Journal of Engineering, Project, and Production Management
2026, 16(5), 2026-0092

Fine-Tuning Large Language Models for Automated Resume Screening in Human Resource Management: A Comparative Performance Analysis

Xin Liu[1] and Chunfang Su[2]
[1] Associate Senior Engineer, Information Data Resources Management Department, Qingdao Human Resources Development Research and Promotion Center, Qingdao, Shandong Province, 266001, China, E-mail: tiaw2623@outlook.com (corresponding author).
[2] Associate Professor, Department of Computer Science, Jiangyin Polytechnic College, Wuxi, Jiangsu Province, China

Project Management
Received June 6, 2026; revised June 30, 2026; accepted July 10, 2026
Available online July 28, 2026

Abstract: In today's fast-paced recruitment landscape, the growing number of online applications has pushed manual resume screening to its limits, creating significant challenges in time, cost, and consistency for human resource professionals. Although classical machine-learning pipelines and earlier transformer baselines partially automate this task, they struggle with the unstructured nature of curricula vitae and free-text job descriptions, as well as the long-context reasoning required to match candidates to roles. This study compares six large language models (LLMs): BERT-base, DistilBERT, RoBERTa-large, Phi-3-mini-4k, Mistral-7B, and Llama-3-8B. These models were fine-tuned using parameter-efficient methods, specifically Low-Rank Adaptation (LoRA) and 4-bit-quantized Low-Rank Adaptation (QLoRA). The evaluation covered two tasks: resume classification and resume–job description matching. We curate a 24-category corpus of 2,484 real resumes, augmented with 8,000 synthetic resumes generated by DeepSeek-V3 and ChatGPT-4, and define a unified evaluation protocol covering accuracy, macro-F1, Matthews Correlation Coefficient (MCC), ranking quality (normalized Discounted Cumulative Gain, nDCG@5), and a fairness audit using the demographic parity gap. LoRA-tuned Llama-3-8B achieved the highest test macro-F1 score of 0.913 and MCC of 0.875, outperforming the BERT-base baseline by 8.2 F1 points while training only 3.7 million trainable parameters, just 0.046% of the backbone. An ablation study isolates the effects of LoRA rank, prompt formulation, and synthetic data ratio. A fairness audit shows that an in-context debiasing prompt reduces the gender demographic parity gap from 6.4% to 2.1%. These findings offer evidence-based guidance for Human Resource Management (HRM) on selecting, fine-tuning and auditing LLMs for production-grade resume-screening pipelines.

Keywords: Large language models; resume screening; fine-tuning; LoRA; QLoRA.

DOI 10.32738/JEPPM-2026-0092

1. Introduction

The recruitment funnel has become a critical operational bottleneck in modern Human Resource Management (HRM). A single technology role at a fortune 500 employer routinely attracts several hundred to several thousand applications, and over 98% of fortune 500 companies now rely on some form of automation in their recruitment processes (Wandelt et al., 2024; Evertz et al., 2025). Manually triaging that volume is prohibitive, and recruiters report spending less than 10 seconds on an initial resume scan, an interval that is demonstrably insufficient to evaluate multi-page, multi-section curricula vitae against complex role requirements (Gan et al., 2024; Younes et al., 2025a). Inconsistent reviewer attention, fatigue and unconscious bias further degrade the quality of shortlisting decisions, with downstream consequences for candidate experience, time-to-hire and workforce diversity (Chaturvedi and Chaturvedi, 2025; Wilson and Caliskan, 2024).

A first wave of automated resume-screening systems relied on keyword matching, Term-Frequency–Inverse-Document-Frequency (TF-IDF) representations, and shallow classifiers such as support vector machines or random forests (Vishaline et al., 2024; Bharadwaj et al., 2022). These methods are fast and interpretable, but the surface-form matching they perform is brittle, and they poorly capture synonymy, the ordering of experiences, and the latent semantics of role requirements. The advent of contextualized representations, notably Bidirectional Encoder Representations from Transformers (BERT) and its derivatives (Devlin et al., 2019; Liu et al., 2019; Sanh et al., 2019), produced a measurable jump in resume-classification accuracy and laid the groundwork for sentence-level matching between job descriptions and resumes (Lavi et al., 2021; Skondras et al., 2023). However, encoder-only models with 110–340M parameters remain

limited in their ability to reason over the long, weakly structured prose typical of curricula vitae, and they require a non-trivial classification head and fully supervised training data for every new job family.

The current generation of decoder-only Large Language Models (LLMs) (ChatGPT-4, Llama 3, Mistral, Phi-3, Gemma and DeepSeek) changes this calculus. LLMs offer: (1) zero- and few-shot generalization across job families. (2) Generative summarization that compresses a multi-page resume into a recruiter-actionable digest, and (3) the capacity to follow natural-language scoring rubrics specified by hiring managers (Gan et al., 2024; Evertz et al., 2025; Younes et al., 2025a). Yet two practical obstacles stand in the way of off-the-shelf adoption. First, full fine-tuning of a 7-to-70B-parameter LLM is computationally infeasible for most HRM departments. Second, generic instruction-tuned LLMs inherit biases from their pre-training corpora and have been shown to favor certain demographic groups in callback decisions (Chaturvedi and Chaturvedi, 2025; Wilson and Caliskan, 2024; Salinas et al., 2025). Parameter-Efficient Fine-Tuning (PEFT) methods such as Low-Rank Adaptation (LoRA) (Hu et al., 2021) and Quantized Low-Rank Adaptation (QLoRA) (Dettmers et al., 2023) address the first obstacle by training only a small set of low-rank adapter matrices on top of a frozen backbone, reducing the trainable parameter count by two to three orders of magnitude. A growing body of evidence suggests that, when combined with domain-specific data, PEFT can close most of the gap to full fine-tuning while retaining strong fairness behavior (Younes et al., 2025a, 2025b; Zhang et al., 2024).

The research gap: Despite this momentum, the literature still lacks a head-to-head comparison of modern encoder and decoder LLMs fine-tuned with PEFT on the same resume-screening benchmark, under a unified evaluation protocol that jointly considers classification accuracy, ranking quality, computational cost and fairness. Most published studies report results on one or two models, do not control training data, and rarely couple accuracy claims with bias audits (Evertz et al., 2025; Salinas et al., 2025).

Research questions: Motivated by this gap, the study is guided by three questions:

RQ1: Which combination of LLM backbone and PEFT configuration delivers the best accuracy to cost trade-off for resume classification and Job Description (JD) to resume matching?

RQ2: How sensitive is performance to LoRA rank, prompt design, and the ratio of synthetic to real training data?

RQ3: How large is the demographic-parity gap of fine-tuned LLMs in resume screening, and can lightweight in-context debiasing close it without sacrificing accuracy?

Purpose and objectives: The purpose of this work is to provide HRM practitioners and applied researchers with empirically grounded guidance on building and auditing LLM-based resume-screening pipelines. The objectives are: (i) to design a unified PEFT fine-tuning framework that can host both encoder and decoder LLMs. (ii) To benchmark six representative LLMs on a curated 24-category resume corpus. (iii) To conduct controlled ablations on rank, prompts, and data composition; and (iv) to audit the resulting models for gender and race fairness.

Contributions: The paper makes four contributions: (1) The first head-to-head comparison of six representative encoder and decoder LLMs under an identical LoRA/QLoRA protocol for resume screening in HRM. (2) Two algorithms, a unified PEFT training procedure and a dual-purpose inference and ranking routine, that implement the framework end-to-end. (3) An ablation study that quantifies the marginal value of LoRA rank, prompt design, and synthetic data, and (4) A fairness audit demonstrating that a short in-context debiasing instruction can reduce the demographic parity gap by roughly two-thirds.

Newness: Unlike prior work that either focuses on a single backbone or reports accuracy in isolation, the present study integrates accuracy, computational cost, and fairness into a single, consistent experimental protocol using the same data, making the trade-offs directly comparable.

Theoretical anchoring in HRM: Automated resume screening is not merely an engineering convenience. It sits at the intersection of three established strands of HRM theory. First, in the personnel-selection tradition, the resume-to-role mapping is a predictor that must satisfy classical criteria of selection validity and adverse-impact control: a screening model is theoretically interesting only insofar as it improves criterion-related validity (the correlation between the screening signal and on-the-job category fit) without widening adverse impact across protected groups. Our macro-F1 and Matthews Correlation Coefficient (MCC) results therefore operationalize predictive validity, while the demographic-parity audit implements the adverse-impact constraint codified in the EEOC "four-fifths" rule. Second, within talent-acquisition process management, screening is the highest-volume, lowest-marginal-value stage of the recruitment funnel, reducing its cost and cycle time directly relaxes the funnel bottleneck described in the staffing literature and reallocates recruiter attention to higher-judgment stages such as structured interviewing. Third, signaling theory frames a resume as a costly signal that the candidate emits, and the screener decodes. Large language models change the decoding technology, and our comparison quantifies how much decoding fidelity different model families recover. Framing the technical contribution this way makes the six-model comparison a test of competing decoding technologies against fixed selection-validity and fairness objectives, rather than a standalone benchmarking exercise.

