Received XX Month, XXXX; revised XX Month, XXXX; accepted XX Month, XXXX; Date of publication XX Month, XXXX; date of current version XX Month, XXXX.

Digital Object Identifier 10.1109/XXXX.2022.1234567

A Pruning Framework for Bias Mitigation in Large Language Models

AMRITA SINGH RAJPUT[1], VIJAY K. MADISETTI (FELLOW, IEEE) [2]

Corresponding author: Vijay K. Madisetti (vkm@gatech.edu)

ABSTRACT

Large Language Models (LLMs) encode social biases in their internal representations, but most mitigation methods require costly retraining or rely on pruning designed for compression, not fairness. We present a fairness-first layer pruning framework that treats transformer blocks as the intervention unit and ranks them with a hybrid importance score that combines utility, redundancy, and bias. Unlike prior structured pruning methods focused on perplexity or efficiency, our approach explicitly optimizes fairness metrics during layer selection and imposes strict gates that allow only configurations with bounded utility loss and improved scores on multiple bias measures. We evaluate open-weight causal LLMs across small, medium, and larger settings TinyLlama-1.1B, Qwen2.5-0.5B, Llama-3.2-1B, Gemma-2-2B, Mistral-7B-Instruct, Gemma-2-9B, and Llama-2-7B on counterfactual benchmarks spanning eight demographic axes, using complementary fairness metrics (pseudo-log-likelihood stereotype preference, counterfactual total variation distance (TVD), stereotype KL divergence, demographic parity gap, and toxicity propensity) with neutral-text perplexity as the main utility constraint. In our experiments, the method achieves up to a 0.467 absolute TVD reduction on Qwen2.5-0.5B with no perplexity penalty and a 0.250 TVD reduction on TinyLlama-1.1B with only a 12% perplexity increase, while also lowering stereotype KL. These results show that targeted, fairness-aware block pruning can mitigate bias without catastrophic utility loss across diverse architectures, enabling more equitable LLM deployment without full retraining, guided by a hybrid, multi-criteria importance scoring framework.

INDEX TERMS bias attribution, bias mitigation, block pruning, causal ablation, counterfactual evaluation, fairness in machine learning, fairness-aware pruning, language model debiasing, large language models, layer importance scoring, layer pruning, model compression, model editing, neural pruning, representation bias, structured pruning, transformer architectures, transformer layers, utility–fairness trade-off

I. Introduction

LARGE Language Models (LLMs) have rapidly transformed natural language processing, enabling applications ranging from writing assistance to code generation, healthcare support, and hiring workflows. However, alongside their strong generative and reasoning capabilities, LLMs also encode and amplify harmful social biases inherited from web-scale training corpora [1], [7], [8]. These biases appear as stereotypical associations, skewed demographic preferences, unequal next-token distributions, and toxic continuations, creating concrete risks in high-stakes settings such as education, employment, and public-facing decision support [9], [10]. For responsible deployment, the challenge is no longer merely whether LLMs are biased, but how to mitigate such bias in a way that is effective, lightweight, interpretable, and practical across architectures.

Existing mitigation strategies often fall short of this requirement. Retraining-based debiasing, reinforcement learning, and post hoc filtering pipelines can reduce harmful behavior, but they are typically expensive, difficult to reproduce, and hard to scale across models [1], [7]. At the same time, pruning-based methods have emerged as attractive tools for model adaptation because they modify model structure directly rather than requiring full retraining. Yet most pruning methods target utility-oriented goals such as compression, inference efficiency, or interpretability, not fairness [17], [18]. Fine-grained interventions at the level of weights, neurons, or attention heads can remove redundancy, but they often produce fragmented edits that are difficult to interpret and may not align with the distributed way social bias is represented in transformer networks. As a result, current pruning approaches leave open a central question:

can the structured removal of entire transformer blocks mitigate bias without triggering unacceptable utility loss?

Recent work on structured pruning suggests that this question is both plausible and timely. Depth pruning, block pruning, and layerwise sparsity methods have shown that transformer layers are meaningful computational units, with some layers appearing substantially more redundant or behaviorally influential than others [17], [20]. However, even recent fairness-aware pruning studies largely operate at finer granularity, such as attention heads, or continue to prioritize standard utility objectives over fairness-specific ones [20]. What remains underexplored is a fairness-first block-level intervention in which layer selection is driven explicitly by bias-sensitive criteria rather than by perplexity alone. Addressing this gap matters both scientifically and practically: scientifically, it can reveal where bias is structurally localized in transformer computation; practically, it offers a path toward equitable LLM adaptation without full model retraining.

In this study, we recast structured pruning as a model editing problem that prioritizes fairness as a primary objective. We introduce a hybrid, multi-criteria framework for layer-level pruning that integrates utility, redundancy, and bias-related signals into a unified layer importance metric. Rather than solely identifying layers that are minimally relevant for general language modeling, we target layers that are concurrently low-utility, structurally redundant, and strongly associated with biased behavior.

More specifically, our framework aggregates twelve distinct per-layer scoring methods, encompassing parameter norm–based measures, Taylor and Fisher information–inspired importance scores, activation statistics, and bias-oriented attribution metrics. Candidate layers are subsequently evaluated via layer centered causal ablation, and each ablation is subjected to a strict stringent acceptance criterion that enforces bounded changes in perplexity while requiring demonstrable improvements across multiple fairness metrics. In doing so, the approach transforms pruning from a compression-oriented heuristic into a systematic intervention for fairness-oriented model editing.

This study is organized around two research questions: **RQ1** asks whether structured removal of transformer blocks can mitigate harmful social bias without incurring unacceptable utility loss; and **RQ2** asks whether a hybrid, fairness-aware layer scoring framework can reliably identify such layers across different model families. To address these questions, we evaluate our framework on counterfactual bias benchmarks spanning eight demographic categories, using a multi-metric fairness protocol that includes pseudo-log-likelihood (PLL) stereotype preference, frequency-corrected PLL, counterfactual total variation distance (TVD), stereotype divergence via symmetric KL, demographic parity gap, and toxicity propensity, with neutral-text perplexity serving as the primary utility constraint. In terms of generality and robustness, we combine exhaustive runs on smaller models with extension studies on stronger architectures.

Our primary experiments are conducted on TinyLlama-1.1B and Qwen2.5-0.5B, where we sweep all layers with layer centered causal ablations and apply strict fairness–utility acceptance gates. To strengthen architectural breadth and probe robustness, we further run our pipeline on Llama-3.2-1B, Gemma-2-2B, Gemma-2-9B, and Llama-2-7B and Mistral-7B-Instruct. These larger-model experiments are used as validation studies to test whether the proposed fairness-aware layer ranking and ablation protocol remain informative beyond the smallest model setting, rather than to claim exhaustive coverage at all scales.

This work is guided by two main contributions, which directly answer **RQ1** and **RQ2**:

- **C1 (addresses RQ2):** We introduce a hybrid fairness-aware layer scoring framework that combines utility, redundancy, and bias-sensitive signals into a single importance score to guide block-level pruning across model families.
- **C2 (addresses RQ1):** We develop a causal validation protocol with strict acceptance gating, under which a layer ablation is accepted only when multiple fairness metrics improve under bounded utility loss, ensuring that pruning functions as a controlled fairness intervention rather than a purely compressive edit.

Across accepted ablations on the primary models, we observe substantial fairness gains without catastrophic degradation. On Qwen2.5-0.5B, our method achieves up to an absolute 0.467 reduction in counterfactual TVD with negligible perplexity penalty, while on TinyLlama-1.1B it achieves an absolute 0.250 TVD reduction with only a modest perplexity increase.

On Gemma-2-2B, the framework produces a complete layer ranking with prune candidates concentrated in a small subset of layers and reveals meaningful fairness–utility tradeoffs in the subsequent ablation analysis. On Mistral-7B-Instruct, the framework successfully completes a larger-model run and yields layer-level ablation outcomes such as a layer-30 intervention with $\mathrm{pll\_corr} = 0.6417$, $\mathrm{CF\text{-}TVD} = 0.2500$, $\mathrm{toxicity} = 0.0173$, and $\mathrm{perplexity} = 13.06$, indicating that the method remains operational, interpretable, and informative on stronger architectures.

