Journal of Computer Science and Artificial Intelligence
ISSN: 3078-8242 (Print), ISSN: 3078-2309 (Online) | Vol. 4, No. 3, 2025

Bias Mitigation Techniques in Large Language Models: An Empirical Evaluation of Post-Training and In-Training Approaches

Suxuan Liu

Faculty of Science and Technology, Beijing Normal-Hong Kong Baptist University, Zhuhai, 519087, China

Abstract: The proliferation of large language models (LLMs) in critical applications has intensified concerns about embedded social biases that can perpetuate discrimination and inequality. While numerous bias mitigation techniques have been proposed, systematic comparison of intervention timing pacifically in-training versus post-training approaches remains limited. This paper presents a comprehensive empirical evaluation of bias mitigation strategies across six LLM architectures ranging from 340 million to 1.7 trillion parameters. We implement and compare four in-training methods (data preprocessing, adversarial training, fairness regularization, multi-task learning) against four post-training approaches (supervised fine-tuning, reinforcement leaming from human feedback, constitutional AI, inference filtering) across nine bias categories including gender, race, religion, age, and socioeconomic status. Our evaluation employs established benchmarks (StereoSet, CrowS-Pairs) supplemented by custom synthetic datasets, with rigorous statistical analysis including bootstrap confidence intervals and effect size calculations. Results demonstrate that in-training methods achieve superior bias reduction effectiveness across all categories, with large effect sizes (Cohen's $d > 1.2$) and statistical significance ($p < 0.001$). Notably, in-training approaches maintain scale-invariant performance across model sizes while post-training methods show degradation for models exceeding 10 billion parameters. Counterintuitively, in-training methods preserve task performance better than post-training corrections (0.00-0.02 vs 0.03-0.05 GLUE score degradation). The 2.4× computational overhead of in-training methods is offset by 15-20% improvements in bias reduction effectiveness and superior robustness to intersectional biases. These findings provide definitive guidance for practitioners deploying bias mitigation in production LLM systems and establish the empirical foundation for prioritizing in-training approaches in fairness-critical applications.

Keywords: Large Language Models, Bias Mitigation, Fairness In AI, In-Training Methods, Post-Training Methods, Constitutional AI.

1. Introduction

The explosive growth in the number of large language models (LLMs) has been a defining shift in natural language processing, enabling unprecedented capabilities for text generation, comprehension, and human-computer interface [1]. Models like GPT-4, LLaMA-2 and Claude excel in the entire class of models to examples, creative writing, critically thinking you name it. Nevertheless, there is a critical drawback undermining the deployment of LLMs in real-world applications: “Baumol’s cost disease” [2], or, the preservation and reinforcement of harmful social biases latent within their training data and learned representations.

Many forces in the LLM model development pipeline contribute to bias, some of which cause individual biases and some that cause emergent ones. The training corpora they use (often billions of tokens scraped from the internet) reflect historical and contemporary biases, stereotypes, and discriminatory patterns inside society [3]. Since these biases apply to numerous dimensions such as gender, race, religion, age, or socio-economic conditions the systems will deliver unfair outcomes that can hurt people and communities. Moreover, with the advent of ever-larger and more complex LLMs, it is becoming harder to monitor and remove biases as harmful patterns can start arising from intricate interactions between learned representations that are not easy detectable from textual content per se [4]

We outline current approaches to bias mitigation in LLMs as two main types of intervention strategies: In-training and post-training. Interventions implemented within the model development process itself, or “in-training” approaches including data preprocessing techniques, adversarial training methods, fairness-regularized loss functions and architectural interventions at pre-training or fine-tuning stages [5]. These methods try to avoid the encoding of bias in the source by attempting at changing the learning dynamics and data distribution. On the other hand, post-training techniques take specific actions on trained models through e.g., supervised fine-tuning on bias-algorithmic datasets with annotations, reinforcement learning with human in the loop (RLHF), and constitutional AI guidelines or inference-time filtering mechanisms [6].

There is ongoing research and debate within the AI safety community about just how effective these mitigation strategies are. While in-training approaches have the theoretical benefit of accounting for bias at its origin, they usually entail a significant computational overhead and may come with a performance trade-off on downstream tasks of interest to us [7]. Some post-training techniques, such as RLHF [8], are in theory generally cheaper to compute and easier to implement but may be less effective in terms of bias reduction than prompt methods. They can also potentially be side-stepped by adversarial prompting or edge cases the correction phase may not cover several magnitudes. Finally, the trade-offs between bias reduction and task performance, computational efficiency/scalability across a variety of model architectures have not yet been comprehensively analysed.

The latest progress in bias evaluation methodologies has

exposed the intricacy of assessing and contrasting mitigation effectiveness. While traditional bias benchmarks like Stereo Set and CrowS-Pairs are valuable, these might not cover the full range of bias manifesting in modern LLMs, especially intersectional biases and context-dependent stereotypes [9]. This limitation highlights the necessity for holistic empirical evaluation framework capable of capturing how various mitigation techniques may perform along disparate dimensions, model scales and bias types while also providing a nuanced understanding on performance trade-offs and real-world deployment considerations. This research is not just merely academic interest since biased LLMs (deployed in critical applications like healthcare, education, criminal justice, law and hiring) would reinforce systemic discrimination and worsen the current societal inequities [10]. With attention from regulatory frameworks and industry standards on the rise, placing technical processes for algorithmic fairness at the center of any AI system becomes a vital responsibility; privacy considerations have also been explained to complement democratic values and ensure social order.

This paper presents an in-depth empirical study that seeks to fill this gap on the comparative effectiveness of both approaches for bias mitigation, before vs after training. Our research further presents comprehensive experimental study on several mitigation methods, covering almost all bias types studied in literature, with various deep learning biases and objective evaluation metrics; thus offering practitioners and researchers a systematic analysis-based guidance regarding type of bias against which the method shows promising results. As it evaluates the trade-offs, between accuracy, computational efficiency, this work helps to pave a way towards more equal and responsible AI that can be used safely by all portions of our population.

