LARGE LANGUAGE MODELS SYSTEMATICALLY FAVOR POPULAR OPTIONS: EVIDENCE AND MITIGATION ACROSS MULTIPLE CHOICE TASKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multiple–choice questions (MCQs) are widely used for benchmarking large language models (LLMs). We show that modern LLMs systematically favor *popular* distractors over less–popular correct options. We introduce **PopMCQ**, a *strategy technique* of six stress/control manipulations for MCQs that alter option popularity while keeping the gold label fixed. We apply these strategies to the PlausibleQA evaluation built from NQ, TriviaQA, MuSiQue, and QASC, and quantify bias via the Spearman rank correlation between correctness and *relative* popularity surplus. We then introduce **PopDebias**[1], an inference–time correction that removes a label-free popularity prior and requires *no LLM fine-tuning* (with an optional lightweight calibration step). When averaged across all datasets and strategies, PopDebias improves the accuracy of all 23 models evaluated. This finding holds true at the individual dataset level as well, with the method boosting accuracy for at least 20 of 23 models on every dataset we tested (NQ: 23/23, QASC: 22/23, MuSiQue: 22/23, and TriviaQA: 20/23), demonstrating broad effectiveness.

1 INTRODUCTION

Multiple-choice questions (MCQs) are a standard format for eliciting discrete decisions from large language models (LLMs) in both benchmarks and real-world applications (Hendrycks et al., 2020; Zhong et al., 2023; Zheng et al., 2023; Huang et al., 2023; Wei et al., 2022). However, the construction of MCQs introduces uncontrolled variables. Options often differ in non-semantic properties, such as the general popularity of the entities they contain (Mallen et al., 2023; Kandpal et al., 2023; Ni et al., 2025; Abe et al., 2025). This raises a critical, underexplored question: **Are LLM decisions in MCQ tasks biased by option popularity?** Our investigation suggests that models indeed learn to employ popularity as a heuristic for correctness, Adopting such a heuristic may be a rational strategy (Gigerenzer et al., 2000), as popularity often correlates with correctness in the vast corpora on which these models are trained, making it often a better-than-random strategy under uncertainty. However, this reliance on statistical shortcuts (Geirhos et al., 2020) creates a predictable vulnerability, leading to systematic failures on questions where the factually correct answer contains an obscure entity. This is demonstrated in Figure 1 which displays a sample MuSiQue question (Trivedi et al., 2022) where **Llama-3.1-8B** assigns the highest probability to the most popular distractor (Frank Sinatra, Wikipedia popularity=1.00; Probability=0.408) instead of to the correct yet much less popular (Carol Richards, popularity=0.011), yielding a *Popularity Gap* (POPGAP)(§3.2) of $+0.989$. Isolating such a bias is challenging with standard benchmarks where entity popularity is an incidental and uncontrolled property. We design *six MCQ strategies* within **PopMCQ** (S1–S6;

```
Question: Who sings with
the artist with the best
selling single of all
time on Silver Bells?
Options (Wikipedia
popularity in
parentheses):
A. Pat Boone (0.964)
B. Frank Sinatra (1.0)
C. Carol Richards (0.011)
D. Vera Lynn (0.818)
Answer: C
Model picks: B (p=0.408)
```

Figure 1: MuSiQue (S1). Llama-3.1-8B chooses the most popular distractor (Frank Sinatra, pop. 1.00) over the correct but much less popular (Carol Richards) (0.011), a *PopGap* of $+0.989$.

[1]Code and data will be made available upon publication.

Table 1: Manipulating option popularity (S1–S6) induces notable accuracy fluctuations and strongly negative popularity–correctness correlation under popularity pressure. Cells show the metric (top) and the change relative to S1 Baseline (bottom; red = increase, blue = decrease).

MuSiQue	S1 Baseline	S2 Pop Trap	S3 Pop Grad	S4 Direct	S5 Reverse	S6 None
Accuracy (%)	25.4	26.2	25.1	25.6	26.6	0.20
	(—)	(+0.8)	(-0.3)	(+0.3)	(+1.2)	(-25.2)
Correlation	-0.114	-0.492	-0.536	-0.282	0.137	-0.126
	(—)	(-0.378)	(-0.421)	(-0.168)	(+0.251)	(-0.012)
HPSR (%)	46.5	47.4	64.3	47.1	43.0	62.2
	(—)	(+1.0)	(+17.9)	(+0.7)	(-3.5)	(+15.8)

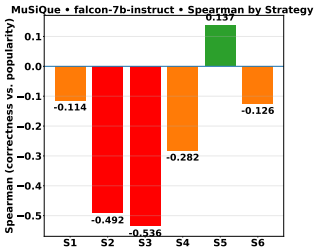

Figure 2: Popularity pressure induces *negative* correctness–popularity correlation.

§ 2.3) that tag options with a popularity signal and vary the popularity configuration under a fixed gold label, enabling causal tests of popularity-driven model behavior.