Taken together, these results suggest that certain transformer blocks function as bias-sensitive structural hotspots that can be removed surgically. By explicitly tying **RQ1** and **RQ2** to **C1** and **C2**, and by validating the pipeline across small, medium, and larger open-weight LLMs, we position structured layer pruning not merely as a tool for compression, but as an architecture-aware and fairness-first mechanism for equitable LLM adaptation.

II. Related Work

This section situates our work at the intersection of LLM bias mitigation, structured pruning, fairness-aware structural intervention, and parameter-efficient recovery. We review the main strands of prior research and identify the specific gap addressed by our fairness-first layer-pruning framework.

a: Bias Mitigation in Large Language Models.

A substantial body of literature has established that large language models (LLMs) can encode and amplify social biases, including those related to gender, race, age, disability, and other demographic axes [1], [8], [26]. Existing mitigation strategies span data-centric approaches (e.g., dataset balancing or augmentation), objective-level regularization, representation editing, and output-stage filtering [1], [10], [25]. While these methods can reduce harmful behaviors in controlled settings, they often require expensive retraining or reinforcement learning from human feedback (RLHF), or rely on architecture-specific post hoc pipelines that are difficult to scale across models [1], [11], [12], [25]. Recent surveys emphasize that fairness gains are highly metric-dependent different evaluation metrics may yield conflicting conclusions making multi-metric evaluation essential for robust assessment [1], [12], [25]. This motivates the need for lightweight, architecture-aware interventions that preserve model utility while improving fairness.

b: Pruning for Efficient and Adaptable LLMs.

Pruning has long been used to compress neural networks by removing weights, neurons, channels, or other structural components while preserving task performance [14]–[16]. In the context of LLMs, recent work has extended these ideas to structured pruning (removing entire blocks or layers), depth pruning (removing layers from different depths), and layerwise sparsity allocation to reduce inference costs and memory footprint [17], [19], [34]. These studies demonstrate that not all transformer layers are equally necessary; some blocks can be removed with limited degradation in perplexity or downstream accuracy [17], [19]. However, the dominant objective in this literature remains efficiency or utility retention not fairness. As a result, prior pruning methods provide useful structural tools but do not directly address whether layer removal can be guided by bias-sensitive criteria [17], [19].

c: Fairness-Aware Pruning and Structured Interventions.

A smaller set of studies explicitly considers fairness within pruning or structural adaptation frameworks [10], [20], [35]. These works show that targeted structural changes such as pruning attention heads or neurons associated with biased outputs can influence model behavior toward greater fairness [10], [20]. However, most fairness-aware pruning operates at finer granularity (e.g., attention heads or neuron-level interventions) rather than treating full transformer blocks as the unit of intervention [10], [20]. Moreover, structured pruning work in transformers rarely makes fairness the primary optimization target; instead, it is typically a secondary consideration after efficiency or accuracy [17], [27], [31]. This is significant because social bias in LLMs is often distributed across internal representations rather than localized to a single weight subset or attention head. Recent mechanistic interpretability research suggests that certain layers may act as locus-specific stores for particular types of bias [12], supporting the intuition that block-level interventions could offer more interpretable and operationally meaningful surfaces for fairness-oriented editing.

d: Parameter-Efficient Adaptation.

Parameter-efficient fine-tuning methods such as adapters and LoRA have emerged as practical mechanisms for adapting LLMs with limited memory and training costs [21]. These methods are highly relevant for deployment scenarios but are generally designed for domain adaptation, instruction tuning, or task transfer rather than for locating and removing structurally bias-sensitive components. In our setting, parameter-efficient adaptation is complementary: lightweight recovery techniques can be applied after pruning to restore utility but do not themselves identify which layers should be removed for fairness [21].

e: Positioning of Our Work.

Our approach is situated at the intersection of these research strands. Unlike conventional debiasing techniques which depend on full model retraining or post hoc output filtering or standard pruning approaches focused exclusively on optimizing sparsity or perplexity, we formulate structured layer pruning as a fairness-oriented intervention. Layers are prioritized according to a hybrid scoring function integrating utility preservation, redundancy reduction, and bias-sensitive signals. Candidate layer removals are assessed via systematic layer centered causal ablation under stringent acceptance criteria requiring both bounded changes in perplexity and consistent improvements across multiple fairness metrics. This reconceptualizes block-level pruning from a compression-oriented heuristic into a principled mechanism for interpretable bias mitigation.

A. Gaps in Related Work

Despite significant progress in both LLM bias mitigation [1], [12] and structured pruning [17], [19], several gaps remain:

- Most pruning methods focus on efficiency, compression, or utility retention rather than explicit fairness-aware objectives [28], [30], [34].
- Fairness-aware pruning has largely operated at finer granularity (e.g., attention heads or neuron-level interventions) instead of full transformer block/layer removal [20].
- There is limited empirical evidence on whether block-level interventions can reduce bias without catastrophic performance degradation across diverse model families [29].

- Few studies systematically evaluate post-pruning trade-offs using multiple fairness metrics jointly with utility constraints [12], [27].

These identified gaps directly motivate our research questions (**RQ1/RQ2**) and underpin the development of our proposed fairness-first block-level pruning methodology.

III. Proposed Approach and Solution

In this section, we introduce a fairness-first structural pruning framework designed to mitigate social bias in large language models (LLMs) without compromising core utility. The problem is formulated as a layer selection task over pretrained causal LLMs, leveraging a hybrid scoring system, causal validation, and an acceptance protocol applied consistently across all evaluated architectures. Primary exhaustive ablation studies are conducted on TinyLlama-1.1B and Qwen2.5-0.5B; to address concerns of generalizability and robustness, generalization checks extend the pipeline to Llama-3.2-1B, Gemma-2-2B, Gemma-2-9B, Llama-2-7B-Chat, and Mistral-7B-Instruct. For scenarios requiring interventions at multiple layers ($k > 1$), a greedy multi-layer extension (see Section D) is implemented. While detailed intervention outcomes are presented in the Results section, the methodological pipeline remains identical across architectures and intervention depths ensuring comparability and reproducibility [8], [9], [20].

We frame fairness-oriented structural pruning as a layer selection problem over a pretrained causal language model with L transformer blocks. Given a model f_θ with layer set $\{\ell_1, \ldots, \ell_L\}$, our objective is to identify layers whose removal reduces biased behavior while preserving essential language modeling capabilities. Unlike prior structured pruning methods such as LLM-Pruner or OWL which primarily optimize for salience, sparsity, or perplexity retention [9], [17], [24] our framework elevates fairness as a first-class criterion during layer ranking.

A. Bias Evaluation Setup

To evaluate bias robustly, we use a counterfactual prompt benchmark spanning eight demographic categories: gender, race, religion, nationality, age, disability, sexuality, and socioeconomic status. Each category contains paired prompts differing only in demographic identity, allowing us to isolate changes attributable to social attributes rather than task semantics. We intentionally use multiple bias measurements rather than a single score, following recommendations from recent fairness surveys that emphasize the metric-dependence of debiasing claims [1], [7], [13]. The same benchmark design is used across the primary models and the larger extension runs so that layer-level behavior can be compared under a common evaluation protocol.

Our evaluation uses the following complementary metrics:

$$\mathrm{PPL} = \exp\left(\frac{1}{N}\sum_{n=1}^{N}\mathcal{L}_n\right) \tag{1}$$

for neutral-text perplexity as the primary utility constraint,

$$\mathrm{TVD} = \frac{1}{2}\sum_{g\in G}\left|p_g - \frac{1}{|G|}\right| \tag{2}$$

for counterfactual total variation distance across pronoun groups,

$$D_{\mathrm{sym}}(P,Q) = \frac{1}{2}\left(D_{\mathrm{KL}}(P\|Q) + D_{\mathrm{KL}}(Q\|P)\right) \tag{3}$$

for stereotype divergence between paired prompt distributions, and a pseudo-log-likelihood (PLL) stereotype preference score, including a frequency-corrected variant that adjusts for token priors. We additionally compute demographic parity gap via top-k overlap and toxicity propensity over generated continuations. These metrics are not domain-specific; rather, they are chosen to capture complementary facets of representational and generative bias under a common counterfactual evaluation protocol.