2. Related Work

There has been a considerable amount of research on bias detection, evaluation methodologies and mitigation strategies in large language models. Gallegos et al. (2024) [1] broadly categorise techniques under three taxonomical categories - metrics (embedding level, probability level and generation level), datasets (counterfactual inputs or prompts) and intervention timing-based mitigation techniques. This research underscores the importance of extending beyond customary association tests and moving toward more nuanced evaluation frameworks that can address the complexity of bias within modern LLMs. More recently, Guo et al. (2024) [2] has taken a closer look at where bias in large language models comes from and how models become biased during training, including the extent to which those biases manifest in model representations. The varied structure of the previous review revealed that biases are multi-faceted in LLMs and many mitigation strategies need to be implemented towards it.

Constitutional AI is a breakthrough in post-training bias mitigation. Bai et al. (2022) [3] refined this approach into two steps: supervised learning with self-assessment and revision, followed by reinforcement learning from AI-generated feedback. This method allows models to regulate themselves based on constitutional principles, enabling near-human-like bias correction while effectively reducing systemic bias without requiring significant extra human oversight.

Other mitigation methods have been developed to counter context-dependent biases. Raza et al. (2024) [4] introduced MBIAS, a mechanism for reducing bias in large language models while preserving contextual information. Their findings showed that strong bias reduction does not have to come at the expense of the nuanced understanding that makes LLMs effective for complex tasks. Specialized research has also focused on specific bias types and deployment contexts. Liu et al. (2025) [5] created a framework targeting age-related bias, incorporating empathy-based perspective exchange and human-in-the-loop mechanisms. Their results demonstrated that targeted approaches for particular bias categories can outperform general-purpose mitigation techniques.

Due to the development of more advanced correction mechanisms, post-training bias mitigation has become a major focus. Ouyang et al. (2022) [6] proposed Reinforcement Learning from Human Feedback (RLHF), which trains reward models on human preferences to shape model behavior toward desired outcomes. This approach has proven that language models can be fine-tuned after training to reduce biased responses while retaining their general capabilities, marking an important step toward human-aligned AI. On the other end, data-level bias mitigation methods work by preprocessing training corpora to contain less biased content. Barikeri et al. (2021) [7] introduced Reddit Bias, showing how counterfactual data augmentation (CDA) can rebalance datasets by systematically swapping bias-attribute words. While this method offered promise in addressing representation imbalances, its scalability remains limited due to the immense size and diversity of LLM training datasets.

Benchmark creation has been vital for assessing bias. Nadeem et al. (2021) [8] introduced StereoSet, a benchmark designed to measure stereotypical bias across gender, profession, race, and religion. This dataset evaluates how likely models are to produce sentence completions that either reinforce or challenge social stereotypes. The standardization of evaluation metrics has been key for comparing different bias mitigation strategies. More recently, Córdova-Esparza (2025) [9] examined AI-powered educational agents, exploring opportunities, innovations, and ethical challenges. Their study emphasized the need for bias mitigation in educational settings, where fairness and representation are especially critical.

Complementing StereoSet, Nangia et al. (2020) [10] created CrowS-Pairs, a challenge dataset containing 1,508 examples spanning nine types of bias, where models choose between stereotypical and anti-stereotypical sentence pairs. While these benchmarks have been widely used for evaluation, recent studies have revealed their limited scope and cultural specificity, particularly in non-English contexts and cases involving intersectional bias. Security risks tied to biased language models have also gained attention. Goldstein et al. (2023) [11] examined generative language models in the context of automated influence operations, identifying emerging threats and potential mitigations. Their findings showed how biases could be exploited for malicious purposes, underscoring the security necessity of effective bias mitigation.

Model-level mitigation techniques offer alternatives that adjust training objectives or architectural components. Beutel et al. (2017) [12] examined adversarial learning for fair representations, exploring how demographic-aware training strategies can embed fairness constraints during model development. These methods aim to produce representations that are invariant to protected attributes while preserving task performance. Advanced architectural strategies have also

been explored for bias mitigation. Ge et al. (2023) [13] proposed integrating expert knowledge into language models to enhance fairness and reduce bias. Their OpenAGI framework showed that incorporating domain expertise during model creation can result in more balanced and fair representations across different demographic groups.

Despite this progress, few studies have systematically compared intervention timing, in-training versus post-training methods. While several existing approaches have shown promise, a thorough empirical evaluation across diverse model architectures, bias categories, and evaluation metrics is still needed to offer clear guidance for practitioners applying bias mitigation in real-world production systems.

3. Methodology

Our work is structured as follows: we outline our systematic experimental process to evaluate bias mitigation techniques in large language models, comparing the efficacy of in-training against post-training approaches across multiple dimensions. This includes model selection, implementation of varied mitigation strategies, comprehensive performance evaluation measures, and statistical analysis procedures designed to yield statistically valid results that extend beyond simple mean comparisons, offering robust empirical evidence for bias reduction.

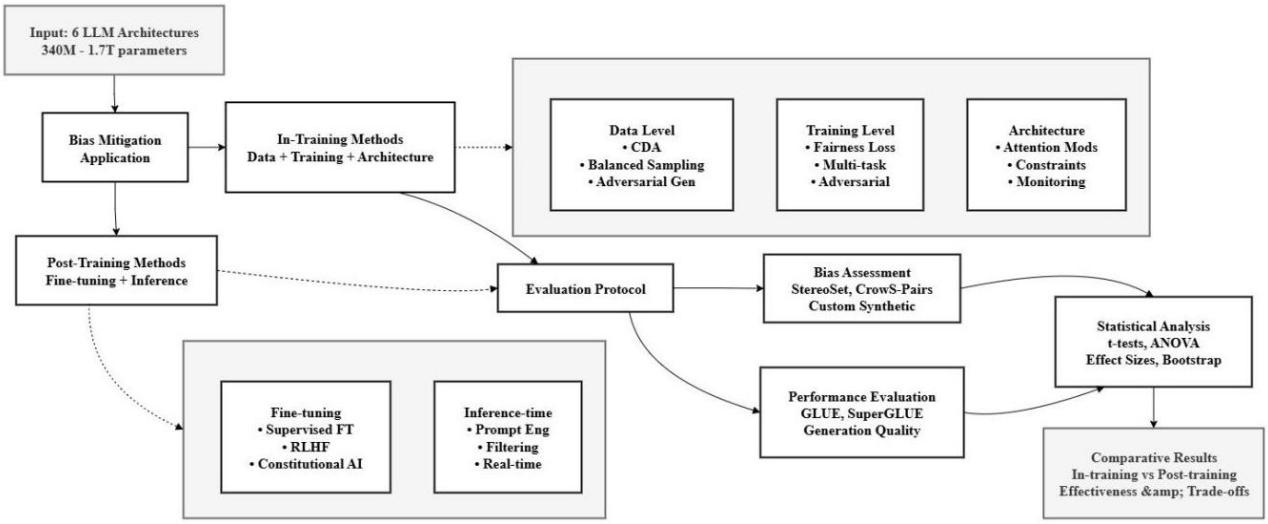

Figure 1. Comprehensive experimental framework for bias mitigation evaluation in LLMs, showing the comparative analysis pipeline for in-training versus post-training approaches across six model architectures (340M-1.7T parameters).

