Accuracy-Efficiency Trade-offs in Resume Classification: A Comparative Study of Lightweight Transformers and Parameter-Efficient Fine-Tuning

Nezar Abdilah Prakasa

Abstract: The growing reliance on online recruitment platforms has increased the need for automated resume classification systems that are both accurate and computationally efficient. Prior work has shown that large language models (LLMs) such as Gemma 1.1 2B can achieve strong accuracy on this task, but their computational cost limits practical deployment for organizations with constrained resources. This study examines whether compact transformer encoders, trained under full fine-tuning and Low-Rank Adaptation (LoRA), can match the performance of such large models on a 43-category resume classification task using the ResumeAtlas dataset. We evaluate MobileBERT (25M parameters), DistilBERT (66M parameters), and ModernBERT-base (150M parameters), reporting accuracy, macro F1-score, trainable parameter ratio, training time, and inference latency. Results show that fully fine-tuned compact models, including one as small as 25M parameters, match or exceed the 92% top-1 accuracy reported for a 2-billion-parameter LLM baseline while using up to two orders of magnitude fewer parameters. However, LoRA fine-tuning with a uniform default configuration (rank = 8) shows a substantial and irregular accuracy gap relative to full fine-tuning across all three models, with no consistent relationship to model size, indicating that fixed LoRA hyperparameters may not transfer uniformly across compact encoder architectures.

Keywords: Resume Classification, Parameter-Efficient Fine-Tuning, LoRA, Lightweight Transformers, Human Resource Analytics.

I. INTRODUCTION

Online recruitment platforms receive large volumes of resumes on a daily basis, making manual categorization into job-relevant categories impractical at scale. Automated resume classification, the task of assigning each resume to one of several predefined occupational categories, has therefore become an important component of applicant tracking and candidate recommendation systems [10].

Recent work has approached this task using increasingly large language models. Heakl et al. [1] introduced the ResumeAtlas dataset, comprising 13,389 resumes spanning 43 categories, and showed that a fine-tuned 2-billion-parameter LLM (Gemma 1.1 2B [8]) can reach a top-1 accuracy of 92%. While effective, such models require substantial GPU memory and inference cost, which can be impractical for smaller organizations or for high-throughput screening pipelines.

An alternative direction is to use compact transformer encoders, which require far fewer parameters but have historically lagged behind larger models in accuracy. Parameter-efficient fine-tuning (PEFT) methods such as Low-Rank Adaptation (LoRA) [2] further reduce training cost by updating only a small subset of model parameters, while keeping the bulk of the pretrained weights frozen.

This study addresses two questions. RQ1: Can compact encoder models, when fully fine-tuned, match the accuracy of a large LLM baseline on resume classification, and how small can such a model be while retaining this property? RQ2: Does LoRA provide a consistent accuracy-efficiency trade-off across models of different sizes? To answer these questions, we fine-tune MobileBERT [5], DistilBERT [4], and ModernBERT-base [6] on the ResumeAtlas dataset under both full fine-tuning and LoRA, and compare the resulting accuracy, macro F1-score, trainable parameter ratio, training time, and per-sample inference latency.

II. RELATED WORK

Heakl et al. [1] curated the ResumeAtlas dataset and benchmarked several classification approaches, with their best configuration based on Gemma 1.1 2B reaching a top-1 accuracy of 92% and a top-5 accuracy of 97.5%. Their work established a strong reference point for resume classification but relied on a model with approximately two billion parameters.

Low-Rank Adaptation (LoRA), introduced by Hu et al. [2], reduces the number of trainable parameters during fine-tuning by injecting small, low-rank update matrices into selected weight matrices of a pretrained model, typically the query and value projections of the attention mechanism. LoRA has been widely adopted as a default strategy for adapting large pretrained models at low cost, with rank values such as $r = 8$ commonly used as a starting configuration across many model families and tasks. A related comparative study by Nwaiwu [9] evaluated LoRA alongside other PEFT methods on DistilBERT for low-resource text classification, finding that PEFT performance relative to full fine-tuning can vary considerably depending on task and data conditions.

On the model side, all evaluated architectures build on the Transformer encoder introduced by Vaswani et al. [11]. DistilBERT [4] is a distilled version of BERT [3] that retains most of its language understanding capability with roughly 40% fewer parameters, MobileBERT [5] is an encoder architecture explicitly designed for resource-constrained deployment through bottleneck structures and knowledge distillation, and ModernBERT [6] is a more recent encoder architecture that incorporates modern training and architectural improvements over the original BERT design while remaining substantially smaller than billion-parameter LLMs. ALBERT [7] and RoBERTa [12] represent two further directions within the same encoder family, the former reducing parameters through cross-layer sharing and the latter optimizing the original BERT training recipe, both illustrating that encoder architectures and training choices, not only model size, shape downstream performance.