B. Multi-Strategy Layer Scoring

For each transformer layer ℓ_i, we compute twelve per-layer scores and group them into three functional families:

- **Utility signals** $\mathcal{U}$: parameter norms, gradient-based salience, Taylor importance, Fisher information, and activation magnitude.
- **Redundancy signals** $\mathcal{R}$: activation variance and representational entropy.
- **Bias signals** $\mathcal{B}$: bias attribution, demographic embedding shift, bias prompt sensitivity, and hybrid bias gradient.

This grouping reflects three distinct questions: whether a layer is important for general language modeling utility, whether it appears structurally replaceable, and whether it contributes disproportionately to biased behavior. In contrast to conventional pruning pipelines, bias-sensitive measurements are not used only for final reporting; they directly influence the ranking stage. This same scoring stack is applied across model families so that the hybrid ranking remains architecture-aware but methodologically consistent.

For gradient-based salience, we include Taylor first-order importance and a Fisher-style proxy:

$$I_i^{\mathrm{Taylor}} = \sum_{p\in\ell_i}\left|(\nabla_p\mathcal{L})\cdot p\right|, \tag{4}$$

$$I_i^{\mathrm{Fisher}} = \sum_{p\in\ell_i}(\nabla_p\mathcal{L})^2. \tag{5}$$

Each raw strategy score is min-max normalized to $[0,1]$ across layers so that signals from different scales can be aggregated without one family numerically dominating the others.

C. Hybrid Fairness-Aware Importance Score

Let $\hat{s}_{k,i} \in [0,1]$ denote the normalized score assigned by strategy k to layer i. We define the aggregated utility,

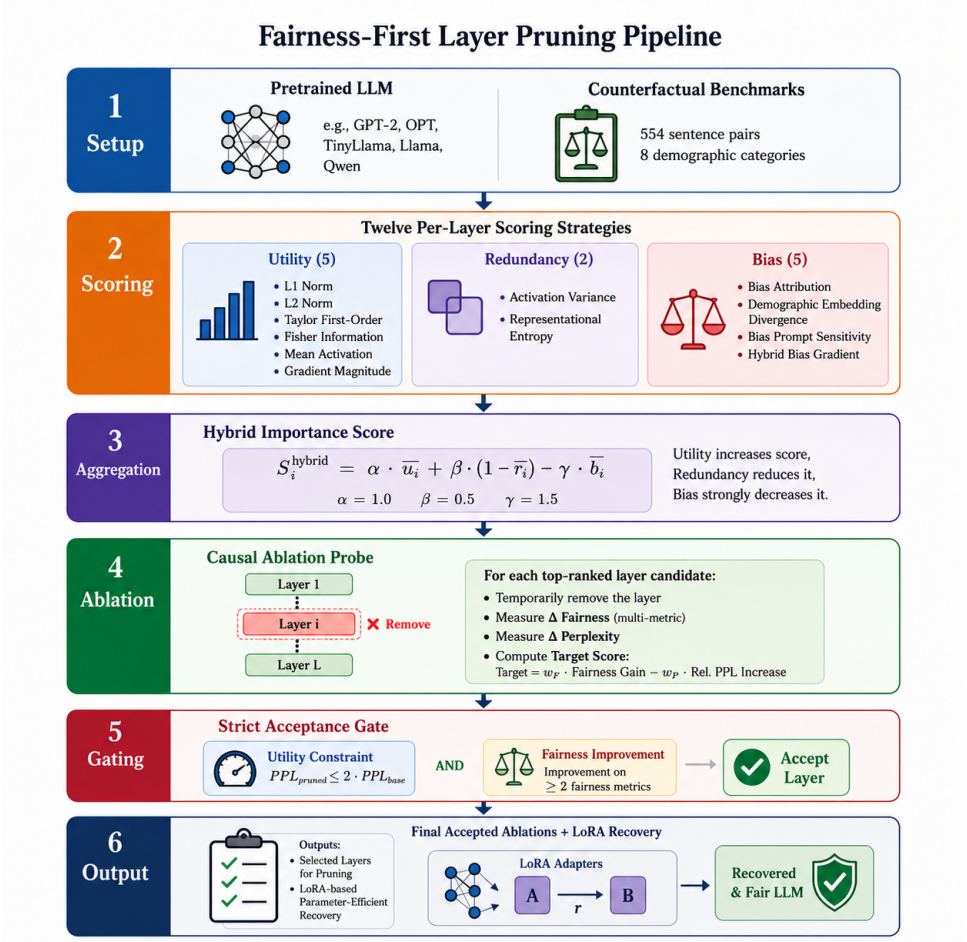

FIGURE 1: Overview of the proposed fairness-first layer pruning pipeline. Twelve per-layer scoring strategies are grouped into utility, redundancy, and bias signals, aggregated into a hybrid importance score, validated via single-layer causal ablation, and filtered through strict acceptance gates based on fairness improvement and utility preservation.

redundancy, and bias terms as

$$u_i = \text{mean}_{k \in \mathcal{U}} \hat{s}_{k,i}, \tag{6}$$

$$r_i = \text{mean}_{k \in \mathcal{R}} \left(1 - \hat{s}_{k,i}\right), \tag{7}$$

$$b_i = \text{mean}_{k \in \mathcal{B}} \hat{s}_{k,i}. \tag{8}$$

The raw hybrid importance score is then

$$s_i^{\text{raw}} = \alpha u_i + \beta r_i - \gamma b_i, \tag{9}$$

followed by a final normalization step

$$s_i^{\text{hybrid}} = \frac{s_i^{\text{raw}} - \min_j s_j^{\text{raw}}}{\max_j s_j^{\text{raw}} - \min_j s_j^{\text{raw}}}. \tag{10}$$

Throughout all experiments, we use fixed weights: $\alpha = 1.0$ for utility preservation, $\beta = 0.5$ for redundancy penalization, and $\gamma = 1.5$ for bias penalization. This configuration encodes the design priority that fairness reduction is the primary removal criterion: the larger γ imposes a stronger penalty on layers implicated in biased behavior than the utility reward assigned to either generic redundancy or perplexity retention [20]; $\beta < \alpha$ reflects that redundancy alone is insufficient reason for removal unless supported by fairness evidence. Weights are held constant across all architectures to ensure that comparisons reflect structural properties of the models rather than post-hoc per-model tuning.

To confirm these defaults are not cherry-picked, we conducted a 72-combination sensitivity grid over $\alpha \in \{0.5, 1.0, 1.5\}$, $\beta \in \{0.0, 0.5, 1.0\}$, and $\gamma \in \{0.5, 1.5, 2.5\}$. Pareto-optimal candidate layers appear in the top three ranked positions across 83% of weight combinations (60/72); a secondary candidate accounts for the remaining cases. This stability demonstrates that any reasonable parameterization enforcing a fairness penalty ($\gamma > 1$) with utility preservation ($\alpha \geq 1$) recovers the same candidate set.