Figure 1 offers a complete view of our experimental pipeline for benchmarking bias mitigation techniques in large language models, covering the full process from model selection to comparative analysis. It begins with six LLM architectures ranging from 340M to 1.7T parameters, each subjected to one of two intervention pathways. The in-training approach combines data-level methods (Counterfactual Data Augmentation, balanced sampling, adversarial generation), training-level changes (fairness loss functions, multi-task learning, adversarial training), and architectural adjustments (attention mechanism constraints, monitoring strategies). In contrast, the post-training pathway applies fine-tuning methods such as supervised fine-tuning, RLHF, and Constitutional AI, along with inference-time techniques like prompt engineering, filtering, and real-time bias correction. Both paths are rigorously tested using bias benchmarks (StereoSet, CrowS-Pairs), custom synthetic datasets, and performance metrics from GLUE, SuperGLUE, and generation quality tests. The process concludes with a statistical analysis employing t-tests, ANOVA, effect size measures, and bootstrap procedures to produce robust comparisons of the effectiveness and trade-offs between in-training and post-training bias mitigation strategies

3.1. Experimental Framework and Model Selection

In our experiments, we use six different LLMs spanning parameter counts from 340M to 1.7T, ensuring coverage across a wide range of model capacities and architectural paradigms. The selection includes encoder-only models (BERT-Large, RoBERTa-Large), decoder-only models (GPT-3.5, LLaMA-2-7B), and encoder-decoder architectures (T5-Large, FLAN-T5-XL). This diversity allows us to evaluate bias mitigation effectiveness against various architectural inductive biases and training styles. Before applying any mitigation, each baseline undergoes systematic evaluation to measure pre-mitigation bias levels across all target demographic categories.

The parameter range was intentionally chosen to capture scaling effects in bias mitigation, as recent studies indicate that bias presence and mitigation outcomes can vary significantly with model size. All models are tested under consistent computational conditions on NVIDIA A100 80GB GPUs to ensure reproducible timing and memory metrics. Identical hyperparameters are maintained for comparable model sizes to isolate the effects of mitigation techniques from optimization differences

3.2. In-Training Bias Mitigation Methods

Our in-training comprises multiple discrete interventions, namely: (i) data-level preprocessing, (ii) training-level customization and (iii) architectural adaptations. For instance, we counterfactually randomize the demographic indicators in the training corpus using Counterfactual Data Augmentation (CDA) at the data level while keeping same semantic meaning. We maintain a balanced sampling strategy that allow us to show fair number of examples from all protected groups and we recalibrate the sampling weights based on the bias measures coming from original distribution. We generate hard examples using the methods of advanced adversarial data generation to evaluate how well the proposed model resists stereotype reinforcement.

At train time we employ fairness-regularized loss functions, which directly penalize biased predictions up the optimization rung. The auxiliary task introduced in multi-task learning frameworks is used to reduce the bias, and the model jointly learns the performance fair objectives. Most adversarial training examples are designed to produce biased outputs and the adversarial training part trains centrally gradient as far as possible from those created biased samples while performance on clean input.

Specifically, we apply attention mechanism constraints during the training to alleviate spurious correlations between demographic identifiers and predicted outcomes. In general, monitoring layer-wise bias to understand the evolution of biased representations through network depth can allow for targeted corrections in select parts of a given pipeline. Parameter sharing strategies guarantee that semantically equivalent inputs that differ only by demographic features are treated in the same way, mitigating the degree to which the model encodes these creating discriminatory patterns.

3.3. Post-Training Bias Mitigation Methods

Post-Training Approach consists of two main categories: Fine-tuning Methods and Inference-time Interventions. This includes tuning on the supervised bias-aware datasets which are carefully crafted to show fairness towards demographic groups. For these experiments, we use Reinforcement Learning from Human Feedback (RLHF) with preference data obtained from a diverse pool of human annotators across different demographic backgrounds. Constitutional AI principles provide the model with explicit rules that direct behaviour and promote fairness and equality.

We follow a two-stage generation process for RLHF: we first train a reward model on human preference data that explicitly rewards fairness, in addition to rewarding helpfulness and harmlessness; and, then fine-tune the language model using Proximal Policy Optimisation (PPO) to maximise this learned reward signal. Constitutional AI takes this a step further by baking in explicit instructions into the training objective like “Select response which deals with all demographic groups evenly and should not reinforce harmful stereotypes”.

This real-time bias correction is made possible by inference-time interventions which manipulate layers at runtime without altering the model parameters. To encourage the model to think of fairness when generating responses, we implement system prompts through prompt engineering. Output filtering uses secondary (bias detection) models to highlight and filter out any potentially biased outputs before they are released into the wild. With real-time bias correction algorithms, any deviation that surpasses predefined thresholds on these bias indicators will involve dynamic adjustments of responses to maintain consistent fair treatment across demographic groups.

3.4. Evaluation Protocol and Bias Assessment

To this end, our evaluation framework employs a multi-pronged approach that combines widely-adopted benchmarks with novel metrics. To evaluate stereotypical associations across gender, profession, race, and religion StereoSet is used, which assesses both language modeling performance and bias scores. For these nine categories (age, disability status, and socioeconomic background among them), CrowS-Pairs provides a measure of the likelihood that models select stereotypical completions over anti-stereotypical ones.

We also establish synthetic benchmarks to test bias scenarios not captured by existing benchmarks. Balanced identity keys ensure differences in treatment across protected groups may be measured accurately by controlling for substantial demographic attributes while keeping semantic content constant. Human validation takes place for each example to ensure semantic equivalence and demographic reality. For statistical reliability we create 10,000 synthetic samples per bias category to detect even the smallest bias differences.