Most existing PEFT studies evaluate a single model scale or architecture family, leaving open the question of whether a fixed LoRA configuration generalizes across encoder models of different sizes for a specific downstream task such as

resume classification. This study aims to provide preliminary evidence on this question.

III. METHODOLOGY

This section describes the dataset, models, training configurations, and evaluation procedure used in this study. Fig. 1 illustrates the overall experimental pipeline, from dataset preparation through model fine-tuning to final evaluation.

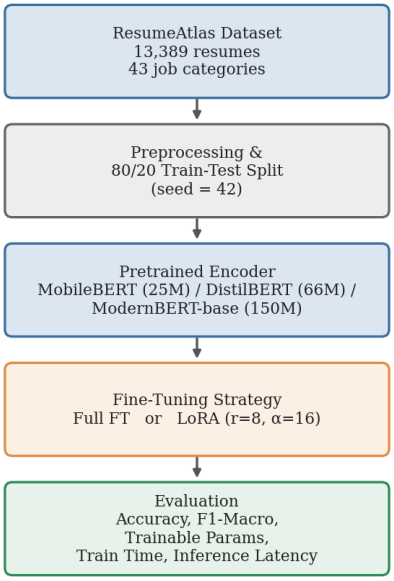

Fig. 1. Overview of the experimental pipeline, from dataset preparation to evaluation.

A. Dataset

We use the ResumeAtlas dataset, which contains 13,389 resumes labeled with one of 43 job categories (e.g., Accountant, Data Science, HR). The resume text has been pre-processed by the dataset authors (lowercased, punctuation removed). As the publicly available release contains a single split, we constructed an 80/20 train-test split with a fixed random seed (42) for all experiments reported in this study.

B. Models

Three compact transformer encoders were evaluated: (1) MobileBERT [5], with approximately 25 million parameters, (2) DistilBERT-base-uncased [4], with approximately 66 million parameters, and (3) ModernBERT-base [6], with approximately 150 million parameters. For each model, a sequence classification head with 43 output units was attached and initialized from scratch, following the standard Hugging Face AutoModelForSequenceClassification setup.

C. Training Configurations

Two fine-tuning regimes were compared for each model, as illustrated schematically in Fig. 2.

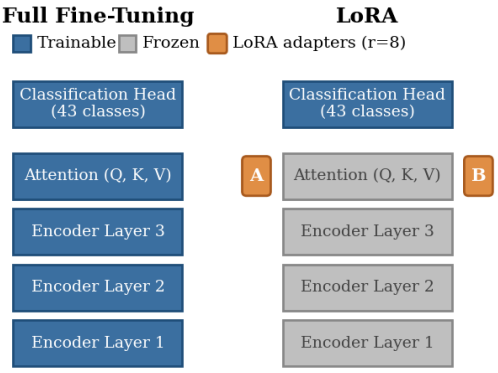

Fig. 2. Schematic comparison of full fine-tuning (all parameters trainable) and LoRA fine-tuning (frozen backbone with low-rank adapter matrices A and B inserted into the attention projections).

Each model was fine-tuned under two regimes. In the Full Fine-Tuning (Full FT) regime, all model parameters, including the backbone encoder and the classification head, were updated during training. In the LoRA regime, the backbone encoder was frozen and low-rank adapters were inserted into the attention projection layers (query and value projections for MobileBERT and DistilBERT; the fused query-key-value projection for ModernBERT), using rank $r = 8$, scaling factor $\alpha = 16$, and a dropout of 0.05, following the default configuration commonly recommended in the LoRA literature. The classification head remained trainable in both regimes. All models were trained for 3 epochs with a batch size of 8 and a maximum sequence length of 512 tokens.

D. Evaluation Metrics

For each configuration, we report classification accuracy and macro-averaged F1-score on the held-out test split, the number and proportion of trainable parameters, wall-clock training time, and average inference latency per sample, measured on the same GPU (NVIDIA T4).

E. Reference Baseline

As an external reference point, we report the 92% top-1 accuracy achieved by Gemma 1.1 2B on the ResumeAtlas dataset, as reported by Heakl et al. Because the exact train-test split used in the original study is not publicly available, this comparison is indicative rather than a strictly controlled replication; both studies, however, draw from the same underlying dataset of 13,389 resumes across 43 categories.