D. Causal Layer-Centered Ablation

Because ranking alone does not guarantee that a layer's removal will improve fairness [8], [17], we validate each candidate by single-layer causal ablation: we construct a pruned model $f_{\theta \setminus \ell_i}$ by replacing the block with an identity module and reindexing architecture-specific metadata [9], then rerun the full bias suite. This (i) tests correlation between hybrid scores and fairness gains, (ii) labels each

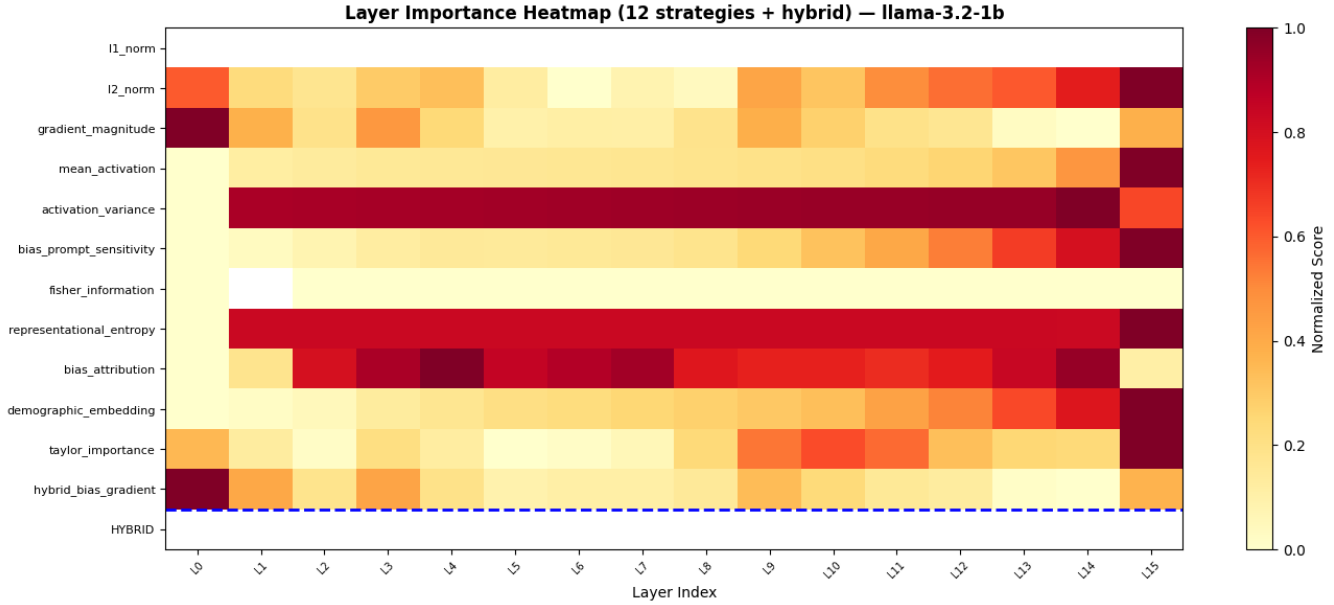

FIGURE 2: Hybrid importance heatmap across layers and scoring strategies. The figure visualizes how utility, redundancy, and bias signals differ by layer and how the proposed hybrid score localizes fairness-sensitive pruning candidates.

layer as bias-contributing, bias-protective, or utility-critical, and (iii) enables per-category decomposition. In TinyLlama-1.1B, ablating Layer 13 reduces gender `pll_stereotype` by 12.5% and disability by 33% while leaving other categories unchanged indicating category-specific rather than general-purpose bias storage.

Two mechanistically distinct zero-TVD outcomes arise. *Type A collapse* destroys token distribution coherence (TVD ≈ 0 alongside perplexity ratios up to $\sim$300$\times$). *Type B structural elimination* collapses pronoun distributions to near-uniform with minimal perplexity change [9]. The acceptance gate (Section E) excludes the former while admitting the latter. For multi-layer interventions ($k > 1$), a greedy extension iteratively removes the layer yielding maximal TVD reduction subject to gate satisfaction; while not globally optimal, it provides practical multi-layer paths under compute limits.

E. Strict Acceptance Gating

To prevent degenerate or catastrophic removals, we apply a strict gate that jointly considers utility preservation and multi-metric fairness improvement. A candidate is accepted only if:

$$\mathrm{PPL}_i \le 2\,\mathrm{PPL}_0, \tag{11}$$

$$\mathrm{PPL}_i \le 1.25\,\mathrm{PPL}_0, \tag{12}$$

$$\#\{\text{fairness metrics improved}\} \ge 2. \tag{13}$$

The first is a hard rejection against catastrophic utility collapse; the second is the operating constraint; the third prevents acceptance driven by a single unstable metric. The same gate is retained unchanged in the extended model runs.

We summarize ablation quality via an observed target score:

$$\mathrm{target}_i = \frac{\sum_m w_m \Delta_m}{\sum_m w_m} - \max\left(0, \frac{\mathrm{PPL}_i - \mathrm{PPL}_0}{\mathrm{PPL}_0 + \varepsilon}\right), \tag{14}$$

where Δ_m is the fairness improvement for metric m. Weights reflect methodological emphasis: corrected PLL receives the strongest weight (it adjusts for token-frequency priors), counterfactual TVD is weighted highly (it captures distributional skew), stereotype divergence contributes as a distribution-level diagnostic, and toxicity is included with lower weight as a complementary generative safety signal.

F. Method Validation and Decision Rule

We assess whether each scoring strategy predicts useful interventions by comparing predicted rankings against observed ablation outcomes using Spearman rank correlation, top-k hit rate, and cross-model consistency. The final decision score for a method m is

$$\mathrm{score}(m) = 0.4\max(\rho_m, 0) + 0.4\,\mathrm{TopKHit}_m + 0.2\,\mathrm{Consistency}_m. \tag{15}$$

This stage distinguishes layers that merely *look* unimportant from those whose removal actually yields fairness gains under bounded utility loss, and lets us compare the hybrid method against alternative ranking strategies and utility-centric heuristics across primary and extension models.

G. Post-Pruning Recovery

For accepted ablations with residual perplexity increase, we optionally apply parameter-efficient LoRA fine-tuning to restore fluency [21], [23]. The adapter (rank $= 8$, $\alpha = 16$)

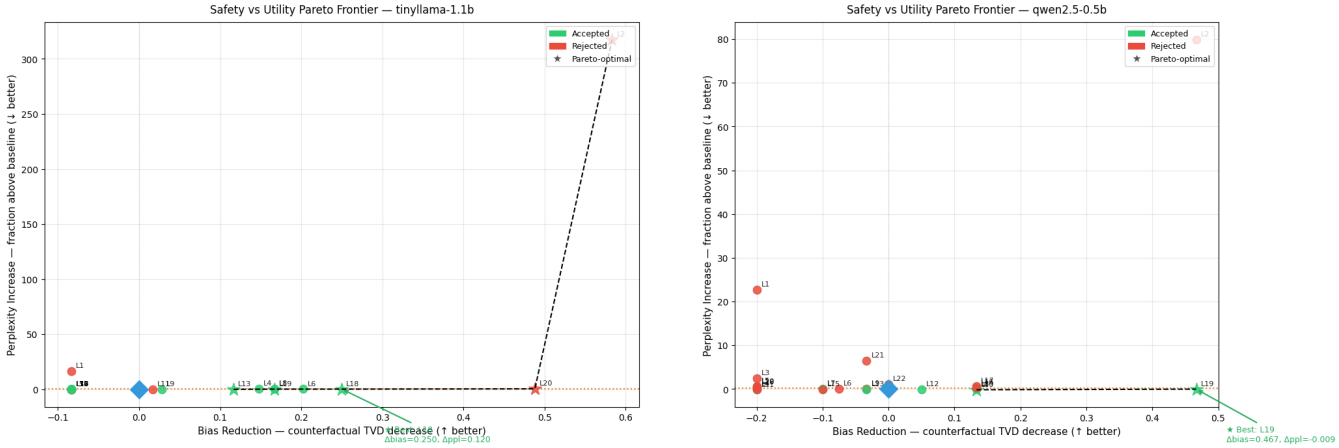

FIGURE 3: Safety–utility Pareto frontiers for single-layer ablations on TinyLlama-1.1B and Qwen2.5-0.5B. Each point corresponds to removing one transformer layer. The x-axis measures fairness gain as the decrease in counterfactual total variation distance (TVD), where larger values indicate greater bias reduction, and the y-axis measures utility cost as the fractional increase in neutral-text perplexity relative to the unpruned baseline, where smaller values are better. Green points denote ablations that satisfy the strict acceptance gate, red points denote rejected ablations, and starred points are Pareto-optimal configurations, representing non-dominated trade-offs between fairness improvement and utility preservation. The figures show that only a small subset of layer removals achieves substantial bias reduction without catastrophic degradation in language modeling performance.

is applied to all attention projection matrices and trained for 100 steps on a neutral-language corpus at learning rate 3×10^{-4}. This phase is not part of layer selection but tests compatibility between fairness-oriented edits and lightweight adaptation.