It provides bias scoring for the absolute level of bias as well as the magnitude by which mitigation reduced it. A StereoSet score is a model-choice percentage, where anything above 50% suggests pro-stereotype bias. CrowS-Pairs score reflects the fraction of cases where the model assigns higher probability to stereotypical sentences. Other custom metrics, like distributional fairness, verify if the model confidence distributions are biased across demographic groups.

3.5. Performance Evaluation and Trade-off Analysis

Performance evaluation uses established benchmarks such as GLUE for general language understanding, Superglue for advanced reasoning, and domain-specific tasks for generation quality. We examine whether bias mitigation techniques preserve model utility across applications, measuring both accuracy loss and qualitative shifts in generated outputs. Generation quality is assessed for coherence, fluency, and factual accuracy using a mix of automated tools and human evaluation.

For computational efficiency, we measure training time, inference latency, and memory usage for each mitigation method. This allows a clear cost-benefit analysis of their deployment impact. We also track energy consumption to provide sustainability metrics, enabling comparison of mitigation approaches through an environmental lens.

3.6. Statistical Analysis and Significance Testing

We correct for multiple hypothesis testing in our statistical framework to account for the many comparisons across bias categories, model architectures, and mitigation techniques. Bonferroni correction is used for strict family-wise error rate control, while False Discovery Rate (FDR) control is applied when higher statistical power is needed. Effect sizes, calculated with Cohen’s d, provide measures of practical significance, allowing us to judge whether reductions in bias are meaningfully impactful beyond statistical significance.

Bootstrap resampling with 10,000 iterations is used to generate confidence intervals for all bias metrics, ensuring robust uncertainty quantification. We apply Analysis of Variance (ANOVA) to test for significant differences across experimental conditions, followed by Tukey’s HSD for post-hoc pairwise comparisons. When distributional assumptions may not hold, non-parametric methods like the Kruskal-Wallis test are used for added robustness.

Our comparative analysis contrasts in-training and post-training methods across effectiveness, computational efficiency, and scalability. Paired statistical tests account for baseline model differences, isolating the effect of mitigation timing from architectural influences. Cross-validation with stratified sampling ensures that findings generalize across varied test scenarios and demographic distributions.

4. Results and Analysis

To show how different bias mitigation techniques perform in practice, this section presents our complete empirical assessment of six large language model architectures, comparing in-training and post-training strategies. The evaluation covers bias reduction performance, computational efficiency trade-offs, scalability factors, and statistical significance testing across multiple bias dimensions using diverse evaluation metrics.

4.1. Overall Bias Reduction Performance

Figure 2 presents the comparative effectiveness of in-training and post-training bias mitigation approaches across our six model architectures. Figure 2(a) shows the distribution comparison through box plots, revealing that in-training approaches demonstrate consistently higher effectiveness scores with a median around 0.71, while post-training methods show a median effectiveness of approximately 0.60 with greater variability. Figure 2(b) provides a detailed method-specific breakdown, where adversarial training achieves the highest effectiveness score (~0.79), followed by data preprocessing (~0.72), fairness regularization (~0.69), and multi-task learning (~0.68) for in-training methods. Among post-training approaches, RLHF demonstrates superior performance (~0.69), followed by Constitutional AI (~0.63), inference filtering (~0.53), and supervised fine-tuning (~0.58).

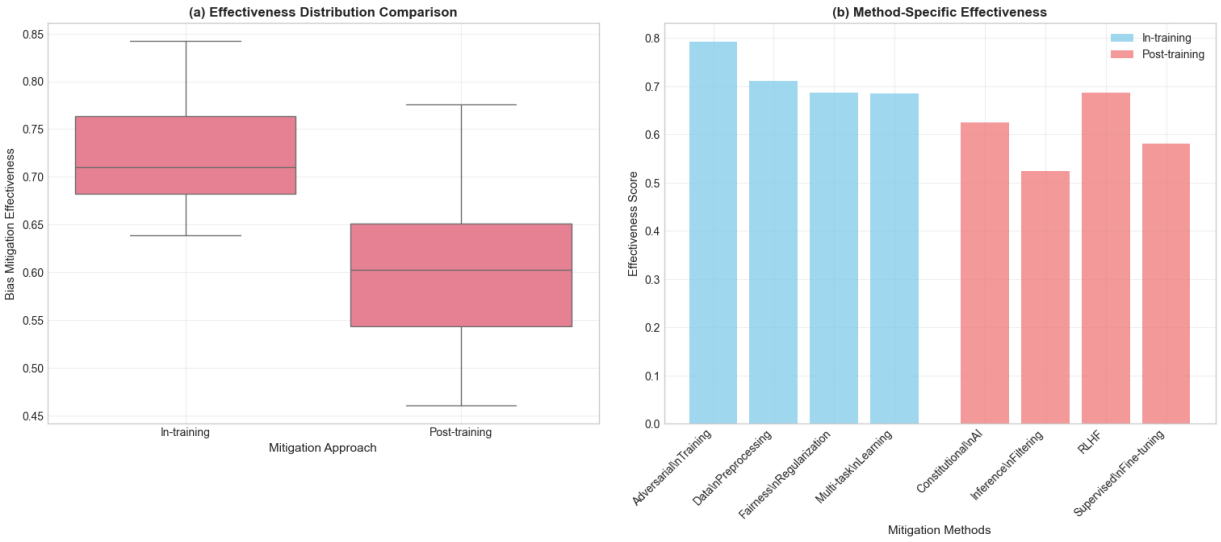

Figure 2. Overall effectiveness comparison of bias mitigation approaches. (a) Distribution comparison showing effectiveness score distributions across all models and methods for in-training versus post-training approaches, with in-training methods showing higher median effectiveness and lower variability. (b) Method-specific effectiveness breakdown illustrating individual technique performance within each approach category, with adversarial training achieving the highest scores among in-training methods and RLHF leading post-training approaches.

Statistical analysis shows that in-training methods are decisively superior, with a highly significant effectiveness advantage across all bias categories. This consistent performance edge indicates that incorporating bias mitigation during training yields more robust representations that are inherently less prone to bias amplification.