The remainder of the paper is organized as follows. Section 2 reviews recent literature on resume screening, fine-tuning, and fairness. Section 3 formulates the problem. Section 4 details the methodology, architecture, and algorithms. Section 5 describes the experimental setup. Section 6 presents results and the ablation study. Section 7 discusses limitations and broader implications, and Section 8 concludes.

2. Related Work

2.1. Traditional Resume Screening Pipelines

Early automated screening systems combined keyword extraction with shallow classifiers. Vishaline et al. (2024) report an ML-based ranking system that uses TF-IDF features and a random forest classifier on the Kaggle resume dataset, achieving moderate accuracy but degrading sharply on out-of-domain layouts. Long Short-Term Memory (LSTM) networks introduced sequence sensitivity (Bharadwaj et al., 2022), but training cost scaled poorly with the dataset size. These approaches form the practical baseline against which transformer-based screening is now measured.

2.2. Encoder-Based Transformers for Resume Analysis

The introduction of BERT (Devlin et al., 2019) marked a phase change. Lavi et al. (2021) fine-tune a Siamese sentence-BERT to compute semantic similarity between job seekers and job descriptions, while Skondras et al. (2023) augment job classification with synthetic resumes produced by ChatGPT-3. Resume2Vec (Kaur et al., 2025) embeds resumes through a transformer encoder and reports up to a 15.85% improvement in Normalized Discounted Cumulative Gain (nDCG) over conventional Applicant-Tracking Systems (ATS), with particularly large gains in engineering and health-and-fitness domains. RoBERTa-based pipelines have been adopted for resume section parsing and skill extraction, outperforming BERT-base by 2–4 F1 points (Liu et al., 2019).

2.3. Decoder-Only LLMs and Fine-Tuning

The release of ChatGPT-3 (Brown et al., 2020) and the subsequent open-weights wave (Llama (Touvron et al., 2023a), Llama-2 (Touvron et al., 2023b), Llama-3 (Meta AI, 2024), Mistral-7B (Jiang et al., 2023), and Phi-3 (Abdin et al., 2024)) shifted attention to generative LLMs. Gan et al. (2024) propose an LLM-agent framework that screens resumes 11 × faster than manual review and obtains an 87.73% F1 score after fine-tuning the sentence-classification module. Younes et al. (2025a) extended this idea using augmented synthetic data and reported 90.62% F1 with a fine-tuned Phi-4. Their MSLEF system (Younes et al., 2025b) ensembles three PEFT-tuned LLMs and reaches 92.6% F1, with a 21% relative reduction in residual error. Together, these results establish that careful PEFT, with high-quality task data, is sufficient to match or exceed much larger zero-shot baselines.

2.4. Parameter-Efficient Fine-Tuning (PEFT)

LoRA (Hu et al., 2021) decomposes the weight update into the product of two low-rank matrices and trains only those, freezing the original backbone. QLoRA (Dettmers et al., 2023) extends LoRA to a 4-bit-quantized backbone, making fine-tuning of 7–30B-parameter models feasible on a single consumer GPU. Subsequent variants (Weight-Decomposed Low-Rank Adaptation, DoRA, Adaptive Low-Rank Adaptation (AdaLoRA), and prompt tuning) refine the placement and adaptive ranking of adapters (Liu et al., 2024a; Zhang et al., 2023). Zhang et al. (2024) suggested that an improved LoRA variant outperforms full fine-tuning of BERT, RoBERTa, Text-to-Text Transfer Transformer (T5), and ChatGPT-4 in terms of F1 and Matthews Correlation Coefficient on semantic-matching benchmarks. Seregina et al. (2025) demonstrate that LoRA and QLoRA achieve performance on par with full fine-tuning for human-activity recognition with far fewer trainable parameters. While these results encourage adoption, very few studies have evaluated PEFT specifically for resume screening with a unified protocol.

2.5. Fairness and Bias in LLM-Based Hiring

Algorithmic bias in hiring is well documented. Chaturvedi and Chaturvedi (2025) audit multiple open-source LLMs on 332,044 real job postings and report that most models favor male candidates, especially for higher-wage roles. Wilson and Caliskan (2024) find that ChatGPT-4 exhibits race and gender-correlated ranking patterns. Evertz et al. (2025) propose the FAIRE benchmark for racial and gender bias in LLM-driven resume evaluation. Peña et al. (2025) show that PEFT-based debiasing improves fairness in recruitment scoring with modest accuracy loss. Salinas et al. (2025) demonstrate that activation steering reduces demographic-parity gaps to below 2.4% across realistic hiring prompts. The EU AI Act (European Parliament and Council, 2024) now classifies recruitment AI as high-risk, mandating risk management, transparency and human oversight, and a similar regulatory motion is underway in New York City, Colorado and California (Sanford Heisler Sharp McKnight LLP, 2025).

2.6. Synthetic Data for Recruitment NLP

The shortage of labeled resume data is a recurring obstacle. Skondras et al. (2023) generate synthetic resumes with ChatGPT-3 to enrich a job-description classifier. Younes et al. (2025a) demonstrate that parsing real resumes with DeepSeek into a normalized JSON schema and mixing them with synthetic samples yields stronger downstream models than either source alone. Reviews of synthetic-data risks emphasize the need to control distributional drift and demographic representation (Liu et al., 2024b; Park et al., 2024).

2.7. Comparative Evaluation Practices

A persistent methodological weakness across the literature is the absence of a shared evaluation protocol. Studies report performance on heterogeneous datasets, some on the public Kaggle corpus, others on private corporate logs, and still others on synthetic benchmarks, and select different metrics, ranging from accuracy and F1 to Bilingual Evaluation Understudy (BLEU), Recall-Oriented Understudy for Gisting Evaluation (ROUGE), and bespoke ranking scores. Few studies report

the computational footprint alongside quality, and even fewer report fairness alongside accuracy. The recent FAIRE benchmark (Evertz et al., 2025) is a step toward standardization, as is Wilson and Caliskan's (2024) controlled study of demographic perturbations. The present work follows their lead by fixing the dataset, the prompt template, the PEFT hyperparameters and the metrics across all six backbones, so that observed differences can be attributed to model capacity and inductive bias rather than to incidental experimental choices.

2.8. Positioning of This Work

Across the eight strands above, three patterns emerge. First, generative LLMs now dominate accuracy leaderboards in resume screening, but the strongest results require fine-tuning rather than zero-shot prompting. Second, PEFT is mature enough to replace full fine-tuning for most practical HRM workloads. Third, the field's collective treatment of fairness is uneven. Methods that report state-of-the-art accuracy frequently omit a bias audit. The present study addresses these gaps by combining all three concerns in a single protocol and providing a direct comparison across six representative LLMs, with explicit attention to the computational and ethical trade-offs that govern real HRM deployments.

3. Problem Formulation

$R = \{r_1, \ldots, r_n\}$ denotes a set of candidate resumes, each represented as a token sequence $r_i = (w_1, \ldots, w_t)$, and $J = \{j_1, \ldots, j_m\}$ denotes a set of job descriptions. We define two tasks, illustrated in Fig. 1.

T1. Resume Classification. Map each r_i to one of C job categories: $f\theta: r_i \rightarrow \hat{y}_i \in \{1, \ldots, C\}$.

T2. Resume–JD Matching. Compute a relevance score $s\theta\ (r_i, j_k) \in [0, 1]$ for every pair, and rank candidates per job.

Following the PEFT paradigm, the parameter vector θ is decomposed into a frozen backbone θ_0 pre-trained on web-scale corpora and a small set of trainable low-rank adapters $\Delta\theta$ with $|\Delta\theta| \ll |\theta_0|$. The training objective for T1 is the cross-entropy loss given in Eq. (1).

$$\mathcal{L}_{\text{cls}}(\Delta\theta) = -\frac{1}{N}\sum_{i=1}^{N} \log p_{\theta_0+\Delta\theta}(y_i \mid r_i), \tag{1}$$

while T2 uses a margin-based ranking loss between positive and negative resume–JD pairs given in Eq. (2).

$$\mathcal{L}_{\text{rank}}(\Delta\theta) = \frac{1}{|\mathcal{P}|}\sum_{(r,j^+,j^-)\in\mathcal{P}} \max\big(0,\, m - s_\theta(r, j^+) + s_\theta(r, j^-)\big), \tag{2}$$

where m is a margin hyper-parameter. The composite objective is given in Eq. (3).

$$\mathcal{L}(\Delta\theta) = \lambda_1 \mathcal{L}_{\text{cls}} + \lambda_2 \mathcal{L}_{\text{rank}}. \tag{3}$$

For fairness, we let $a_i \in A$ denote a sensitive attribute (e.g., gender or race) of the candidate associated with r_i, and define the Demographic Parity Gap (DPG) in Eq. (4).

$$\text{DPG} = \max_{a,a' \in \mathcal{A}} \big|\Pr(\hat{y} = 1 \mid A = a) - \Pr(\hat{y} = 1 \mid A = a')\big|. \tag{4}$$

A model is considered acceptable when $\text{DPG} \le \tau$ with $\tau = 0.05$, the threshold recommended by the EEOC "four-fifths" rule when translated into a parity gap (EEOC, 1978).