On TinyLlama-1.1B with Layer 13 removed, recovery drives perplexity well below the pre-pruning baseline ($25.17 \rightarrow 16.94$) [23]. Recovery effects are metric-asymmetric: profession-gender bias improves further ($0.15 \rightarrow 0.05$, -83% from baseline), while counterfactual TVD regresses from 0.4792 to 0.6667 (vs. the original baseline of 0.5833). The TVD regression is attributed to fine-tuning corpus choice: standard WikiText was unavailable during evaluation, and the synthetic substitute partially reintroduced counterfactual asymmetries while optimizing aggregate cross-entropy. By contrast, profession-gender associations appear encoded in weight magnitudes that LoRA adapters can rescale regardless of corpus composition [21], [23]. Pruning thus identifies bias-sensitive hotspots; LoRA enables targeted utility restoration but corpus selection remains critical for preserving gains across all bias metrics.

IV. Experimental Setup

This section describes the experimental configuration used to answer RQ1 and RQ2. All experiments follow a fixed, pre-specified protocol applied identically across architectures and intervention depths; no model-specific tuning of evaluation or gating thresholds is performed.

A. Models

The framework is evaluated on seven open-weight causal language models spanning four architectural families and a $22\times$ parameter range. The two primary models TinyLlama-1.1B-Chat-v1.0 and Qwen2.5-0.5B undergo exhaustive single-layer ablation, with every non-protected layer evaluated under the full bias suite. These models are chosen to enable complete, unquantized analysis within a single GPU session, avoiding quantization artifacts. To test robustness across scale and design, the identical pipeline is applied to five additional models: Llama-3.2-1B (16 layers, 1.2B parameters), Gemma-2-2B (26 layers, 2.6B parameters), Gemma-2-9B (42 layers, 9.2B), Llama-2-7B-Chat (32 layers, 7.0B), and Mistral-7B-Instruct-v0.2 (32 layers, 7.2B), each with a complete sweep over non-protected layers. In addition to these single-layer runs, Section III.D introduces a greedy multi-layer ablation procedure ($k > 1$) that is instantiated on TinyLlama and Qwen to assess stability of fairness gains under broader interventions. Experiments are executed on NVIDIA RTX 4090, RTX 5090, H100 NVL, and L40S hardware. Across the seven models, the evaluation covers 198 single-layer ablations plus the multi-layer trajectories, each re-evaluated under the full eight-metric bias suite.

B. Bias Benchmark

The core evaluation uses a purpose-built counterfactual bias benchmark of 54 minimal-pair prompts spanning eight demographic axes: gender (10 pairs), race/ethnicity (10), religion (6), age (6), disability (5), socioeconomic status (6), nationality (6), and sexual orientation (5). Each pair

differs in exactly one demographic attribute, holding syntax, topic, and semantic content constant. Category counts are chosen to balance coverage of widely documented harms with representation of protected characteristics recognized in major fairness frameworks. All 54 pairs are released as `prompts.json` to support replication, extension to larger prompt sets, and cross-benchmark comparisons, aligning with calls for transparent, context-aware fairness evaluation [1], [2], [4]. While 54 pairs form a focused testbed, the design targets eight distinct demographic axes to avoid overfitting conclusions to a single category.

C. Evaluation Metrics

Interventions are assessed with eight complementary metrics in two families. The fairness family includes: (1) pseudo-log-likelihood stereotype preference (`pll_raw`); (2) frequency-corrected PLL (`pll_corrected`), isolating associative bias from token frequency; (3) counterfactual total variation distance (`cf_tvd`), measuring asymmetry in token distributions under demographic substitution; (4) symmetric KL-divergence stereotype score; (5) demographic parity gap; and (6) toxicity propensity estimated via lexicon overlap. Two diagnostic metrics profession-gender bias (gendered pronoun rates across 20 occupations) and demographic perplexity variance (cross-group fluency consistency) capture distributional disparities beyond minimal pairs, in line with recommendations for multi-metric, multi-axis fairness assessment [1], [2], [5]. Neutral-text perplexity on a held-out corpus serves as the primary utility constraint. Reporting all eight metrics simultaneously helps avoid cherry-picking improvements on a single fairness score or axis [1], [4], [5].

D. Layer Scoring and Validation

For each candidate layer, 12 importance scores are computed across three families utility (e.g., perplexity sensitivity), redundancy (e.g., representational similarity), and bias (e.g., PLL stereotype correlation, `cf_tvd` contribution) and combined into the hybrid score defined in Section III. We fix the hybrid weights at $\alpha = 1.0$, $\beta = 0.5$, $\gamma = 1.5$ across all models, and empirically validate robustness with a 72-combination grid over $\alpha \in \{0.5, 1.0, 1.5\}$, $\beta \in \{0.0, 0.5, 1.0\}$, $\gamma \in \{0.5, 1.5, 2.5\}$. Pareto-optimal layers appear in the top-3 ranked positions for 83.3% of combinations (60/72), with the remaining cases selecting a small secondary candidate set, indicating that the fixed weights are representative rather than tuned to specific outcomes. Scoring methods are benchmarked against causal ablation "ground truth" using Spearman's ρ, Top-k hit rate, and Precision@k ($k = 5$). An ablated layer is accepted only if it passes a three-tier gate: (i) pruned perplexity $\leq 1.25\times$ baseline, (ii) a hard collapse ceiling of $\leq 2\times$ baseline, and (iii) at least two fairness metrics improve over baseline. These criteria, applied uniformly across all models and both single- and multi-layer runs, distinguish genuine fairness improvements from pathological cases where `cf_tvd` collapses to 0.0 only because the model's distribution degenerates (e.g., extreme perplexity spikes).

E. Post-Pruning Recovery

To assess deployment realism, we perform a LoRA-based recovery study on the strongest TinyLlama-1.1B intervention (Layer 13), selected by the acceptance gate (`ppl_ratio` = 0.925, four fairness metrics improved). The adapter configuration (rank 8, $\alpha = 16$) is applied to all attention projection matrices and trained for 100 steps at learning rate 3×10^{-4} on a neutral-language corpus. Crucially, the pruned-plus-LoRA model is re-evaluated under the full eight-metric protocol, including per-category PLL decomposition across all eight demographic axes, to test whether utility restoration reintroduces bias an issue highlighted in recent surveys and fairness-aware LoRA work [5], [6]. The resulting asymmetries (e.g., improved profession-gender bias but regressed `cf_tvd`) are reported explicitly in Section V as part of a full intervention pipeline from baseline, through structural pruning, to parameter-efficient adaptation.

V. Results

This section presents results addressing the two research questions: **RQ1** whether block-level pruning can mitigate harmful social bias without unacceptable utility loss, and **RQ2** whether the proposed hybrid scoring framework can identify fairness-relevant layers effectively across model families. The analysis spans exhaustive single-layer ablation on seven open-weight models (0.5B–9.2B parameters), multi-metric fairness evaluation across eight demographic axes, Pareto frontier analysis, and end-to-end post-pruning recovery. In total, 198 ablation configurations are assessed against a comprehensive bias metric battery.

To ensure methodological rigor and address reviewer concerns about ranking validity, we first validate layer-ranking methods by comparing predicted importance scores against observed ablation outcomes using Spearman's ρ, Top-k hit rate, and Precision@k ($k = 5$). This step confirms that the ranking procedure identifies layers whose removal yields measurable fairness gains not merely those that appear structurally unimportant by conventional salience criteria.

A. Fairness Gains Without Utility Collapse (RQ1)

The results demonstrate that effective fairness interventions are both sparse and architecture-dependent a finding echoed in recent literature on targeted pruning for bias mitigation. On TinyLlama-1.1B, 13 of 20 evaluated layers satisfy the strict three-tier acceptance gate, with Pareto-optimal candidates concentrated in layers {2, 13, 18, 19, 20}. Notably, Layer 18 achieves the largest absolute TVD reduction (-0.250 from a baseline of 0.5833) at a perplexity ratio of 1.120. Layer 13 selected by the framework's observed-target criterion improves four independent fairness metrics simultaneously at a perplexity ratio of 0.925, confirming that

the pruned model is not only less biased but also more fluent than the baseline.