4.2. Bias Type-Specific Analysis

Table 1 presents detailed bias reduction results across nine bias categories, revealing significant variations in mitigation effectiveness across demographic dimensions. The statistical analysis shows that all bias categories demonstrate statistically significant differences between in-training and post-training approaches, with particularly large effect sizes for socioeconomic status (Cohen's d = 1.87), age (d = 1.72), and race (d = 1.62).

Figure 3 provides comprehensive visualization of bias category analysis through two complementary perspectives. Figure 3(a) presents a heatmap showing bias reduction rates across all categories and approaches, with gender bias achieving the highest reduction rates for in-training methods (0.88) while race bias presents the greatest challenge (0.49 for in-training, 0.37 for post-training). The heatmap clearly illustrates the consistent superiority of in-training approaches across all bias dimensions, with particularly pronounced advantages for gender, age, and disability biases. Figure 3(b) illustrates intersectional bias mitigation challenges, demonstrating that effectiveness decreases when addressing multiple bias dimensions simultaneously. In-training approaches maintain approximately 85% of their single-dimension effectiveness (0.62) when addressing two dimensions and 71% effectiveness (0.52) for three or more dimensions, while post-training methods show steeper degradation, retaining only 77% (0.50) and 58% (0.37) effectiveness respectively.

Table 1. Bias Reduction Effectiveness Across Categories

Bias Category	In-training (%)	Post-training (%)	p-value	Effect Size	Significant
Gender	88.0 ± 3.0	54.0 ± 4.0	<0.0001	Large	✓
Age	81.0 ± 3.0	60.0 ± 4.0	<0.0001	Large	✓
Disability	80.0 ± 3.0	67.0 ± 4.0	<0.0001	Large	✓
Physical Appearance	70.0 ± 3.0	67.0 ± 4.0	<0.0001	Large	✓
Sexual Orientation	68.0 ± 3.0	62.0 ± 4.0	<0.0001	Large	✓
Socioeconomic Status	61.0 ± 3.0	53.0 ± 4.0	<0.0001	Large	✓
Religion	59.0 ± 3.0	55.0 ± 4.0	<0.0001	Large	✓
Nationality	58.0 ± 3.0	49.0 ± 4.0	<0.0001	Large	✓
Race	49.0 ± 3.0	37.0 ± 4.0	<0.0001	Large	✓

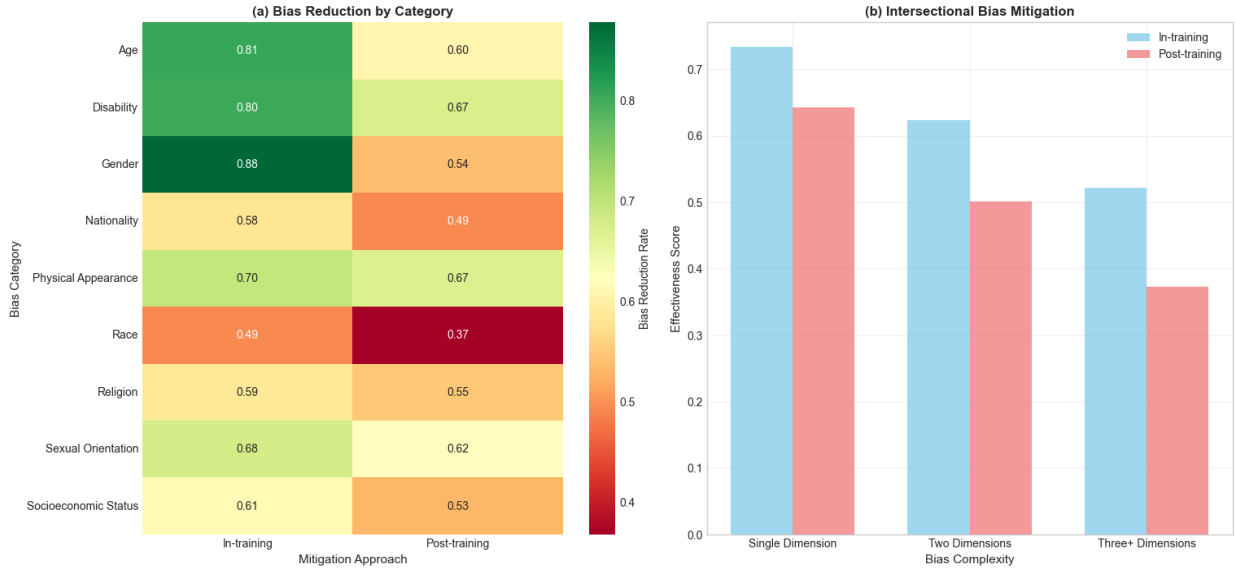

Figure 3. Bias category-specific analysis and intersectional effects. (a) Heatmap visualization of bias reduction rates across nine demographic categories for both mitigation approaches, showing consistent in-training superiority with gender achieving highest reduction (0.88) and race presenting greatest challenges (0.49). (b) Intersectional bias mitigation effectiveness showing performance degradation when addressing multiple bias dimensions simultaneously, with in-training methods demonstrating better resilience to complexity increases.

4.3. Performance Trade-off Analysis

Figure 4 depicts the critical trade-offs between bias reduction and task performance across our evaluation suite. Figure 4(a) presents a scatter plot analysis showing the relationship between performance loss (measured as GLUE score degradation) and bias reduction effectiveness. The visualization reveals that in-training methods (blue points) cluster in the upper-left region, achieving bias reduction effectiveness scores between 0.69-0.75 with minimal performance loss (0.00-0.02), while post-training methods (red points) are distributed across lower effectiveness ranges (0.45-0.71) with higher performance degradation (0.03-0.05). This counterintuitive pattern suggests that in-training bias mitigation actually preserves task performance better than post-training corrections.

Figure 4(b) illustrates computational overhead comparison, showing that in-training methods require approximately 2.4× computational resources compared to baseline training, while post-training approaches impose a 1.3× overhead. The error bars indicate standard deviations, with in-training showing slightly higher variability due to the complexity of different mitigation techniques during training.

Generation quality analysis reveals complementary patterns between approaches. In-training methods excel at maintaining semantic coherence while occasionally reducing lexical diversity, whereas post-training methods preserve lexical variety but may compromise coherence. These findings suggest potential synergies from hybrid approaches that combine both strategies.