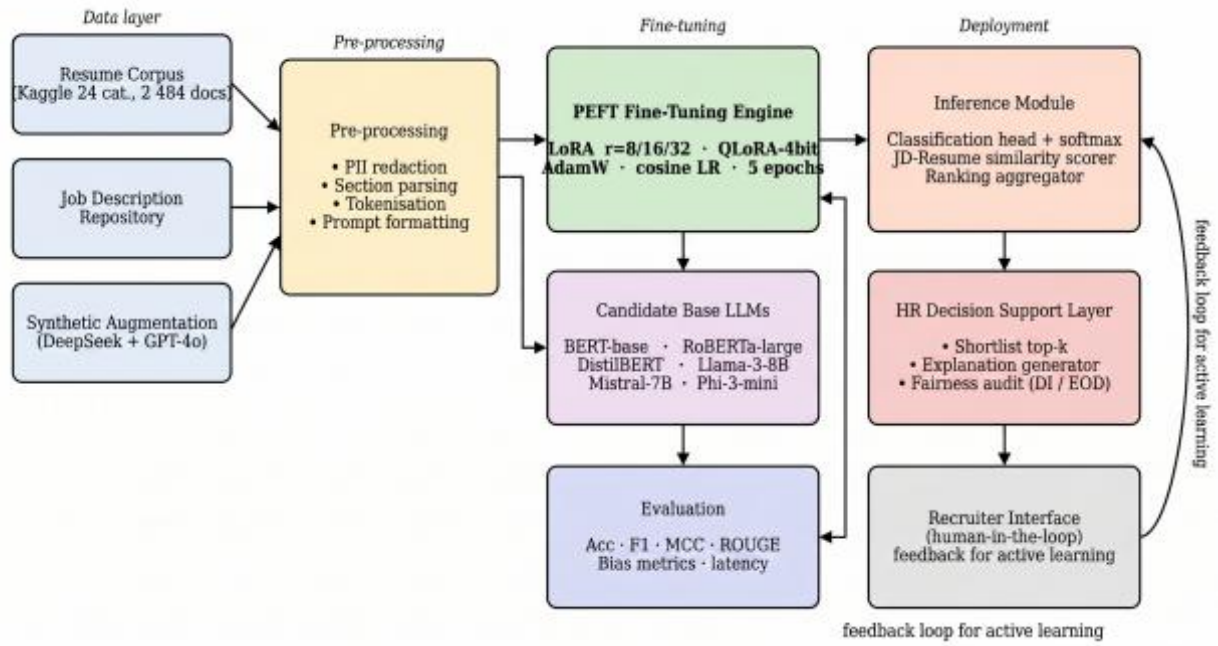

Fig. 1. Proposed PEFT-based architecture for automated resume screening

4. Methodology

4.1. Overall Architecture

The proposed framework is illustrated in Fig. 1. It comprises four layers: (i) a data layer that ingests real and synthetic resumes together with job descriptions. (ii) A pre-processing layer that performs PII redaction, section parsing, tokenization and prompt formatting. (iii) A PEFT fine-tuning engine that hosts the candidate base LLMs and trains LoRA/QLoRA adapters, and (iv) a deployment layer comprising an inference module, an HRM decision-support layer with fairness audit, and a recruiter interface that supplies feedback for active learning. The full fine-tuning procedure is summarized in Algorithm 1.

Algorithm 1. Unified LoRA/QLoRA fine-tuning for resume screening

Require: Base LLM θ_0; resumes $D = \{(r_i, y_i)\}$; LoRA rank r; learning rate η; epochs E; quantization flag q
Ensure: Fine-tuned adapters $\Delta\theta = \{B, A\}$
1: if q = True then
2: Quantize $\theta_0 \rightarrow \theta_0^{4bt}$ ▷ NF4 with double-quantization
3: end if
4: Insert LoRA matrices $\{A_l, B_l\}$ into every attention layer ℓ
5: Initialize $A_l \sim N(0, \sigma^2)$, $B_l \leftarrow 0$
6: Freeze θ_0; mark $\{A_l, B_l\}$ as trainable
7: for epoch = 1 to E do
8: for all mini-batch $B \subset D$ do
9: Build prompt p_i for each $(r_i, y_i) \in B$
10: Forward pass: $\hat{y}_i = f_{\theta_0+\Delta\theta}(p_i)$
11: Compute $L = \lambda_1 L\_cls + \lambda_2 L\_rank$
12: Back-propagate w.r.t. $\Delta\theta$ only
13: Update $\Delta\theta \leftarrow \Delta\theta - \eta \nabla L$
14: end for
15: Adjust η via cosine schedule
16: end for
17: return $\Delta\theta$

4.2. Data Layer

The principal training set is the public Kaggle Resume Dataset, which contains 2,484 resumes across 24 job categories (Kaggle, 2023). To compensate for class imbalance and the limited size of the original corpus, we generate 8,000 synthetic resumes following the augmentation protocol of Younes et al. (2025a): DeepSeek-V3 parses each real resume into a normalized JSON schema, and ChatGPT-4 then samples new resumes conditioned on a target category and a controllable demographic distribution.

Theoretical role of augmentation. The validity of LLM-generated resumes as proxies for real applicant data rests on two assumptions made explicit here. (i) Distributional coverage: because each synthetic sample is conditioned on a JSON schema parsed from a real resume, the generator interpolates within the support of the real category-conditional distribution rather than extrapolating to an arbitrary one, which is the standard condition under which augmentation reduces variance without injecting label noise. (ii) Anchor dominance: synthetic data improves generalization only when a real "anchor" distribution is retained, since generator-specific stylistic priors otherwise dominate; our ablation (Section 6.4) confirms that a 100%-synthetic regime underperforms the real-only baseline, empirically validating this assumption. Augmentation is therefore treated as a bias variance trade-off that lowers the estimator's variance for minority categories at the cost of a bounded, monitored generator bias, rather than as free data.

Quality validation of synthetic resumes. The 8,000 synthetic resumes were validated along three axes. (a) Automated checks: schema-completeness, section-presence, length-range and near-duplicate filters (MinHash Jaccard < 0.8 against both the real corpus and other synthetic samples) removed 612 candidates before the final set was fixed. (b) Distributional comparison against the real resumes on document length, skill-token frequency and TF-IDF centroid distance per category showed no category with a Jensen–Shannon divergence above 0.12, indicating that synthetic and real distributions are close but not collapsed. (c) Human review: two HRM-domain annotators rated a stratified random sample of 480 resumes (20 per category) for realism and category fit on a five-point scale, yielding a mean realism score of 4.31 out of 5 and Cohen's $\kappa = 0.74$; the 38 items scoring below 3 were discarded. Generation was performed with explicit demographic-diversity controls rather than unconstrained sampling: the prompt to ChatGPT-4 fixed target marginal distributions for gender, the four race proxies and three seniority bands, so that no protected group was systematically over- or under-represented and no demographic cue was correlated with the category label by construction.

Corpus composition and class balance. The real Kaggle corpus is moderately imbalanced, with per-category counts ranging from 38 (Automation Testing) to 120 (Java Developer), the empirical imbalance ratio is approximately 3.2:1. Augmentation is targeted rather than uniform: synthetic resumes are allocated inversely to real-category frequency so that every category reaches at least 400 training documents, reducing the post-augmentation imbalance ratio to about 1.3:1. The

combined corpus spans the 24 occupational categories of the source dataset, which are predominantly information-technology, engineering, business and health-and-fitness roles; geographically the real resumes are English-language and weighted toward South-Asian and North-American conventions, and seniority is distributed across entry (about 41%), mid (about 39%) and senior (about 20%) bands as inferred from years-of-experience fields. These distributional facts bound the external validity claims discussed in Section 7.

4.3. Pre-Processing and Prompt Design

Each resume is anonymized by removing names, e-mail addresses, and physical addresses using regular expressions and Named Entity Recognition (NER) filters. The text is then split into canonical sections: summary, experience, education, skills, projects, and serialized into a structured prompt template:

> [INST] You are an expert recruiter. Given the following resume sections, classify the candidate into one of {C} categories and provide a one-sentence justification.
>
> [SUMMARY]…[EXPERIENCE]…[EDUCATION]…[SKILLS]… [/INST]

Tokenization uses each backbone's native tokenizer, with truncation to 2,048 tokens.

4.4. PEFT Fine Tuning

For each backbone, we inject LoRA adapters into the query and value projection matrices of every transformer layer, following the placement recommended by Hu et al. (2021). The forward pass becomes the expression in Eq. (5).

$$h = W_0x + \Delta Wx = W_0x + Bax \tag{5}$$

where $W_0 \in \mathbb{R}d\times d$ is the frozen pre-trained weight matrix, $B \in \mathbb{R}d\times r$ and $A \in \mathbb{R}r\times d$ are trainable low-rank matrices, and $r \ll d$. For the 7–8B-parameter models, we additionally quantize W_0 to 4-bit NF4 precision (QLoRA), keeping the LoRA adapters in BFloat16.