TABLE I PERFORMANCE COMPARISON ON THE RESUMEATLAS 43-CLASS RESUME CLASSIFICATION TASK

Model	Strategy	Trainable Params	Accuracy	F1-Macro	Train Time (min)	Infer. (ms/sample)
*Gemma 1.1 2B**	–	*≈ 2,000M*	*92.00%*	–	–	–
MobileBERT	Full FT	24.6M (100%)	92.01%	91.13%	26.1	1.97
MobileBERT	LoRA	0.19M (0.78%)	46.08%	42.40%	21.9	2.37
DistilBERT	Full FT	67.0M (100%)	91.15%	90.37%	32.1	0.40
DistilBERT	LoRA	0.77M (1.14%)	78.08%	74.04%	23.8	0.63
ModernBERT-base	Full FT	149.6M (100%)	92.98%	92.02%	90.6	1.18
ModernBERT-base	LoRA	0.57M (0.38%)	67.44%	66.20%	76.9	2.21

** Reference value from Heakl et al.; reported under a different train-test split and provided here as an indicative external baseline rather than a directly controlled comparison.*

IV. RESULTS

Table I summarizes the performance of the six trained configurations alongside the external Gemma 1.1 2B reference. Under full fine-tuning, all three compact models achieve accuracy close to or exceeding the reference value: MobileBERT reaches 92.01%, DistilBERT reaches 91.15%, and ModernBERT-base reaches 92.98%, with both MobileBERT and ModernBERT-base slightly exceeding the 92.00% reported for the 2-billion-parameter baseline. Under LoRA, all three models show a marked decrease in accuracy relative to their full fine-tuning counterparts. Notably, the magnitude of this decrease does not vary monotonically with model size: it is largest for MobileBERT (45.93 points), smallest for DistilBERT (13.07 points), and intermediate for ModernBERT-base (25.54 points). Fig. 3 visualizes these results relative to the reference baseline.

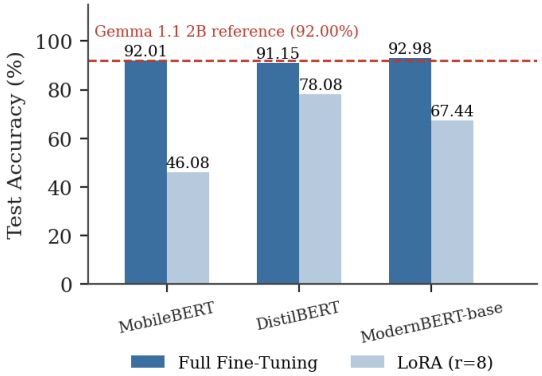

Fig. 3. Test accuracy of each configuration relative to the Gemma 1.1 2B reference (92.00%).

In terms of efficiency, LoRA reduces the number of trainable parameters substantially across all models, to roughly 0.78%, 1.14%, and 0.38% of the total model size for MobileBERT, DistilBERT, and ModernBERT-base, respectively. However, the corresponding reduction in training time is modest and inconsistent across models, and per-sample inference latency under LoRA is higher than under full fine-tuning for all three models, likely due to the additional computation introduced by the unmerged adapter matrices during the forward pass.

V. DISCUSSION

A. Compact Models Are Competitive with Large-Scale LLMs

The first main finding addresses RQ1. All three compact encoders, when fully fine-tuned, achieve accuracy close to or exceeding the 92% reported for Gemma 1.1 2B, despite having 13 to 80 times fewer parameters. MobileBERT, at only 25 million parameters, is particularly notable: it reaches 92.01% accuracy with roughly 80 times fewer parameters than the LLM baseline, and ModernBERT-base slightly surpasses the reference accuracy with roughly 13 times fewer parameters. This suggests that for a well-defined, single-label text classification task such as resume classification, the additional capacity of billion-parameter LLMs may not be strictly necessary, and that even very compact encoder models can represent a practical, lower-cost alternative for deployment in resource-constrained settings such as small organizations or applicant tracking systems with high throughput requirements. Fig. 4 places all configurations on a single accuracy-versus-parameters plot to illustrate this trade-off.

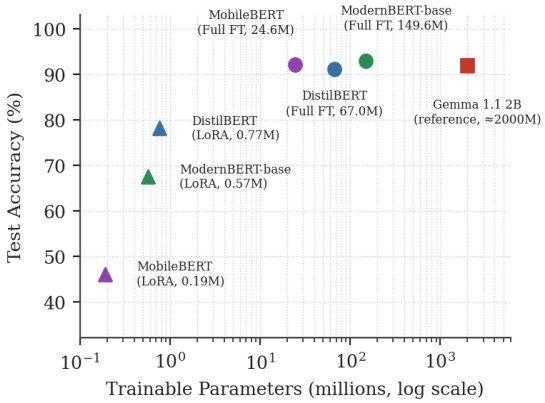

Fig. 4. Accuracy vs. trainable parameters (log scale), illustrating the accuracy-efficiency trade-off across configurations.