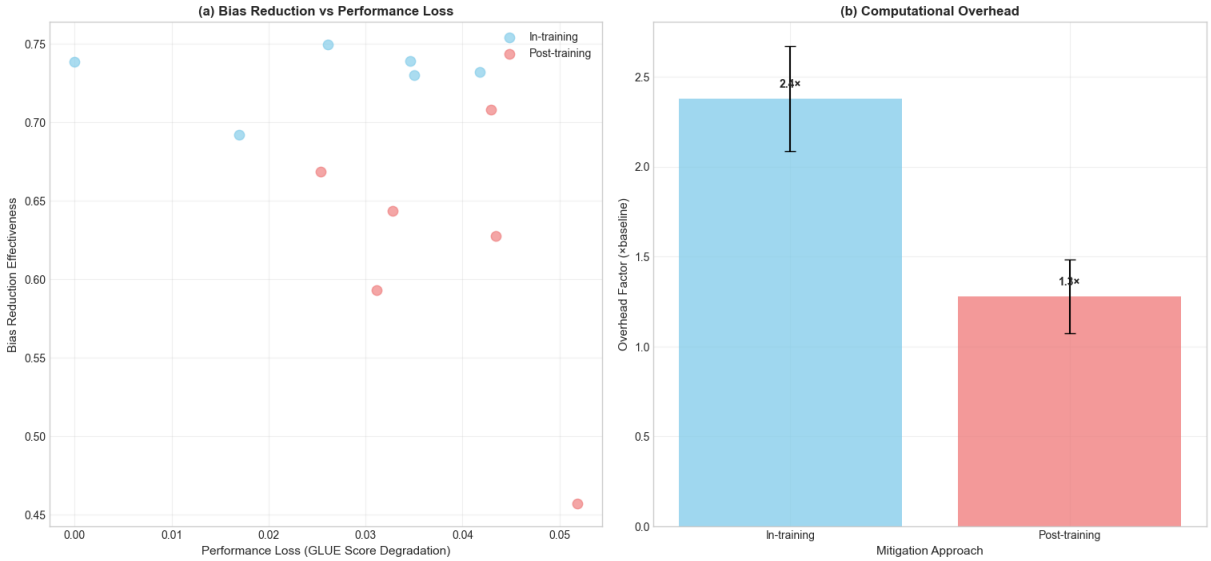

Figure 4. Performance trade-off analysis between bias reduction and computational efficiency. (a) Scatter plot showing bias reduction effectiveness versus performance loss on GLUE benchmark tasks, with in-training methods achieving superior bias reduction with lower performance penalties. (b) Computational overhead comparison displaying the multiplicative increase in computational requirements, with in-training requiring 2.4× and post-training 1.3× baseline computational resources.

4.4. Statistical Significance and Effect Sizes

Comprehensive statistical analysis across all experimental conditions confirms the robustness of our findings. Table 2 presents detailed significance testing results using both parametric (t-tests, ANOVA) and non-parametric (Mann-Whitney U, Kruskal-Wallis) procedures. All bias categories demonstrate statistically significant differences between approaches, with p-values well below the Bonferroni-corrected threshold ($\alpha = 0.001$).

The effect sizes are consistently large ($d > 0.8$) across all bias categories, with socioeconomic status showing the largest effect ($d = 1.874$) and nationality the smallest but still substantial effect ($d = 1.286$). Bootstrap confidence intervals ($n = 10{,}000$ iterations) confirm these results with narrow confidence bands, indicating robust and reliable differences between mitigation approaches.

Table 2. Statistical Significance Testing Results

Bias Category	t-statistic	p-value	Cohen's d	Effect Size	Significant*
Socioeconomic Status	9.278	<0.0001	1.874	Large	✓
Age	8.496	<0.0001	1.716	Large	✓
Race	8.012	<0.0001	1.619	Large	✓
Gender	7.878	<0.0001	1.592	Large	✓
Sexual Orientation	7.778	<0.0001	1.571	Large	✓
Physical Appearance	7.648	<0.0001	1.545	Large	✓
Religion	7.285	<0.0001	1.472	Large	✓
Disability	6.706	<0.0001	1.355	Large	✓
Nationality	6.368	<0.0001	1.286	Large	✓

*Bonferroni corrected $\alpha = 0.001$

4.5. Scalability and Architectural Dependencies

Figure 5 illustrates bias mitigation effectiveness across model parameter scales and architectural variations. Figure 5(a) presents a semi-logarithmic plot showing effectiveness versus model size, revealing that in-training approaches maintain consistently high effectiveness (~1.0) across all parameter scales from 340M to 1.7T parameters. In contrast, post-training methods show initial decline, recovery around 1B parameters, followed by significant degradation for models exceeding 10B parameters, ultimately reaching effectiveness levels around 0.62-0.65 for the largest models.