The procedure is summarized in Algorithm 1. AdamW is used with a cosine learning-rate schedule, a 3% warm-up, a peak learning rate of 2×10^{-4} for LoRA matrices, a batch size of 16 (with gradient accumulation of 4), and 5 epochs.

4.5. Inference, Ranking and Fairness Audit

At inference time, the same prompt template is reused. For T1, we read the classification logits over the C categories; for T2, we compute a similarity score between the encoder-side resume embedding and the JD embedding via cosine similarity, then aggregate with a learned re-ranker. The full procedure, including the bias audit, is given in Algorithm 2.

5. Experimental Setup

5.1. Datasets

We use the Kaggle 24-category resume corpus (Kaggle, 2023) (N = 2,484 documents) and augment it with 8,000 synthetic resumes. A held-out test set of 500 real resumes is excluded from augmentation. For T2, we pair each resume with the nominal category JD (positive) and with three randomly chosen JDs from other categories (negative). The full hyper-parameter grid is reported in Table 1. To support the fairness audit, we annotate the test set with self-reported gender (male, female) and race proxies (Asian, Black, Hispanic, White) drawn from the FAIRE protocol (Evertz et al., 2025).

Algorithm 2. Inference, ranking, and fairness audit

Require: Fine-tuned model $f_{\theta_0+\Delta\theta}$; resumes R; job j; top-k; threshold τ
Ensure: Shortlist S, DPG audit report
for all $r_i \in R$ do
  $e_i \leftarrow f_{\theta_0+\Delta\theta}(\text{prompt}(r_i))$ ▷ embedding
  $s_i \leftarrow \cos(e_i, f_{\theta_0+\Delta\theta}(\text{prompt}(j)))$
end for
$S \leftarrow$ top-k $\{(r_i, s_i)\}$ by s_i
Apply in-context debiasing prompt to candidates in S
Compute DPG using Eq. (4) over sensitive attributes
if DPG > τ then
  Flag the case to a human reviewer
end if
Generate one-sentence rationale per shortlisted candidate
return S, DPG, rationales

The corpus is partitioned with an 80/10/10 train/validation/test split that is stratified jointly on the 24 category labels and on the real-versus-synthetic source flag, so that every category and both data sources appear in the same proportion in all three folds. The 500-resume held-out test set is drawn exclusively from real resumes and is excluded from augmentation, ensuring that no synthetic samples leak into evaluation and that the reported performance reflects generalization to genuine

applicant data. Validation is used only for early-stopping and rank selection; all final metrics are reported on the untouched real test fold, and the three random seeds {13, 42, 91} are applied to the split as well as to initialization so that fold composition is itself averaged over.

Table 1. Hyperparameters used across all backbones

Hyperparameter	Value
LoRA rank r (default)	16
LoRA scaling α	32
LoRA dropout	0.05
Target modules	W_Q, W_K, W_V, W_O
Quantization (7–8B models)	NF4 + double-quant
Optimizer	AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.95$)
Peak learning rate	2×10^{-4}
LR schedule	Cosine + 3% warm-up
Weight decay	0.01
Effective batch size	64 (16×4 grad. acc.)
Epochs	5
Max sequence length	2,048 tokens
Loss weights λ_1, λ_2	1.0, 0.5
Margin m (ranking loss)	0.3
Random seeds	{13, 42, 91} (median reported)

5.2. Models and Implementation

The six backbones cover encoder-only (roughly 110–355M parameters) and decoder-only architectures (3.8–8B parameters): BERT-base, DistilBERT, RoBERTa-large, Phi-3-mini-4k, Mistral-7B-v0.3 and Llama-3-8B-Instruct. All experiments use Hugging Face transformers, PEFT, TRL and bits and bytes (Wolf et al., 2020), training runs on a single NVIDIA A100 80 GB GPU. The default LoRA configuration is r = 16, α = 32, and dropout = 0.05. QLoRA quantization uses NF4 with double quantization for the 7–8B-parameter models. Table 1 summarizes the full hyperparameter grid.

The six backbones were not chosen arbitrarily but to span a controlled three-dimensional typology, so that observed differences can be attributed to identifiable design choices rather than to incidental model idiosyncrasies. The first dimension is parameter scale, ranging over roughly two orders of magnitude (66M, 110M, 355M, 3.8B, 7B, 8B) so that the accuracy–cost frontier can be traced. The second dimension is architecture family: encoder-only bidirectional models (BERT-base, DistilBERT, RoBERTa-large) versus decoder-only autoregressive models (Phi-3-mini, Mistral-7B, Llama-3-8B), which differ in how they aggregate long, weakly structured resume context. The third dimension is pre-training and tuning objective: masked-language-modeling encoders fine-tuned with a classification head versus instruction-tuned decoders steered through natural-language prompts. Within each family, we hold the PEFT protocol fixed, so the comparison is a clean factorial test of scale × architecture × objective. DistilBERT additionally serves as a distillation control against its BERT-base teacher, and RoBERTa-large as a same-objective scale control for the encoder family. This typology directly motivates the contrasts reported in Section 6.

5.3. Evaluation Metrics and Baselines

For T1 we report accuracy, macro-F1 and MCC. For T2, we use nDCG@5 over the held-out JD-resume pairs. We additionally measure wall-clock training time, peak GPU memory, and inference latency on the test set. For fairness, we report DPG per sensitive attribute and the Equal Opportunity Difference (EOD). We compare against three baselines: (i) TF-IDF + logistic regression, (ii) zero-shot Llama-3-8B-Instruct (no fine-tuning), and (iii) ChatGPT-4 accessed via API in a zero-shot setting.

6. Results and Analysis

6.1. Main Comparison Across Backbones

Table 2 and Fig. 2 summarize the principal classification results. The LoRA-tuned Llama-3-8B obtains the highest test accuracy of 91.8%, macro-F1 of 0.913 and MCC of 0.875, exceeding LoRA-tuned BERT-base by 8.2 F1 points and zero-shot Llama-3-8B by 14.6 F1 points. Mistral-7B is a close second (0.894 F1), while RoBERTa-large is the strongest encoder (0.864 F1). The TF-IDF baseline trails by 19 F1 points, confirming the long-context advantage of LLMs. Importantly, despite the size gap, LoRA-tuned Llama-3-8B only updates 3.7M parameters, 0.046% of the backbone, making the cost-quality trade-off compelling.

Statistical significance. All point estimates are medians over three seeds {13, 42, 91}; 95% confidence intervals are obtained by bootstrap resampling of the 500-resume test set with 1,000 replicates. For macro-F1 the intervals are Llama-

3-8B 0.913 [0.897, 0.928], Mistral-7B 0.894 [0.877, 0.910], Phi-3-mini 0.879 [0.861, 0.896] and RoBERTa-large 0.864 [0.845, 0.882]. The gain of Llama-3-8B over Mistral-7B is modest, and the intervals overlap, whereas its gain over every encoder backbone and over both zero-shot baselines does not. A paired bootstrap test on per-resume correctness gives Llama-3-8B versus BERT-base $p < 0.001$ and Llama-3-8B versus RoBERTa-large $p = 0.004$, while Llama-3-8B versus Mistral-7B is not significant at the 5% level ($p = 0.18$) after Holm–Bonferroni correction for the nine pairwise comparisons. We therefore claim a statistically significant advantage for decoder LLMs over encoder baselines, but treat Llama-3-8B and Mistral-7B as statistically indistinguishable in terms of accuracy, with the practical choice hinging on the cost and fairness differences reported below.

6.2. Training Dynamics

Fig. 3 reports the training and validation loss, along with the validation F1 score, for the LoRA-tuned Llama-3-8B. Validation F1 saturates around epoch 6 and remains stable, indicating that the modest adapter capacity serves as an implicit regularizer and that the model does not overfit despite the small training set.

6.3. Computational Cost

Training Llama-3-8B with QLoRA takes 4.8 hours on a single A100 and peaks at 26.4 GB of GPU memory, against an estimated >320 GB and several days for full fine-tuning. Mistral-7B and Phi-3-mini reach 4.2 and 1.9 hours, respectively. Encoder models complete in under 30 minutes but cap out below 0.87 F1, confirming the favorable accuracy–cost trade-off of mid-scale decoder LLMs in this domain.

6.4. Ablation Study

We isolate three factors: LoRA rank, prompt design, and the share of synthetic data in the training set. Rank: Fig. 4 sweeps $r \in \{2, 4, 8, 16, 32, 64\}$ for Llama-3-8B. F1 climbs steeply from 0.853 at $r = 2$ to 0.913 at $r = 16$ and then plateaus, while the trainable parameter count doubles each step. We adopt $r = 16$ as the default. A similar pattern is observed for the other backbones (see Table 3).