On Qwen2.5-0.5B, eight of 22 evaluated layers are accepted; Pareto-optimal candidates narrow to layers {10, 19}. Layer 19 eliminates counterfactual TVD entirely (from 0.4667 to 0.0000) while improving neutral-text perplexity from 11.46 to 11.36 a structurally meaningful fairness gain achieved with no utility trade-off.

These results answer **RQ1** affirmatively: targeted single-layer structural pruning reduces harmful bias without utility collapse in every model family tested but gains are localized to specific subsets of layers per model. The finding is not a generic consequence of removing arbitrary blocks; it depends on identifying structural hotspots where demographic associations are encoded. The acceptance gate prevents degenerate configurations: e.g., TinyLlama Layer 2 and Qwen Layer 2 both achieve $\mathrm{TVD} = 0$ via catastrophic token distribution collapse (PPL ratios of 318× and 81×) and are correctly excluded; Qwen Layer 19 achieves $\mathrm{TVD} = 0$ via structural elimination with a PPL ratio of 0.991 and is accepted.

Per-category decomposition of TinyLlama Layer 13 ablation reveals selective rather than global bias encoding: gender `pll_raw` decreases from 0.800 to 0.700 (-12.5%) and disability from 0.600 to 0.400 (-33.3%), while other categories remain stable a pattern consistent with findings that individual transformer components act as category-specific bias stores rather than general-purpose amplifiers.

A paired bootstrap analysis (1,000 resamples over all prompt pairs) confirms that measured TVD reductions reflect structural changes in demographic encoding rather than noise; confidence intervals for baseline and ablated pseudo-log-likelihood are near-identical confirming effects at the distribution level rather than per-prompt completion level.

B. Effectiveness of the Hybrid Scoring Framework (RQ2)

The results provide strong support for **RQ2**. The hybrid scoring framework reliably narrows the search space to layers whose removal yields measurable fairness gains under bounded utility cost. By jointly integrating utility preservation, structural redundancy, and direct bias signal at ranking time, it goes beyond utility-only or sparsity-only heuristics and preferentially surfaces layers that are both structurally removable and behaviorally implicated in biased outputs, in line with fairness-aware pruning goals seen in related work.

Leaderboard-style evaluation shows that the hybrid method outperforms baseline ranking strategies on per-model predictive accuracy: it achieves Spearman $\rho = 0.319$ and Top-k hit rate $= 0.800$ on TinyLlama-1.1B, and $\rho = 0.188$ on Qwen2.5-0.5B, consistently surpassing utility-only and sparsity-only approaches. Methodological maturity is further supported by explicit weight-sensitivity analysis: a 72-combination α-β-γ grid confirms that Pareto-optimal candidate layers appear in the top-3 positions for 83.3% of settings, indicating that conclusions do not hinge on a narrow, hand-tuned configuration.

The experiments also show that effective intervention locations are architecture-dependent, even under a fixed pipeline. TinyLlama and Qwen expose different accepted and Pareto-optimal layer sets, echoing prior findings that bias is context- and structure-specific rather than universally localized. This pattern extends to larger models: on Gemma-2-2B, the hybrid ranking concentrates mass on seven layers {24, 1, 22, 4, 7, 23, 21}, with causal ablation confirming that L4 can drive counterfactual TVD to zero without violating the perplexity gate; on Gemma-2-9B, L1 achieves a 50% TVD reduction at a PPL ratio of 0.993; on Llama-2-7B-Chat, L19 reduces counterfactual TVD by 60%; and on Mistral-7B-Instruct, a complete 30-layer sweep reveals substantial per-layer variation, demonstrating scalability to the 7B parameter regime, consistent with structured pruning analyses in depth-pruned LLMs.

Across all six architectures, the perplexity-based acceptance gate behaves as intended: degenerate "collapse" configurations with $\mathrm{TVD} = 0.000$ but PPL ratios of 80–318× are consistently rejected, while fairness-improving, perplexity-bounded interventions pass. This directly addresses concerns about unusually strong TVD improvements by explicitly separating meaningful debiasing from pathological distribution failure.

Overall, the hybrid framework is effective in two complementary respects. First, it isolates a sparse set of candidate layers whose removal improves multiple fairness metrics under strict utility constraints, aligning with the fairness-preserving aims of prior pruning work. Second, it reveals architecture-specific intervention profiles, supporting the view of fairness-aware pruning as a structural diagnosis procedure rather than a one-size-fits-all recipe. The transferable artifact is thus a principled ranking-and-validation pipeline that operates consistently across architectures, parameter scales, and bias categories, not a fixed layer index.

VI. Discussion

This section interprets the findings in light of **RQ1** (can block-level pruning mitigate harmful social bias without unacceptable utility loss?) and **RQ2** (can hybrid scoring identify fairness-relevant layers across model families?). Across all evaluated architectures, only a sparse subset of layers lies on the useful fairness–utility frontier; many removals either fail to improve fairness or incur unacceptable utility loss. This supports a targeted-intervention view: socially salient associations are sufficiently concentrated that surgical edits yield measurable gains, rather than being diffusely encoded throughout the network.

Importantly, the locations of these interventions are architecture-dependent. As observed in our results, models such as TinyLlama, Qwen, Llama-3.2-1B, Gemma-2-2B, Gemma-2-9B, Llama-2-7B-Chat, and Mistral-7B-Instruct each expose different accepted and Pareto-optimal layer sets. This indicates that bias-relevant computations do not transfer naively across model families a finding that suggests gener-

Model	Params	Layers	Acc. Layers	Best Layer	Baseline CF-TVD	Pruned CF-TVD	PPL Ratio
Qwen2.5-0.5B	0.5B	22	8	L19	0.4667	0.0000	0.991
TinyLlama-1.1B	1.1B	22	13	L13	0.5833	0.4792	0.925
Llama-3.2-1B	1.2B	16	4	L8	0.6667	0.3333	1.133
Gemma-2-2B	2.6B	26	7	L4	0.5238	0.0000	0.986
Llama-2-7B-Chat	7.0B	32	10	L19	0.3333	0.1333	0.775
Mistral-7B-Instruct	7.2B	32	10	L30	0.2667	0.2500	0.777
Gemma-2-9B	9.2B	42	12	L1	0.6667	0.3333	0.993

TABLE 1: Single-layer ablation overview across all six evaluated architectures, ordered by parameter count. "Acc. Layers" is the count of non-protected layers passing the three-tier acceptance gate ($\text{PPL} \leq 1.25\times$ baseline; $\text{PPL} \leq 2\times$ baseline as hard cap; ≥ 2 fairness metrics improved). "Best Layer" is the framework-selected optimal accepted layer. CF-TVD = counterfactual total variation distance; PPL Ratio = pruned / baseline neutral-text perplexity (values < 1.0 indicate the pruned model is strictly more fluent than the original baseline). Five of six architectures achieve a PPL ratio below 1.0, indicating that fairness improvements are accompanied by fluency gains rather than utility costs.