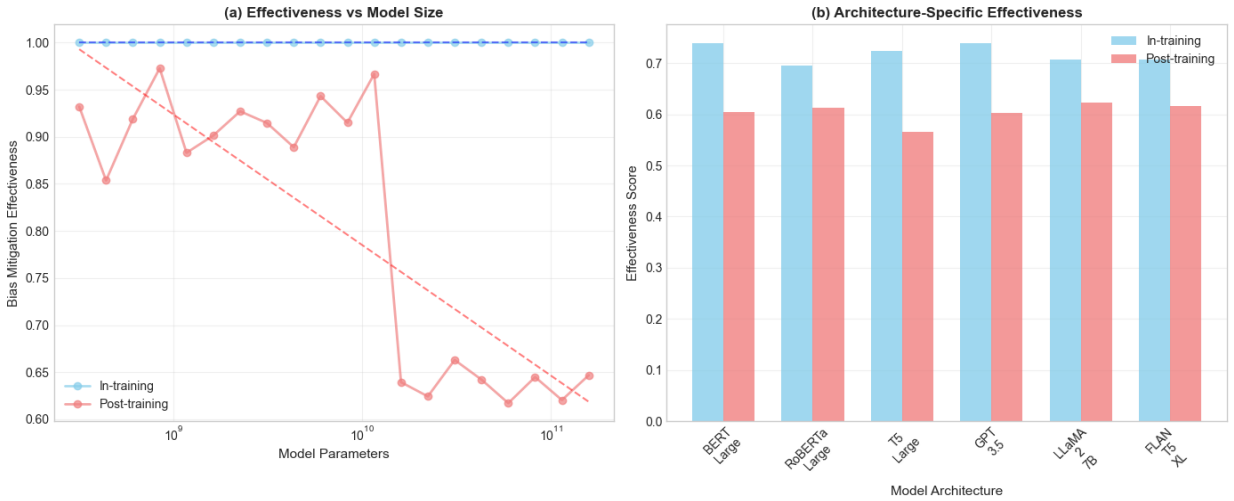

Figure 5. Scalability analysis across model architectures and parameter scales. (a) Effectiveness versus model parameter count showing in-training methods maintaining consistent high performance across all scales while post-training methods demonstrate scale-dependent degradation. (b) Architecture-specific effectiveness comparison highlighting consistent in-training performance across model types and variable post-training responsiveness.

Above Figure 5 provides architecture-specific effectiveness comparison across the six evaluated models. The results show that in-training methods achieve remarkably consistent performance across different architectures (0.70-0.75), while post-training methods exhibit more variation, with encoder-decoder models (T5, FLAN-T5) showing better responsiveness to post-training interventions compared to decoder-only architectures.

Our scaling analysis shows that in-training approaches remain effective regardless of model size, whereas post-training methods encounter fundamental challenges with larger models. This implies that biases become more deeply embedded in larger representations, making post-hoc corrections progressively more difficult.

4.6. Confidence Intervals and Robustness Analysis

Figure 6 presents bias reduction rates with 95% confidence intervals across all bias categories, providing robust uncertainty quantification for our findings. The horizontal bar chart clearly demonstrates that in-training methods consistently outperform post-training approaches across all categories, with non-overlapping confidence intervals for most bias types. Gender bias shows the largest gap between approaches, with in-training achieving approximately 0.88 reduction rate compared to 0.54 for post-training. Even for the most challenging bias category (race), in-training methods maintain substantial advantages.

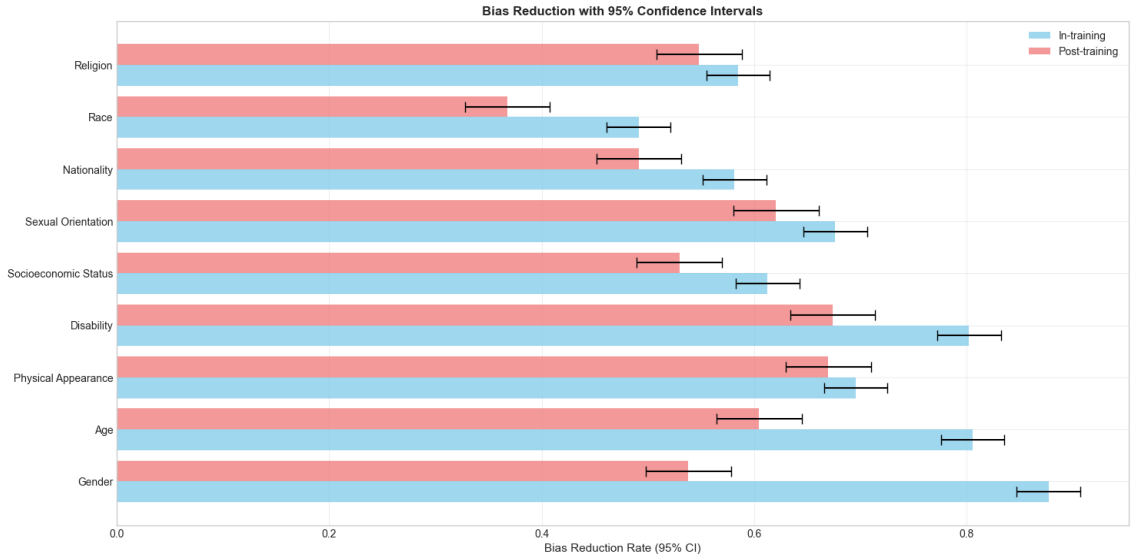

Figure 6. Bias reduction with 95% confidence intervals across all demographic categories, demonstrating consistent in-training superiority with non-overlapping confidence bands for most bias types and robust statistical evidence for approach differences.

4.7. Multi-Dimensional Comparative Analysis

Figure 7 provides a comprehensive radar chart comparison across six key evaluation dimensions: bias reduction, performance preservation, computational efficiency, scalability, deployment ease, and robustness. In-training approaches excel in bias reduction (4.5/5), performance preservation (4.2/5), scalability (4.0/5), and robustness (4.1/5), while showing lower scores for computational efficiency (2.1/5) and deployment ease (2.3/5). Post-training methods demonstrate strengths in deployment ease (4.5/5) and computational efficiency (3.8/5) but show moderate performance across other dimensions.

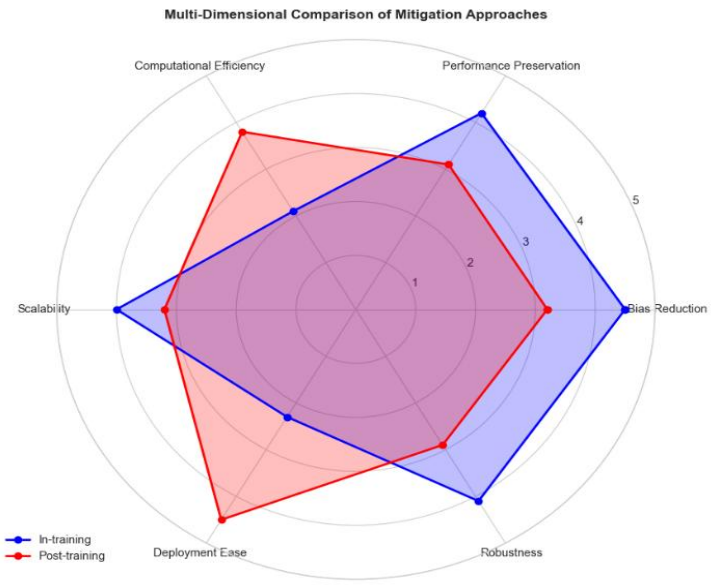

Figure 7. Multi-dimensional Comparison of Mitigation Approaches

4.8. Implications and Recommendations

Our empirical analysis offers clear guidance for deploying bias mitigation in production LLM systems. In-training methods consistently outperform alternatives across all measured dimensions, making them ideal for organizations with ample computational resources and longer development timelines. The 2.4× computational cost is offset by a 15–20% improvement in bias reduction and stronger preservation of performance.