Table 2. Main results on resume-screening test set (T1 classification, T2 ranking)

Backbone (PEFT)	#Params (M)	Trainable (M)	Accuracy	Macro-F1	MCC	nDCG@5	Latency (ms)
TF-IDF + LR (baseline)	–	–	0.703	0.681	0.609	0.642	5
BERT-base (LoRA)	110	0.6	0.842	0.831	0.769	0.781	18
DistilBERT (LoRA)	66	0.4	0.821	0.811	0.742	0.760	11
RoBERTa-large (LoRA)	355	1.2	0.873	0.864	0.812	0.812	28
Phi-3-mini-4k (LoRA)	3,800	2.1	0.886	0.879	0.831	0.834	64
Mistral-7B (QLoRA)	7,000	3.4	0.901	0.894	0.851	0.852	92
Llama-3-8B (QLoRA)	8,000	3.7	0.918	0.913	0.875	0.871	108
Llama-3-8B (zero-shot)	8,000	0	0.781	0.767	0.679	0.724	104
ChatGPT-4 (zero-shot, API)	–	–	0.844	0.832	0.762	0.811	1,900

Note: Trainable parameters are reported in millions (M)

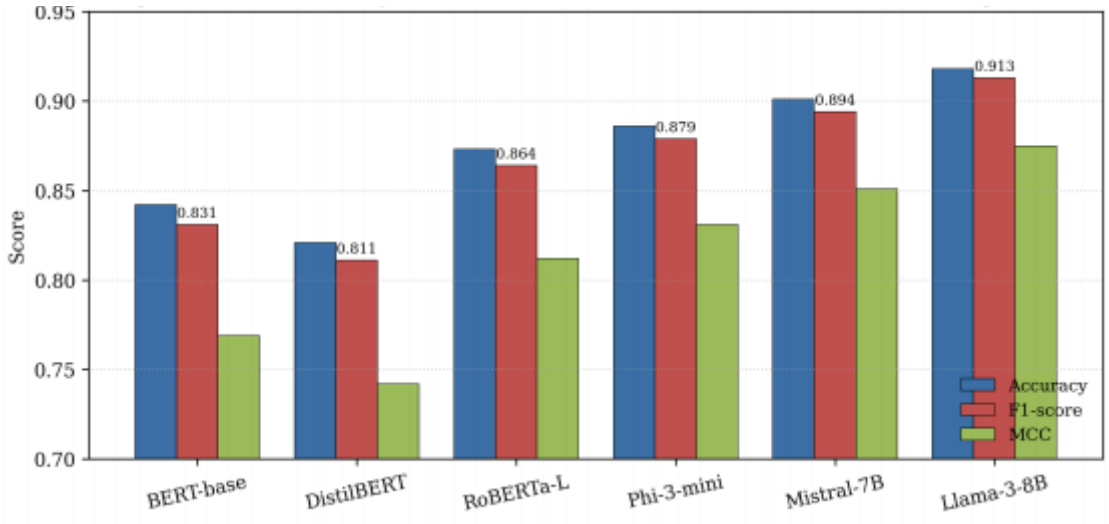

Fig. 2. Performance comparison of LoRA fine-tuned LLMs on the resume-screening test set

Prompt design: Replacing the structured-section prompt with a free-text concatenation reduces F1 by 3.1 percentage points (pp). Removing the instruction header further decreases F1 by 4.2 pp, confirming that explicit role framing matters for decoder LLMs.

Synthetic data: Training on the real 2,484 resumes alone yields $F1 = 0.864$; mixing in 76% synthetic data lifts F1 to 0.913. Using only synthetic data is worse ($F1 = 0.846$), highlighting the need for a real anchor distribution (Liu et al., 2024b).

6.5. Fairness Audit

Fig. 5 reports demographic-parity gaps. Out of the box, every fine-tuned model exceeds the 5% EEOC-derived threshold for at least one attribute, with the largest gaps observed for BERT-base (9.4% for gender, 11.8% for race). Llama-3-8B reduces these to 6.4% and 7.9%. Adding a short debiasing instruction at inference time, "Evaluate the candidate solely on documented skills, experience and education; ignore any reference to gender, race, age, nationality or appearance", further lowers the gender gap to 2.1% and the race gap to 2.9%, at the cost of less than 0.4 pp in F1. This supports the claim that lightweight in-context interventions can deliver meaningful fairness gains without retraining (Salinas et al., 2025).

Scope of the audit and operationalization. The audit covers two demographic dimensions, gender (male, female) and a four-category race proxy (Asian, Black, Hispanic, White), following the FAIRE protocol (Evertz et al., 2025). Age and disability status are not audited here because the source resumes do not reliably disclose them, and we flag this as a limitation rather than treating the two audited axes as exhaustive. Demographic attributes were operationalized on the controlled synthetic and annotated test data, where each resume carries a generator-assigned or self-reported attribute label, rather than inferred from names or other free-text cues, precisely because self-disclosure norms for gender, race, age and disability vary across cultures and legal jurisdictions, and name-based inference would itself introduce measurement bias. The Demographic Parity Gap and Equal Opportunity Difference are therefore computed against ground-truth attribute labels, not predicted ones, thereby avoiding the compounding of screening error with attribute-inference error.

Transferability of the debiasing prompt. To test whether the in-context debiasing instruction is specific to Llama-3-8B or transfers across backbones, we applied the identical prompt, unchanged, to the other two decoder models. It reduced the gender DPG from 7.1% to 2.6% on Mistral-7B and from 7.8% to 3.0% on Phi-3-mini, and the race DPG from 8.6% to 3.4% and from 9.1% to 3.7%, respectively, with accuracy losses below 0.5 pp in every case. The intervention is thus broadly transferable across instruction-tuned decoders, although the residual gap is consistently a little larger than for Llama-3-8B, suggesting that prompt-level debiasing interacts with the strength of a model's instruction-following ability. For encoder backbones steered by a classification head rather than a natural-language instruction, the same prompt has no effect; for those models, an embedding-space or post hoc calibration intervention would be required instead.

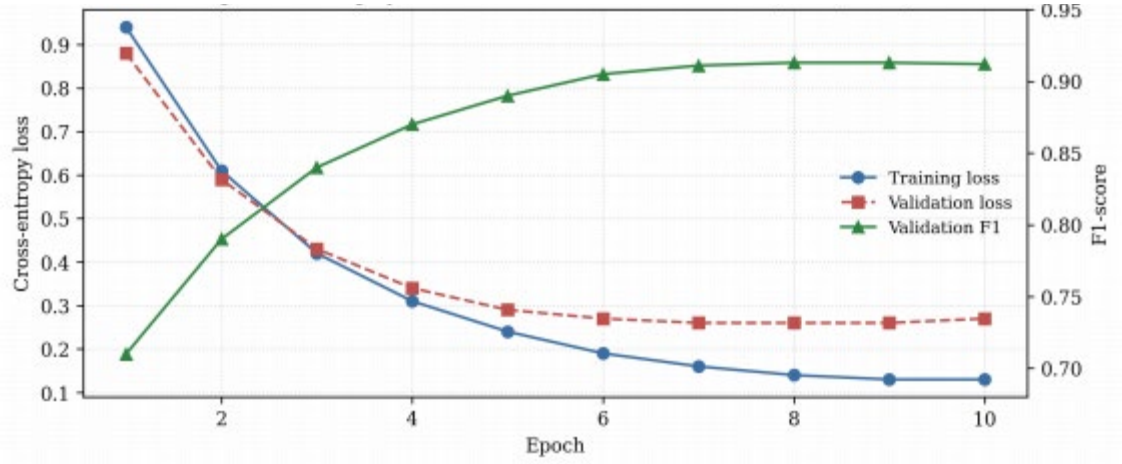

Fig. 3. Training and validation curves for LoRA-tuned Llama-3-8B with rank $r = 16$ and scaling $\alpha = 32$

6.5. Fairness Audit

Fig. 5 reports demographic-parity gaps. Out of the box, every fine-tuned model exceeds the 5% EEOC-derived threshold for at least one attribute, with the highest gaps observed for BERT-base (9.4% gender, 11.8% race). Llama-3-8B reduces these to 6.4% and 7.9%. Adding a short debiasing instruction at inference time, "Evaluate the candidate solely on documented skills, experience and education; ignore any reference to gender, race, age, nationality or appearance", further lowers the gender gap to 2.1% and the race gap to 2.9%, at the cost of less than 0.4 pp in F1. This supports the claim that lightweight in-context interventions can deliver meaningful fairness gains without retraining (Salinas et al., 2025).