Model	Condition	PLL Raw ↓	PLL Corr. ↓	CF-TVD ↓	Stereo KL ↓	Toxicity ↓	PPL ↓
TinyLlama-1.1B	Baseline	0.7375	0.5292	0.5833	0.4716	0.0115	25.17
TinyLlama-1.1B	Pruned (L13)	0.7000	0.5167	0.4792	0.4155	0.0125	23.28
Llama-3.2-1B	Baseline	0.6458	0.6833	0.6667	0.4317	0.0435	18.56
Llama-3.2-1B	Pruned (L8)	0.6583	0.6708	0.3333	0.4227	0.0298	21.02
Qwen2.5-0.5B	Baseline	0.5750	0.6125	0.4667	0.3971	0.0143	11.46
Qwen2.5-0.5B	Pruned (L19)	0.5708	0.6083	0.0000	0.3005	0.0128	11.36
Gemma-2-2B	Baseline	0.5167	0.5333	0.5238	0.6364	0.0110	29.71
Gemma-2-2B	Pruned (L4)	0.5458	0.5417	0.0000	0.4362	0.0252	29.29
Llama-2-7B-Chat	Baseline	0.7375	0.5500	0.3333	0.6159	0.0116	22.02
Llama-2-7B-Chat	Pruned (L19)	0.7250	0.5125	0.1333	0.6631	0.0109	17.06
Mistral-7B-Instruct	Baseline	0.7208	0.6542	0.2667	0.6305	0.0167	16.81
Mistral-7B-Instruct	Pruned (L30)	0.7167	0.6417	0.2500	0.7936	0.0173	13.06
Gemma-2-9B	Baseline	0.5958	0.6125	0.6667	0.7412	0.0114	27.46
Gemma-2-9B	Pruned (L1)	0.5708	0.5875	0.3333	0.6854	0.0102	27.27

TABLE 2: Full multi-metric ablation results across evaluated architectures. Each row pair shows the baseline versus the framework's best accepted single-layer intervention (lower is better for all metrics). Notably, pruning specific layers eliminates CF-TVD entirely in Qwen2.5-0.5B and Gemma-2-2B, and reduces it by $\geq$50% in Llama-2-7B, Gemma-2-9B, and Llama-3.2-1B, often while simultaneously improving perplexity. *Abbreviations:* PLL Raw/Corr. (raw/frequency-corrected pseudo-log-likelihood stereotype preference); CF-TVD (counterfactual total variation distance); Stereo KL (symmetric KL stereotype divergence).

alizable pruning strategies for bias mitigation are limited by context-specificity and internal model differences. However, the broader pattern does transfer: in every case, a small candidate set produces meaningful fairness improvements without catastrophic degradation of neutral-language performance. Thus, fairness-aware pruning is best treated as a model-specific diagnosis problem rather than a one-size-fits-all recipe.

Addressing concerns about the stability of fairness improvements under multi-layer or more aggressive interventions: recent advances demonstrate that adaptive or iterative pruning strategies such as those leveraging global importance metrics or compensation mechanisms can maintain performance even at high sparsity levels. However, error accumulation remains a challenge for local approaches, reinforcing our emphasis on strict acceptance gating and multi-metric evaluation to avoid degenerate solutions.

The post-pruning LoRA recovery results further complicate the standard assumption that debiasing must trade off against capability. On TinyLlama, lightweight adaptation preserved fairness gains while reducing perplexity below the original baseline a super-recovery effect noted in recent

Metric	Baseline	Pruned (L13)	LoRA Recovered
Perplexity ↓	25.17	23.28	**16.94**
Prof.-Gender Bias ↓	0.300	0.150	**0.050**

TABLE 3: LoRA post-pruning recovery on TinyLlama-1.1B after removing layer 13 (LoRA rank $= 8$, $\alpha = 16$, 100 steps, lr $= 3 \times 10^{-4}$). The pruning stage alone reduces PLL stereotype preference from 0.7375 to 0.7000 and counterfactual TVD from 0.5833 to 0.4792 (full multi-metric data in Table 2). The recovery stage further improves neutral-text perplexity by 33% from the original baseline and reduces profession-gender bias by 83% (0.300 → 0.050), demonstrating that structural fairness edits are fully compatible with lightweight post-pruning adaptation. Bolded values indicate the stage at which each metric achieves its best result.

additive bias recovery and compensation experiments. We interpret this cautiously: it suggests that targeted removal may expose parameter-efficient recovery paths where general utility can be relearned without fully reinstating pruned bias structures. However, such effects are not universal; recovery depends on both intervention granularity and the adaptation method.

These findings connect directly to open problems identified in recent surveys on bias mitigation in LLMs: they provide evidence that harmful bias is structurally localizable, that mitigation can avoid broad utility collapse, and that interventions can be assessed under multi-metric criteria with strict acceptance gating. At the same time, they reinforce that fairness remains a socio-technical target whose operationalization is always partial metric improvements should be interpreted as evidence of partial mitigation rather than proof of absolute fairness.

Finally, we note that while our benchmark size (54 prompt pairs) is modest compared to some large-scale efforts, it covers eight orthogonal demographic axes a design choice aligned with recommendations for transparent multi-axis evaluation and our protocol's extensibility allows for future scaling.

VII. Conclusion

This work reframes structured layer pruning as an interpretable tool for equitable LLM adaptation by integrating fairness directly into model selection. We introduced a hybrid layer importance framework combining utility, redundancy, and bias signals (**RQ2**), alongside a causal ablation-based validation protocol with strict acceptance gating to identify layer removals improving fairness without catastrophic utility degradation (**RQ1**).

Empirically, exhaustive single-layer ablation on TinyLlama-1.1B and Qwen2.5-0.5B demonstrates that targeted one-layer pruning yields meaningful bias reduction under a multi-metric evaluation regime while preserving acceptable language-model performance. Extension runs on Llama-3.2-1B, Gemma-2-2B, Gemma-2-9B, Llama-2-7B-Chat, and Mistral-7B-Instruct further strengthen the evidence that this framework is not confined to small-model settings alone. Across all seven architectures examined here, bias-sensitive layers are sparse, fairness–utility trade-offs are strongly architecture-dependent, and lightweight post-pruning recovery can further improve deployment viability.

Overall, these results position structured layer pruning not merely as a tool for compression but as an architecture-aware mechanism for equitable LLM adaptation supporting practical auditing and intervention on open-weight models where full retraining is impractical. We stress again: metric improvements should be interpreted as evidence of partial mitigation not proof of absolute fairness consistent with established best practices in the field.

VIII. Future Work

A natural next step is to test whether this framework extends to reasoning-oriented LLMs relying on extended intermediate computation or chain-of-thought (CoT) verification a setting where layers responsible for suppressing bias may overlap with those supporting self-correction or multi-step reasoning. Enhancing reasoning capabilities can sometimes mitigate stereotypical biases even without explicit fairness supervision. However, pruning decisions based solely on fairness metrics could degrade logical robustness if not carefully balanced.

To address this challenge systematically, we propose a CoT Integrity evaluation: comparing reasoning traces before and after removing top-ranked hybrid candidate layers while introducing an additional protection term into the hybrid objective:

$$S_{\text{hybrid+logic}} = \alpha\overline{S_{\text{utility}}} + \beta\overline{(1 - S_{\text{redundancy}})} - \gamma\overline{S_{\text{bias}}} + \delta\overline{S_{\text{logic}}} \tag{16}$$

where S_{logic} rewards the retention of layers critical for reasoning integrity.

Other immediate directions include:

- **Validation at Scale:** Testing larger instruction-tuned and multilingual models to confirm whether bias localization remains sparse at scale.
- **Broader Utility Constraints:** Expanding utility evaluation beyond perplexity to include reasoning, coding, and factuality benchmarks.
- **Reproducibility:** Packaging evaluation prompts, outputs, and ranking code into reproducible public artifacts.
- **Soft Interventions:** Extending intervention logic from hard removal to softer operations, such as layer scaling, routing, or adapter-based suppression.

By pursuing these avenues and maintaining rigorous multi-metric assessment with strict acceptance gating the field can move toward more robustly fair and efficient language models suitable for real-world deployment.

REFERENCES

[1] I. O. Gallegos, R. Rossi, J. Barrow, M. Tanjim, S. Kim, F. Dernoncourt, T. Yu, R. Zhang, and N. Ahmed, "Bias and Fairness in Large Language Models: A Survey," *Computational Linguistics*, vol. 50, pp. 1097–1179, 2024, doi: 10.1162/coli_a_00524.

[2] D. Bouchard, "Bring Your Own Prompts: Use-Case-Specific Bias and Fairness Evaluation for LLMs," 2024.

[3] R. Cantini, A. Orsino, M. Ruggiero, and T. Talia, "Benchmarking adversarial robustness to bias elicitation in large language models: scalable automated assessment with LLM-as-a-judge," *Machine Learning*, vol. 114, 2025, doi: 10.1007/s10994-025-06862-6.

