved the baseline results, applies linear layers similar to BERT but includes an additional pre-classification transformation. These architectural differences reflect their underlying designs: BERT and DistilBERT as bidirectional encoders (with DistilBERT using half the layers of BERT) and Llama as a decoder-only model.

For causal language modeling, BERT requires architectural modification through the addition of a language modeling head and conversion to unidirectional attention (using `is_decoder=True`), while Llama maintains its native decoder-only architecture. BERT's language modeling head includes a dedicated transformation layer followed by vocabulary projection, adding substantial parameters beyond the base model. These architectural choices significantly impact how each model approaches the classification task: BERT's sequence classification variant directly optimizes for classification through transformed representations, while the CausalLM variants must adapt to a fundamentally different prediction task Jin et al. (2024); Liu et al. (2024). Detailed specifications for each variant are provided in Appendix B.

A.4 Training Methodology

Our training approach incorporates several key methodological elements designed to ensure robust comparison across experimental conditions while maintaining practical applicability.

A.4.1 Base Configuration

All models were trained using the following base configuration:

- **Optimizer**: AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$

- **Learning Rate**: 2e-5 with linear warmup over 100 steps
- **Batch Size**: 32 samples
- **Gradient Accumulation**: 4 steps
- **Weight Decay**: 0.01
- **Epochs**: 8

The training duration was determined through preliminary experiments showing convergence patterns across architectures, with additional steps allocated to account for potential late-stage optimization.

A.4.2 Architecture-Specific Adaptations

Training procedures were adapted for each architectural variant while maintaining fundamental consistency:

- **Sequence Classification**:
  - Binary cross-entropy loss
  - Classification token (`[CLS]`) pooling for BERT
  - Final token representation for Llama
  - Dropout rate of 0.1 maintained across models
- **Causal Language Modeling**:
  - Next-token prediction loss
  - Causal attention masking
  - Label token appending for classification
  - Temperature scaling ($\tau = 0.7$) during inference

B Appendix: Model Architecture Details

B.1 Sequence Classification Architectures

The BERT sequence classification model (BertForSequenceClassification, 109.5M parameters) augments the base BERT encoder with a classification head on top of the pooled output. The base encoder comprises 12 transformer layers with a hidden dimension of 768 and 12 attention heads. Its embedding layer handles 30,522 tokens and includes both positional and token type embeddings, with layer normalization ($\epsilon = 10^{-12}$) and dropout ($p = 0.1$) for regularization.

The DistilBERT sequence classification model (DistilBertForSequenceClassification, 67.0M parameters) follows a similar architecture but achieves parameter efficiency through architectural distillation. It uses 6 transformer layers instead of BERT's 12 while maintaining the same hidden dimension of 768. The model adds a pre-classification transformation layer and uses a higher dropout rate ($p = 0.2$) for the classification head. Both the pre-classifier and classifier are linear transformations, with the pre-classifier maintaining the 768-dimensional representation before the final classification layer.

The Llama sequence classification model (LlamaForSequenceClassification, 1.24B parameters) takes a different approach, performing classification on the last non-padding token's representation. This design choice aligns with other causal models like GPT-2, reflecting Llama's decoder-only architecture. The model comprises 16 transformer layers with a hidden dimension of 2048, uses RMSNorm ($\epsilon = 10^{-5}$) for normalization, and employs rotary positional embeddings. When no padding token is defined, the model defaults to using the final sequence token for classification.

B.2 Causal Language Modeling Architectures

The BERT causal language modeling variant (BertLMHeadModel) modifies the base architecture for CLM fine-tuning by adding a language modeling head and enabling decoder-style processing. The language modeling head consists of a transformation layer maintaining the 768-dimensional

representation, followed by GELU activation, layer normalization, and final projection to the vocabulary size. This architectural modification fundamentally changes how BERT processes sequences, converting its native bidirectional attention to unidirectional attention.

DistilBERT maintains a similar relationship to its base model when adapted for causal language modeling, though with its reduced layer count providing significant parameter efficiency. The adaptation process follows the same pattern as BERT, with the addition of a language modeling head and conversion to unidirectional attention.

Llama's causal language modeling variant (LlamaForCausalLM) represents its natural form as a decoder-only model, using scaled dot-product attention with rotary embeddings. The attention mechanism maintains the model's 2048-dimensional queries while using reduced 512-dimensional keys and values. The model includes a gated MLP with SiLU activation and 8192-dimensional intermediate representations, projecting finally to the vocabulary size of 128,256.