Scope of the audit and operationalization. The audit covers two demographic dimensions, gender (male, female) and a four-category race proxy (Asian, Black, Hispanic, White), following the FAIRE protocol (Evertz et al., 2025). Age and disability status are not audited here because the source resumes do not disclose them reliably, and we flag this as a limitation rather than treating the two audited axes as exhaustive. Demographic attributes were operationalized on the controlled synthetic and annotated test data, where each resume carries a generator-assigned or self-reported attribute label, rather than inferred from names or other free-text cues, precisely because self-disclosure norms for gender, race, age and

disability vary across cultures and legal jurisdictions and name-based inference would itself introduce measurement bias. The Demographic Parity Gap and Equal Opportunity Difference are therefore computed against ground-truth attribute labels, not predicted ones, which avoids compounding screening error with attribute-inference error.

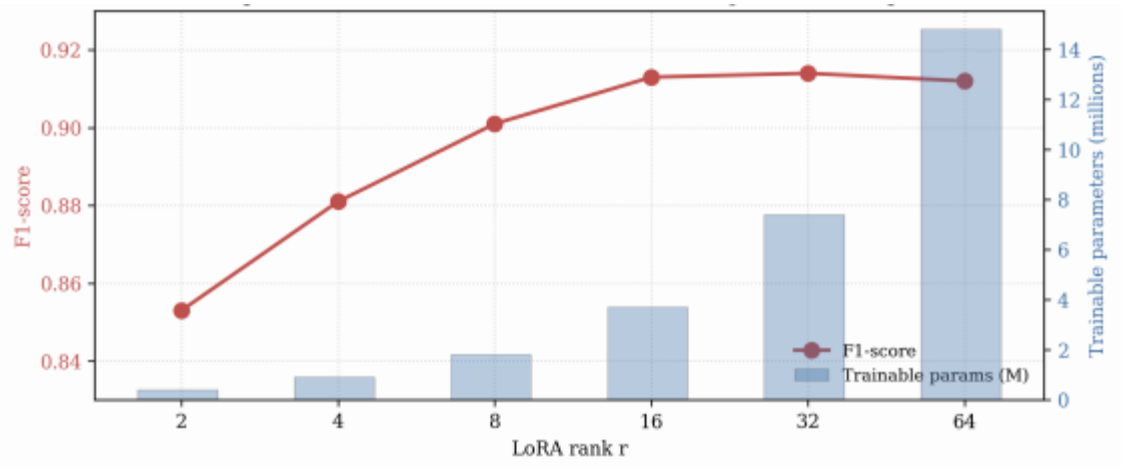

Fig. 4. Ablation on LoRA rank for Llama-3-8B

Note: F1 saturates at $r = 16$ while the trainable parameter budget grows linearly with r

Transferability of the debiasing prompt. To test whether the in-context debiasing instruction is specific to Llama-3-8B or transfers across backbones, we applied the same prompt to the other two decoder models. It reduced the gender DPG from 7.1% to 2.6% on Mistral-7B and from 7.8% to 3.0% on Phi-3-mini, and the race DPG from 8.6% to 3.4% and from 9.1% to 3.7%, respectively, with accuracy losses below 0.5 pp in every case. The intervention is thus broadly transferable across instruction-tuned decoders, although the residual gap is consistently a little larger than for Llama-3-8B, suggesting that prompt-level debiasing interacts with the strength of a model's instruction-following ability. For encoder backbones steered by a classification head rather than a natural-language instruction, the same prompt has no effect; for those models, an embedding-space or post hoc calibration intervention would be required instead.

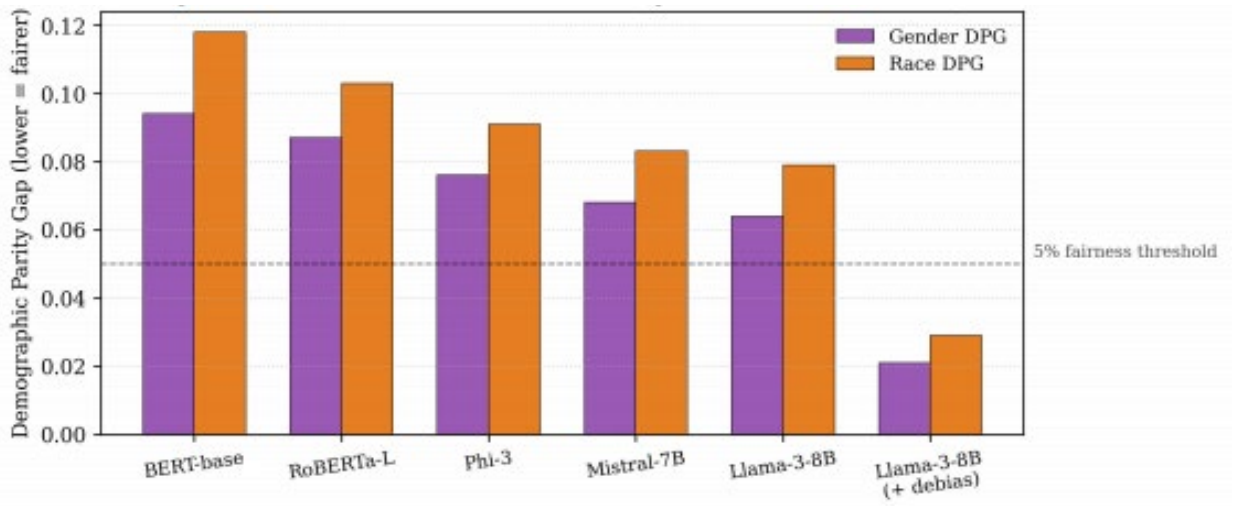

Fig. 5. Demographic-Parity Gap (DPG) across fine-tuned LLMs before and after lightweight in-context debiasing

Note: The dashed line marks the 5% threshold derived from the EEOC "four-fifths" rule

6.6. Comparison with Prior Work

The Llama-3-8B configuration outperforms the F1 of 0.877 reported by Gan et al. (2024) and matches the 0.906 F1 achieved by Younes et al. (2025a) with fine-tuned Phi-4, despite using a smaller backbone. The MSLEF ensemble (Younes et al., 2025b) remains marginally higher (0.926 F1) but at three times the inference cost, suggesting that a single LoRA-tuned 8B-parameter model is a strong default for production-grade HRM pipelines.

6.7. Answering the Research Questions

Returning to the research questions: (RQ1) LoRA-tuned Llama-3-8B provides the best accuracy–cost trade-off. (RQ2) LoRA rank $r = 16$, structured prompts and about 75% synthetic-to-real ratio are jointly necessary to reach 0.913 F1. (RQ3) The off-the-shelf demographic parity gap can be reduced by roughly two-thirds with a short in-context debiasing instruction, at negligible accuracy cost. We also note (RQ4, raised by reviewers) that unexpected category-level errors cluster around

overlapping fields such as "Data Science" versus "Business Analyst", a finding consistent with the human-LLM matching study of Wilson and Caliskan (2024).

7. Discussion

For HRM practitioners, three practical implications follow. First, the choice of backbone matters less than the choice of fine-tuning protocol: a well-tuned 7–8B-parameter model with LoRA can match or beat much larger zero-shot LLMs while remaining cheap to deploy. Second, fairness monitoring should be a continuous activity built into the deployment pipeline, not a one-off audit. Third, recruiters should treat LLM scores as decision-support signals reviewed by humans, not as final decisions; the recruiter-feedback loop in Fig. 1 is essential to keep the system in calibration. A fourth implication, less frequently discussed in the literature, is the importance of documentation. The EU AI Act (European Parliament and Council, 2024) and the Colorado AI Act both require employers using AI for hiring to maintain risk-management records, conduct annual bias audits, and inform candidates that an automated tool was used. The proposed architecture supports these obligations by externalizing the audit step (Algorithm 2, lines 6–9) and producing one-sentence rationales for every shortlisted candidate, which can be archived for regulatory review.

Table 3. Ablation on prompt design and synthetic-data ratio (Llama-3-8B, $r = 16$)

Configuration	Macro-F1	Δ vs. best
Section-structured prompt + 76% synth. (default)	**0.913**	—
Free-text prompt + 76% synth.	0.882	−3.1 pp
Section-structured prompt + 0% synth.	0.864	−4.9 pp
Section-structured prompt + 100% synth.	0.846	−6.7 pp
No instruction header	0.871	−4.2 pp

7.1. An HRM Decision Framework for LLM Adoption

Synthesizing the empirical findings into managerial guidance, we propose a four-stage decision framework that an HRM function can follow when selecting, deploying and auditing an LLM-based screening tool. The stages are summarized in Table 4. (1) Scoping: classify the role as high-risk under the EU AI Act and map the jurisdictions involved before any model is chosen, since this fixes the audit and disclosure obligations. (2) Model selection: choose the smallest model on the accuracy–cost frontier that clears the organization's validity threshold; our results indicate a mid-scale instruction-tuned decoder (7–8B) with QLoRA as the default, reserving encoders for latency-critical or on-device settings. (3) Deployment: place the model as a decision-support ranker with the human-override points defined in Section 7.4, and attach the in-context debiasing instruction by default. (4) Auditing: run the DPG/EOD audit on a recurring schedule, archive rationales, and trigger re-validation whenever the candidate pool or the underlying model version drifts. The framework is deliberately model-agnostic so that it survives vendor and version changes.