B. LoRA's Accuracy Gap Is Large and Architecture-Dependent, Not Simply Scale-Dependent

The second finding addresses RQ2 and is, to our knowledge, less commonly discussed in the PEFT literature. Using an identical default LoRA configuration ($r = 8$, $\alpha = 16$) for all three models, the accuracy gap between LoRA and full fine-tuning varies substantially: 45.93 points for MobileBERT (25M), 13.07 points for DistilBERT (67M), and 25.54 points for ModernBERT-base (150M). Critically, this gap does not increase or decrease monotonically with model size: DistilBERT, the middle-sized model, shows the smallest gap, while the smallest model (MobileBERT) shows the largest. This rules out a simple "smaller model, smaller LoRA capacity, larger gap" explanation and suggests instead that the gap is driven by architecture-specific factors, such as how attention is structured (e.g., MobileBERT's bottleneck attention blocks versus DistilBERT's standard multi-head attention) and how well a given architecture tolerates a frozen backbone with a small number of adapter parameters. If this pattern holds more generally, it would suggest that a single default LoRA configuration does not transfer reliably across compact encoder architectures, and that LoRA rank, target modules, or training schedule may need to be tuned per architecture rather than assumed to follow model size alone. We note that the training loss curves for all three LoRA runs had not fully plateaued by the third epoch, suggesting that part of the observed gap may also reflect slower convergence under LoRA rather than a fundamental capacity limitation; this is discussed further as a limitation below.

VI. LIMITATIONS AND FUTURE WORK

This study has several limitations. First, the analysis is based on three encoder architectures; a broader sweep of model sizes and families, including decoder-only language models, would be needed to confirm whether the observed non-monotonic LoRA gap pattern generalizes. Second, all models were trained for a fixed number of epochs (3); LoRA configurations may require a larger epoch budget or learning-rate schedule to fully converge, and future work should explore training schedules tailored separately to each fine-tuning regime, for example using early stopping based on validation loss. Third, the comparison with the Gemma 1.1 2B reference is indicative rather than a controlled replication, as the original train-test

split was not available. Finally, LoRA target modules and rank were held fixed across all three models; a systematic hyperparameter search per architecture may narrow the observed accuracy gaps and is left for future work.

VII. CONCLUSION

This study compared full fine-tuning and LoRA-based fine-tuning of three compact transformer encoders, MobileBERT, DistilBERT, and ModernBERT-base, on a 43-category resume classification task. Fully fine-tuned compact models, including one as small as 25 million parameters, matched or exceeded the accuracy of a 2-billion-parameter LLM baseline while requiring up to two orders of magnitude fewer parameters, supporting their use as cost-effective alternatives for automated resume classification. In contrast, LoRA with a uniform default configuration showed large accuracy gaps across all three models, with the gap size varying non-monotonically with model scale, indicating that LoRA hyperparameters may need to be adapted to specific encoder architectures rather than applied uniformly or assumed to follow model size alone. These findings motivate further investigation into architecture-aware PEFT configurations for domain-specific text classification tasks.

REFERENCES

[1] A. Heakl, Y. Mohamed, N. Mostafa, R. Harraf, and S. Khairy, "ResumeAtlas: Revisiting Resume Classification with Large-Scale Datasets and Large Language Models," arXiv:2406.18125, 2024.

[2] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," arXiv:2106.09685, 2021.

[3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv:1810.04805, 2018.

[4] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter," arXiv:1910.01108, 2019.

[5] Z. Sun, H. Yu, X. Song, R. Liu, Y. Yang, and D. Zhou, "MobileBERT: a Compact Task-Agnostic BERT for Resource-Limited Devices," arXiv:2004.02984, 2020.

[6] B. Warner et al., "Smarter, Better, Faster, Longer: A Modern Bidirectional Encoder for Fast, Memory Efficient, and Long Context Finetuning and Inference," arXiv:2412.13663, 2024.

[7] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations," arXiv:1909.11942, 2019.

[8] Gemma Team, "Gemma: Open Models Based on Gemini Research and Technology," arXiv:2403.08295, 2024.

[9] S. Nwaiwu, "Parameter-Efficient Fine-Tuning for Low-Resource Text Classification: A Comparative Study of LoRA, IA3, and ReFT," Frontiers in Big Data, vol. 8, 2025.

[10] M. Khaouja, I. Kassou, and M. Ghogho, "Challenges and Opportunities of NLP for HR Applications: A Discussion Paper," arXiv:2405.07766, 2024.

[11] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention Is All You Need," in Advances in Neural Information Processing Systems, 2017.

[12] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv:1907.11692, 2019.