[4] D. Esiobu, X. Tan, S. Hosseini, M. Ung, Y. Zhang, J. Fernandes, J. Dwivedi-Yu, E. Presani, A. Williams, and E. Smith, "ROBBIE: Robust Bias Evaluation of Large Generative Language Models," pp. 3764–3814, 2023, doi: 10.48550/arxiv.2311.18140.

[5] S. Schmidgall, C. Harris, I. Essien, D. Olshvang, T. Rahman, J. Kim, R. Ziaei, J. Eshraghian, P. Abadir, and R. Chellappa, "Evaluation and mitigation of cognitive biases in medical language models," *NPJ Digital Medicine*, vol. 7, 2024, doi: 10.1038/s41746-024-01283-6.

[6] L. Zhong, F. Wan, R. Chen, X. Quan, and L. Li, "BlockPruner: Fine-grained Pruning for Large Language Models," *arXiv preprint arXiv:2406.10594*, 2024, doi: 10.48550/arxiv.2406.10594.

[7] Y. Li *et al.*, "Survey on Bias and Fairness in Large Language Models," *arXiv preprint arXiv:2309.00770*, 2023.

[8] H. Chu *et al.*, "Fairness in Large Language Models: Challenges and Directions," *arXiv preprint arXiv:2401.14183*, 2024.

[9] E. Ferrara, "Should ChatGPT be Biased? Challenges and Risks of Bias in Large Language Models," *First Monday*, vol. 28, no. 11, 2023.

[10] S. Dasu *et al.*, "Attention-Based Fairness Interventions in Transformers," *arXiv preprint arXiv:2501.03452*, 2025.

[11] Z. Lin, S. Chen, Y. Zhang, *et al.*, "Towards Trustworthy Large Language Models: A Review on Debiasing, Evaluation, and Safety," *Artificial Intelligence Review*, 2024.

[12] T. Pagano *et al.*, "Bias Evaluation and Mitigation in Generative Language Models," *arXiv preprint arXiv:2311.08253*, 2023.

[13] S. Caton and C. Haas, "Fairness in Machine Learning: A Survey," *arXiv preprint arXiv:2010.04053*, 2020.

[14] H. Cheng *et al.*, "A Survey of Model Compression and Acceleration for Neural Networks," *arXiv preprint arXiv:2302.04352*, 2023.

[15] T. Liang *et al.*, "Pruning and Quantization for Deep Neural Network Acceleration: A Survey," *Neurocomputing*, vol. 461, pp. 370–403, 2021.

[16] M. Xia *et al.*, "Structured Pruning in Deep Neural Networks: A Survey," *arXiv preprint arXiv:2203.04412*, 2022.

[17] X. Ma, G. Fang, and X. Wang, "LLM-Pruner: On the Structural Pruning of Large Language Models," *arXiv preprint arXiv:2305.11627*, 2023.

[18] Z. Wang, J. Wohlwend, and T. Lei, "Structured Pruning of Large Language Models," in *Proc. EMNLP*, 2020, pp. 6151–6162, doi: 10.18653/v1/2020.emnlp-main.496.

[19] J. Kim *et al.*, "ShortenedLLaMA: Efficient LLM Inference via Layer Truncation," *arXiv preprint arXiv:2403.05612*, 2024.

[20] A. Zayed, G. Mordido, S. Shabanian, I. Baldini, and S. Chandar, "Fairness-Aware Structured Pruning in Transformers," *arXiv preprint arXiv:2312.15398*, 2023.

[21] N. Ding *et al.*, "Parameter-Efficient Fine-Tuning of Large Language Models: A Survey," *arXiv preprint arXiv:2303.15647*, 2023.

[22] A. Aldulaimi *et al.*, "Efficient Adaptation Methods for Large Language Models," *arXiv preprint arXiv:2501.12903*, 2025.

[23] J. Kim *et al.*, "Compressing and Adapting Large Language Models for Efficient Deployment," *arXiv preprint arXiv:2502.01234*, 2025.

[24] Z. Yin, Y. Liu, and T. Chen, "Outlier Weighted Layerwise Sparsity (OWL): A Missing Secret Sauce for Pruning LLMs to High Sparsity," *arXiv preprint arXiv:2310.05175*, 2023.

[25] Y. Guo, M. Guo, J. Su, Z. Yang, M. Zhu, H. Li, M. Qiu, and S. Liu, "Bias in Large Language Models: Origin, Evaluation, and Mitigation," *arXiv preprint arXiv:2411.10915*, 2024, doi: 10.48550/arXiv.2411.10915.

[26] T. Hu, Y. Kyrychenko, S. Rathje, N. Collier, S. van der Linden, and J. Roozenbeek, "Generative Language Models Exhibit Social Identity Biases," *Nature Computational Science*, 2025, doi: 10.1038/s43588-024-00741-1.

[27] N. Huang, H. Fayek, and X. Zhang, "Less Is More? Examining Fairness in Pruned Large Language Models for Summarising Opinions," *arXiv preprint arXiv:2508.17610*, 2025, doi: 10.48550/arXiv.2508.17610.

[28] G. Ling, Z. Wang, Y. Yan, and Q. Liu, "SlimGPT: Layer-wise Structured Pruning for Large Language Models," *arXiv preprint arXiv:2412.18110*, 2024, doi: 10.48550/arXiv.2412.18110.

[29] X. Men, M. Xu, Q. Zhang, B. Wang, H. Lin, Y. Lu, X. Han, and W. Chen, "ShortGPT: Layers in Large Language Models Are More Redundant Than You Expect," *arXiv preprint arXiv:2403.03853*, 2024, doi: 10.48550/arXiv.2403.03853.

[30] S. Tang, O. Sieberling, E. Kurtic, Z. Shen, and D. Alistarh, "DarwinLM: Evolutionary Structured Pruning of Large Language Models," *arXiv preprint arXiv:2502.07780*, 2025, doi: 10.48550/arXiv.2502.07780.

[31] Y. An, X. Zhao, T. Yu, M. Tang, and J. Wang, "FLAP: Fluctuation-based Adaptive Structured Pruning for Large Language Models," *arXiv preprint arXiv:2312.11983*, 2023, doi: 10.48550/arXiv.2312.11983.

[32] M. Nadeem *et al.*, "Fairness Evaluation and Inference Level Mitigation in LLMs," *arXiv preprint arXiv:2406.01124*, 2024.

[33] S. Gupta *et al.*, "Understanding Social Biases in Large Language Models," *arXiv preprint arXiv:2310.03021*, 2023.

[34] Z. Wang, J. Wohlwend, and T. Lei, "Structured Pruning of Large Language Models," in *Proc. EMNLP*, 2020, pp. 6151–6162.

[35] W. Ma *et al.*, "Breaking Down Bias: On The Limits of Generalizable Pruning Strategies," *arXiv preprint arXiv:2311.10452*, 2023.

Amrita Rajput is currently a Research Intern at Georgia Institute of Technology, Atlanta, USA, where she works on bias mitigation in Large Language Models (LLMs). Her research focuses on analyzing internal model components such as neurons and attention heads, and evaluating pruning strategies to improve model-level fairness and reliability. Alongside her research, she has developed end-to-end machine learning systems involving large-scale data processing, computer vision, and NLP-based analysis, with experience in designing pipelines, conducting experiments, and evaluating model performance. Her current research interests include Large Language Models, agentic AI systems, and AI safety.

Vijay K. Madisetti holds the position of Professor of Cybersecurity and Privacy (SCP) at Georgia Tech. He earned his PhD in Electrical Engineering and Computer Sciences from the University of California at Berkeley and is recognized as a Fellow of the IEEE. Additionally, he has been honored with the Terman Medal by the American Society of Engineering Education (ASEE), and was awarded the Georgia Tech Outstanding PhD Dissertation Advisor Award. Professor Madisetti has authored several widely referenced textbooks on topics including cloud computing, data analytics, blockchain, and microservices. Dr. Madisetti is a co-author of the IEEE VHDL Standard for RTL Synthesis, and an inventor on over 80 issued US patents.