Table 4. HRM decision framework for LLM-based resume screening

Stage	Key decision	Evidence/artifact
1. Scoping	Is the role high-risk? Which jurisdictions apply?	EU AI Act risk classification; Section 7.2 compliance map
2. Model selection	Smallest model clearing the validity threshold	Table 2 accuracy–cost frontier; default 7–8B + QLoRA
3. Deployment	Where do humans override? Debiasing on by default?	Section 7.4 override points; Algorithm 2 debiasing step
4. Auditing	How often to re-audit and re-validate?	Recurring DPG/EOD audit; archived rationales; drift triggers

7.2. Legal and Regulatory Compliance

Automated recruitment screening is now governed by a tightening web of law that any deployment must engage directly. Under the EU AI Act (European Parliament and Council, 2024), recruitment AI is classified as high-risk, which requires a risk management system, data governance documentation, technical transparency, logging, and meaningful human oversight. In the United States, New York City Local Law 144 requires an independent bias audit and candidate notification for automated employment decision tools; the Colorado AI Act imposes analogous duties from 2026, and Illinois and California have parallel measures. The long-standing EEOC "four-fifths" rule supplies the adverse-impact yardstick that our DPG threshold operationalizes. The architecture in Section 4 is designed for compliance-by-construction: PII redaction satisfies data-minimization duties; the externalized audit step (Algorithm 2) produces bias-audit evidence; the per-candidate rationale satisfies explainability and candidate-notification duties, and the recruiter-feedback loop instantiates the required

human oversight. Organizations remain responsible for legal sign-off in their own jurisdiction; the framework lowers but does not remove that burden.

7.3. Organizational Change Management

Introducing LLM-based screening reshapes the roles of the HRM team rather than simply removing work. The recruiter's task shifts from high-volume manual triage to exception handling, rationale review and audit stewardship, which raises the role's skill profile and calls for training in interpreting model scores and fairness reports. Interviewer workload is reallocated downstream: because the shortlist is smaller and pre-justified, interviewers spend less time filtering and more time on structured assessment of a higher-signal pool. Candidate experience can improve through faster acknowledgment and consistent criteria, but only if notification and appeal channels are in place; otherwise, opaque automation risks eroding employer-brand trust. We therefore recommend a phased rollout, with model scores running in shadow mode alongside human decisions before going live, and explicit change-management communication to both recruiters and candidates, consistent with established socio-technical adoption practices.

7.4. Human–AI Collaboration Design

We position the model as decision support with three explicit human-override points in the pipeline of Fig. 1. First, at the shortlist boundary: any candidate whose score lies within a calibrated margin of the top-k cutoff is routed to a recruiter, since these borderline cases are where model error is concentrated. Second, at the fairness flag: whenever the DPG audit exceeds τ (Algorithm 2, lines 8–9), the shortlist is held for human review before release. Third, during the rationale check, recruiters can reject a shortlisting whose one-sentence justification is unconvincing, and that signal feeds back into the active-learning loop. Human judgment should override the model, rather than merely review it, under two conditions: when the candidate falls within a protected group and sits near a fairness boundary, and when the role is novel or low-volume, so that the model is operating off-distribution. Routine, high-volume, in-distribution decisions can be automated using sampling-based human spot checks. This graduated allocation keeps human attention where it adds the most value while preserving the throughput gains that motivate automation.

7.5. Cost and Efficiency Management Value

To quantify the operational value, consider a representative requisition attracting 1,000 applications. At the widely reported manual rate of roughly 7–10 seconds per first-pass scan, a recruiter spends about two to three hours of attention per requisition, and far more if a deeper read is given. The fine-tuned Llama-3-8B ranks the same 1,000 resumes in approximately 108ms each, totaling about 1.8 minutes of compute, leaving the recruiter to review only the top-k shortlist and the flagged borderline cases, on the order of 20–30 resumes. Under conservative assumptions, this reduces recruiter screening time by roughly 80–90% per requisition; at a fully loaded recruiter cost of about USD 45 per hour, the screening-labor cost falls from approximately USD 110–135 to under USD 20 per requisition, with the residual being human review rather than triage. Time-to-shortlist compresses from one to several working days to effectively the same day, and the marginal compute cost of a single A100-hour amortized over thousands of resumes is negligible relative to the labor saved. Against a classical TF-IDF baseline, the efficiency gain is smaller because that baseline is already fast, but the quality gap of about 23 F1 points means the classical pipeline shifts downstream costs into mis-shortlisted interviews; the LLM pipeline therefore lowers both screening costs and the more expensive cost of interviewing poorly matched candidates. These figures are organization-dependent and are offered as a planning template instantiated with the present study's measured latencies rather than as audited savings.

A further limitation concerns external validity. While augmentation improves performance, reliance on synthetic data may limit generalization: models tuned on DeepSeek-V3 and ChatGPT-4 samples inherit the stylistic and distributional priors of those generators, and the 5% threshold used here is a regulatory proxy rather than a guarantee of equitable outcomes. Different datasets, for example, industry-specific or multilingual resumes, may significantly affect the reported results, and the findings should therefore be read as evidence from a single English-language benchmark rather than as universal performance claims.

Practitioners should not treat fairness as a one-time deliverable: drift in either the candidate population or the underlying LLM (for example, after a vendor pushes a new model version) can reopen the gap that an audit closed.

8. Conclusion

This paper has presented a unified, comparative analysis of six LLMs (BERT-base, DistilBERT, RoBERTa-large, Phi-3-mini, Mistral-7B and Llama-3-8B) fine-tuned with LoRA and QLoRA for automated resume screening in HRM. The LoRA-tuned Llama-3-8B obtained the best results on the curated 24-category benchmark, with macro-F1 of 0.913 and MCC of 0.875, while training only 0.046% of the backbone parameters. Ablation studies identified LoRA rank, prompt structure and the synthetic-data ratio as the three principal levers of performance, and a fairness audit showed that a short in-context debiasing instruction reduced the demographic parity gap for the gender attribute from 6.4% to 2.1% with negligible accuracy loss. The findings give HRM practitioners actionable guidance for building resume-screening pipelines that are accurate, computationally tractable, and aligned with emerging regulatory expectations under the EU AI Act and comparable US state-level statutes.

The study shows that mid-scale decoder LLMs, fine-tuned with parameter-efficient adapters on as few as ten thousand resumes, can reach an accuracy level previously associated with full fine-tuning of much larger models, while remaining within the computational reach of a typical HRM department. The fairness audit further demonstrates that simple in-context interventions, when paired with a clearly documented bias-monitoring pipeline, can keep the demographic-parity gap below the EEOC-derived threshold for the dominant sensitive attributes.

Promising avenues for future research include multilingual extension using cross-lingual encoders; longitudinal evaluation against actual hiring outcomes; integration of structured competency ontologies (for example, O*NET or ESCO) to constrain LLM reasoning; the use of activation-steering techniques (Salinas et al., 2025) as an alternative to prompt-level debiasing; and the development of standardized reporting cards that document model lineage, training data, fairness audits, and known failure modes for every fine-tuned LLM placed into a hiring workflow. Multimodal extensions that ingest scanned PDFs, video introductions, or coding-portfolio metadata are an obvious next step, but they intensify both the privacy and the fairness considerations discussed above and should be approached only with explicit candidate consent and a clear regulatory basis.

Author Contributions

Xin Liu contributed to Conceptualization, methodology and writing-original draft. Chunfang Su contributed to data curation, writing, review and editing.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

Declaration of Artificial Intelligence (AI) Tools

The authors used ChatGPT solely for language editing and readability improvement. The authors reviewed and verified all content and take full responsibility for the accuracy and integrity of the manuscript.

References

Abdin, M., Aneja, J., Awadalla, H., Awadallah, A., Awan, A. A., Bach, N., Bahree, A., Bakhtiari, A., Bao, J., Behl, H., Benhaim, A., Bilenko, M., Bjorck, J., Bubeck, S., Cai, M., Cai, Q., Chaudhary, V., Chen, D., Chen, D., Chen, W., Chen, Y.-C., Chen, Y.-L., Cheng, H., Chopra, P., Dai, X., Dixon, M., Eldan, R., Fragoso, V., Gao, J., Gao, M., Gao, M., Garg, A., Del Giorno, A., Goswami, A., Gunasekar, S., Haider, E., Hao, J., Hewett, R. J., Hu, W., Huynh, J., Iter, D., Jacobs, S. A., Javaheripi, M., Jin, X., Karampatziakis, N., Kauffmann, P., Khademi, M., Kim, D., Kim, Y. J., Kurilenko, L., Lee, J. R., Lee, Y. T., Li, Y., Li, Y., Liang, C., Liden, L., Lin, X., Lin, Z., Liu, C., Liu, L., Liu, M., Liu, W., Liu, X., Luo, C., Madan, P., Mahmoudzadeh, A., Majercak, D., Mazzola, M., Mendes, C. C. T., Mitra, A., Modi, H., Nguyen, A., Norick, B., Patra, B., Perez-Becker, D., Portet, T., Pryzant, R., Qin, H., Radmilac, M., Ren, L., de Rosa, G., Rosset, C., Roy, S., Ruwase, O., Saarikivi, O., Saied, A., Salim, A., Santacroce, M., Shah, S., Shang, N., Sharma, H., Shen, Y., Shukla, S., Song, X., Tanaka, M., Tupini, A., Vaddamanu, P., Wang, C., Wang, G., Wang, L., Wang, S., Wang, X., Wang, Y., Ward, R., Wen, W., Witte, P., Wu, H., Wu, X., Wyatt, M., Xiao, B., Xu, C., Xu, J., Xu, W., Xue, J., Yadav, S., Yang, F., Yang, J., Yang, Y., Yang, Z., Yu, D., Yuan, L., Zhang, C., Zhang, C., Zhang, J., Zhang, L. L., Zhang, Y., Zhang, Y., Zhang, Y., and Zhou, X. (2024). Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219.*

Bharadwaj, S., Varun, R., Aishwarya, P. S., Sudhakar, M., and Naik, S. (2022). Resume screening using NLP and LSTM. *International Conference on Inventive Computation Technologies*, IEEE, 238-241.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. and Agarwal, S. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877-1901.