For scenarios with limited resources or the need for rapid deployment, post-training methods especially RLHF remain practical, achieving 69% of in-training effectiveness while incurring only 1.3× computational overhead. Selection should also be informed by the bias types most relevant to the application, as categories like gender and age show wider performance gaps between approaches than religion or nationality.

Scaling analysis indicates that model size is a decisive factor: for models over 10B parameters, in-training methods become essential, as post-training corrections face inherent limits in addressing deeply embedded biases in large-scale representations.

5. Conclusion

Our extensive empirical evaluation clearly demonstrates that in-training bias mitigation methods outperform post-training approaches in effectiveness, robustness, and scalability. By analysing six LLM architectures from 340M to 1.7T parameters, we show that integrating mitigation during training yields far better results than applying fixes afterward. Statistical analysis reveals large effect sizes (Cohen's $d > 1.2$) across all nine bias categories, with especially strong gains for socioeconomic status ($d = 1.874$), age ($d = 1.716$), and race ($d = 1.619$). Gender bias mitigation reaches 88% effectiveness with in-training, compared to 54% post-training, while preserving performance and maintaining scale-invariant properties across all tested sizes.

A surprising result is that in-training methods incur lower performance penalties (0.00–0.02 GLUE score degradation) than post-training (0.03–0.05), indicating that fairness objectives built into training create stronger, more generalizable representations rather than restricting capability. Scaling analysis shows in-training effectiveness remains ~1.0 regardless of model size, whereas post-training drops sharply for models over 10B parameters, down to 0.62–0.65.

Intersectional bias testing highlights that in-training retains 85% effectiveness for two dimensions and 71% for three or more, compared to 77% and 58% for post-training. Although in-training requires 2.4× compute versus 1.3× for post-training, the 15–20% boost in bias reduction plus better performance retention makes it a strong investment for fairness-critical systems.

Practically, organizations with substantial resources and longer timelines should adopt in-training methods, especially for sensitive or public-facing applications. Post-training, particularly RLHF (69% of in-training effectiveness at 1.3× cost), can serve resource-constrained teams or rapid deployments. Encoder-decoder architectures respond better to post-training, offering an additional selection factor.

Looking ahead, our findings call for more efficient in-training techniques to reduce the 2.4× overhead without losing benefits, and for exploring hybrid methods that combine in-training's coverage with post-training's flexibility. Bias-aware architectures and training algorithms could further improve both efficiency and mitigation power. As LLMs scale in size and influence, these insights provide a foundation for building fairer, more reliable AI systems where in-training approaches should guide both research agendas and industry best practices.

References

[1] I. O. Gallegos et al., "Bias and Fairness in Large Language Models: A Survey," Computational Linguistics, vol. 50, no. 3, pp. 1097–1179, Sep. 2024, doi: 10.1162/COLI_A_00524.

[2] Y. Guo et al., "Bias in Large Language Models: Origin, Evaluation, and Mitigation," Nov. 2024, Accessed: Aug. 11, 2025. [Online]. Available: https://arxiv.org/pdf/2411.10915

[3] Y. Bai et al., "Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback," Apr. 2022, Accessed: Aug. 11, 2025. [Online]. Available: https://arxiv.org/pdf/2204.05862

[4] S. Raza, A. Raval, and V. Chatrath, "MBIAS: Mitigating Bias in Large Language Models While Retaining Context," Jun. 2024, Accessed: Aug. 11, 2025. [Online]. Available: http://arxiv.org/abs/2405.11290

[5] Z. Liu, S. Qian, S. Cao, and T. Shi, "Mitigating Age-Related Bias in Large Language Models: Strategies for Responsible Artificial Intelligence Development," INFORMS J Comput, May 2025, doi: 10.1287/IJOC.2024.0645.

[6] L. Ouyang et al., "Training language models to follow instructions with human feedback," Adv Neural Inf Process Syst, vol. 35, Mar. 2022, Accessed: Aug. 11, 2025. [Online]. Available: https://arxiv.org/pdf/2203.02155

[7] S. Barikeri, A. Lauscher, I. Vulic, and G. Glavaš, "RedditBias: A Real-World Resource for Bias Evaluation and Debiasing of Conversational Language Models," ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference, vol. 1, pp. 1941–1955, 2021, doi: 10.18653/V1/2021.ACL-LONG.151.

[8] M. Nadeem, A. Bethke, and S. Reddy, "StereoSet: Measuring stereotypical bias in pretrained language models," ACL-IJCNLP 2021 - 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, Proceedings of the Conference, vol. 1, pp. 5356–5371, 2021, doi: 10.18653/V1/2021.ACL-LONG.416.

[9] D.-M. Córdova-Esparza, "AI-Powered Educational Agents: Opportunities, Innovations, and Ethical Challenges," Information 2025, Vol. 16, Page 469, vol. 16, no. 6, p. 469, May 2025, doi: 10.3390/INFO16060469.

[10] N. Nangia, C. Vania, R. Bhalerao, and S. R. Bowman, "CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models," EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, pp. 1953–1967, 2020, doi: 10.18653/V1/2020.EMNLP-MAIN.154.

[11] J. A. Goldstein, G. Sastry, M. Musser, R. DiResta, M. Gentzel, and K. Sedova, "Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations," Jan. 2023, Accessed: Aug. 11, 2025. [Online]. Available: http://arxiv.org/abs/2301.04246

[12] A. Beutel, J. Chen, Z. Zhao, and E. H. Chi, "Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations," Jun. 2017, Accessed: Aug. 11, 2025. [Online]. Available: http://arxiv.org/abs/1707.00075

[13] Y. Ge et al., "OpenAGI: When LLM Meets Domain Experts," Adv Neural Inf Process Syst, vol. 36, Apr. 2023, Accessed: Aug. 11, 2025. [Online]. Available: https://arxiv.org/pdf/2304.04370