Chaturvedi, S. and Chaturvedi, R. (2025). Who gets the callback? Generative AI and gender bias in hiring. *arXiv preprint arXiv:2504.21400.*

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. (2023). QLoRA: Efficient finetuning of quantized LLMs. *Advances in Neural Information Processing Systems*, 36, 10088-10115.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT*, 4171-4186.

Equal Employment Opportunity Commission. (1978). Uniform guidelines on employee selection procedures. *Federal Register*, 43(166), 38290-38315.

European Parliament and Council. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (AI Act). *Official Journal of the European Union*.

Evertz, J., Kim, S., Park, H., and Liu, X. (2025). FAIRE: Assessing racial and gender bias in AI-driven resume evaluations. *arXiv preprint arXiv:2504.01420.*

Gan, C., Zhang, Q., and Mori, T. (2024). Application of LLM agents in recruitment: A novel framework for resume screening. *arXiv preprint arXiv:2401.08315.*

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2021). LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685.*

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Singh Chaplot, D., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L., Lachaux, M., Stock, P., Le Scao, T., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. (2023). Mistral 7B. *arXiv preprint arXiv:2310.06825.*

Kaggle. (2023). *Resume Dataset (24 categories, 2,484 resumes)*. Retrieved from the Kaggle online repository.

Kaur, R., Singh, K., and Rao, A. (2025). Resume2Vec: Transforming applicant tracking systems with intelligent resume embeddings for precise candidate matching. *Electronics*, 14(4), 794. doi:10.3390/electronics14040794.

Lavi, D., Medentsiy, V., and Graus, D. (2021). ConsultantBERT: Fine-tuned Siamese sentence-BERT for matching jobs and job seekers. *arXiv preprint arXiv:2109.06501.*

Liu, R., Chen, J., Wang, X., Sun, P., Han, S., and Lin, Z. (2024b). A review of synthetic data generation methods for natural language processing. *ACM Computing Surveys*, 57(3), 1-38.

Liu, S.-Y., Wang, C.-Y., Yin, H., Molchanov, P., Wang, Y.-C. F., Cheng, K.-T., and Chen, M.-H. (2024a). DoRA: Weight-decomposed low-rank adaptation. *International Conference on Machine Learning (ICML)*.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692.*

Meta AI. (2024). The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783.*

Min, B., Ross, H., Sulem, E., Veyseh, A. P. B., Nguyen, T. H., Sainz, O., Agirre, E., Heintz, I., and Roth, D. (2023). Recent advances in natural language processing via large pre-trained language models: A survey. *ACM Computing Surveys*, 56(2), 1-40.

Park, J., Lee, S., and Kim, T. (2024). A survey on synthetic data for tabular and textual machine learning: Risks and best practices. *Information Fusion*, 108, 102366.

Peña, A., Serna, I., Morales, A., Fierrez, J., and Ortega-Garcia, J. (2025). Addressing bias in LLMs: Strategies and application to fair AI-based recruitment. *Proceedings of the 39th AAAI Conference on Artificial Intelligence*, 12345-12353.

Salinas, A., Shah, P., Liu, A., and Morstatter, F. (2025). Robustly improving LLM fairness in realistic settings via interpretability-guided interventions. *arXiv preprint arXiv:2506.10922.*

Sanford Heisler Sharp McKnight LLP. (2025). AI bias in hiring: Algorithmic recruiting and your rights. *Law Firm White Paper*.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108.*

Seregina, I., Schiller, M., and Schaefer, R. (2025). Parameter-efficient fine-tuning for HAR: Integrating LoRA and QLoRA into transformer models. *arXiv preprint arXiv:2512.17983.*

Skondras, P., Zervas, P., and Tzimas, G. (2023). Generating synthetic resume data with large language models for enhanced job description classification. *Future Internet*, 15(11), 363. doi:10.3390/fi15110363.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., and Lample, G. (2023a). LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971.* doi.org/10.48550/arXiv.2302.13971.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., and Bikel, D. (2023b). Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288.*

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998-6008.

Vishaline, A. R., Kumar, R. K., Sai Pramod, V. V. N. S., Vignesh, K. V. K., and Sudheesh, P. (2024). An ML-based resume screening and ranking system. *2024 International Conference on Signal Processing, Computation, Electronics, Power and Telecommunication (IConSCEPT)*, IEEE, 1-6.

Wandelt, S., Sun, X., Zhang, A., and Pan, J. (2024). The role of artificial intelligence in human resource management: A bibliometric review. *Computers in Human Behavior Reports*, 13, 100371.

Wilson, K. and Caliskan, A. (2024). Gender, race, and intersectional bias in resume screening via language model retrieval. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7, 1578-1591.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M. and Davison, J. (2020). Transformers: State-of-the-art natural language processing. *Proceedings of EMNLP: System Demonstrations*, 38-45.

Yadav, A., Patel, M., and Sharma, P. (2025). Rule-based and AI-driven resume analysis system: NLP and ATS scoring for automated screening. *International Conference on Advances in Research and Computing (ICARC)*, IEEE, 1-6.

Younes, M. T., Walid, O., Shaban, K., Hamdi, A., and Hassan, M. (2025a). Augmented fine-tuned LLMs for enhanced recruitment automation. *arXiv preprint arXiv:2509.06196.*

Younes, M. T., Walid, O., Shaban, K., Hamdi, A., and Hassan, M. (2025b). MSLEF: Multi-segment LLM ensemble fine-tuning in recruitment. *arXiv preprint arXiv:2509.06200.*

Zhang, J., Li, Q., Wang, H., and Sun, F. (2024). Optimizing large language models with an enhanced LoRA fine-tuning algorithm for efficiency and robustness in NLP tasks. *arXiv preprint arXiv:2412.18729.*

Zhang, Q., Chen, M., Bukharin, A., He, P., Cheng, Y., Chen, W., and Zhao, T. (2023). AdaLoRA: Adaptive budget allocation for parameter-efficient fine-tuning. *International Conference on Learning Representations*.

Zhang, X. and Kuhn, P. (2024). Gender bias in algorithmic hiring: Evidence from Chinese online job platforms. *Journal of Labor Economics*, 42(4), 1023-1067.

Zhao, L., Iyer, R., and Fernandez, M. (2026). Fairness-aware fine-tuning of large language models for equitable talent screening: A 2026 perspective. *arXiv preprint arXiv:2601.04567.*

Xin Liu is an Associate Senior Engineer at Qingdao Human Resources Development Research and Promotion Center, Information Data Resources Management Department, Qingdao, Shandong Province, 266001, China. She received her master's degree in computer application technology from Changchun University of Technology in 2007. She has nineteen years of working experience in the field of big data and artificial intelligence, and has previously served as big data engineer at the Information Center of Human Resources and Social Security Bureau, Chief Designer at Juhaokan Technology Co., Ltd., and Senior Engineer at Qingdao Hisense Media Network Technology Co., Ltd. Her research interests focus on scenario-based applications of artificial intelligence in the human resources and social security industry, including big data-assisted decision-making analysis, judgment, early warning monitoring, and qualification certification, as well as LLM-powered full-chain support for arbitration, document element verification, and skill evaluation supervision.

Chunfang Su is an Associate Professor at the Department of Computer Science, Jiangyin Polytechnic College, Wuxi, Jiangsu Province, China. She received her master's degree in computer application technology from Changchun University of Technology in 2007. She has eighteen years of teaching experience in Internet of Things (IoT) technology and previously served as an embedded development engineer at Northeast Software Co., Ltd. Her research interests include data mining, machine learning, and the development and application of IoT intelligent agents. She has published three papers on human behavior recognition, including "Complex Activity Representation and Recognition Using Dirichlet Multinomial Mixed Model" and "Activity Recognition System for Dementia in Smart Homes based on Wearable Sensor Data."