Through extensive empirical evaluation (§3), with 23 LLMs on four diverse benchmarks, we show that **popularity-distractor bias** is a significant and measurable phenomenon. While recent work suggests that knowledge popularity helps LLMs, our results show the opposite in multiple-choice settings when popularity conflicts with truth. For open-domain knowledge, Mallen et al. (2023) and Ni et al. (2025) find that model accuracy tracks the head–tail frequency of facts—accuracy increases with the pre-training document frequency of those facts, and scaling disproportionately benefits head (popular) facts. Explicitly modeling knowledge popularity increases models' confidence and improves boundary perception, again yielding higher accuracy on popular knowledge (Ni et al., 2025). In contrast, we isolate popularity at the answer option level and show that when popular distractors are set against the correct option, models systematically prefer popularity, hurting accuracy and inducing a strong negative correctness-popularity correlation. For example, in Table 1 which summarizes MuSiQue results for **falcon-7b** we observe that under *popularity pressure* (S1–S6) (§2.3), the model is both less accurate and exhibits a strong *negative* association between correctness and *relative* popularity surplus: the **Spearman** rank correlation between per–item correctness and $(\text{PopSelected} - \text{PopTruth})/\max(\text{PopSelected}, \text{PopTruth})$ (§3.2) is $-$**0.492** (S2), $-$**0.536** (S3), and $-$**0.282** (S4). To mitigate this bias, we propose **PopDebias** (§4) , a lightweight, inference-time correction method that does not require LLM fine-tuning. It divides the model's option probabilities by a *popularity prior* built from the options. Specifically, with model's output probabilities p_i and a popularity-derived prior $q_i \propto (\text{pop}_i + \epsilon)^\alpha$, we form debiased probabilities by dividing out the popularity prior and tempering confidence. Despite its simplicity, PopDebias consistently *improves accuracy*, reduces HPSR and PopGap, and makes Corr *less negative* under popularity pressure.

Summary of Contributions (1) We introduce **PopMCQ**, a strategy-driven technique to evaluate popularity bias in LLMs via MCQ benchmark with six option-popularity manipulations across four QA datasets. (2) We conduct **large-scale experiments** across 23 models and show that LLMs systematically favor popular distractors under *popularity pressure* (S2–S4, S6). In contrast, in control settings where popularity aligns with truth (S5) the effect *attenuates or even flips sign*, supporting the validity of our stress/control design.(3) We pinpoint the cause of this bias as a form of **confidence miscalibration** linked to entity familiarity, where models become overconfident in their choices of popular entities. (4) Lastly, we propose **PopDebias**, a label-free, inference-time correction that significantly improves accuracy and reduces popularity bias with negligible computational cost.

2 Dataset and Strategy Construction

2.1 Source corpora and sampling

We study popularity-driven behavior across three question types spanning factoid, multi-hop, and reasoning using the PlausibleQA dataset (Mozafari et al., 2025a), which is constructed from sampled questions from NQ (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), MuSiQue (Trivedi et al., 2022), and QASC (Khot et al., 2020). Instead of relying on existing large-scale MCQ benchmarks, we adopt PlausibleQA questions formatted in an MCQ style. This decision was driven by the need for experimental control. While existing benchmarks (Talmor et al., 2018; Hendrycks et al., 2020) are useful for measuring general performance, the popularity of their distractors is an incidental and uncontrolled property. To make causal claims about the effect of popularity bias, it is necessary to manipulate it as an independent variable while holding other factors, such as the

question and the correct answer—constant. From the PLAUSIBLEQA dataset, we sample $8{,}000$ questions for evaluation, with $2{,}000$ questions drawn from each of the source datasets: NQ, TriviaQA, MuSiQue, and QASC. For NQ/TriviaQA, we convert each question to an MCQ by adding three distractor answers (§2.3). However, for MuSiQue (multi-hop) and QASC (science MCQ), we keep the original question text but regenerate distractors to control the popularity pressure (§2.3).

2.2 WIKIPEDIA POPULARITY SCORING

To quantify popularity for every option (gold + candidates), we map each surface form to a Wikipedia page via title normalization (Mozafari et al., 2025b). For each mapped page, we retrieve monthly page-view counts over a fixed window (Jan 2015–Dec 2024), we aggregate them, remove outliers with an IQR filter, followed by min–max normalization to $[0, 1]$. We then *relativize* within the option set of each MCQ: for options $i \in \{1..4\}$ we report $\mathrm{pop}_i \leftarrow \frac{p_i}{\max_j p_j} \in [0, 1]$, so that the most popular option in the set has $\mathrm{pop}{=}1.0$. If no corresponding Wikipedia page is found, then $\mathrm{pop}{=}0$. The gold option $a^\star$ is scored identically.

2.3 STRATEGY GENERATION

Each base question yields six MCQ variants that differ *only* in the distractor–popularity configuration. All MCQs have four answer options (one correct, three distractors). Let $\mathcal{C}(q)$ be the candidate pool sorted by descending pop and let HIGH denote the top quartile of $\mathcal{C}(q)$ by pop, MED the middle two quartiles, and LOW the bottom quartile.[2] We randomize answer–ID positions uniformly over {A,B,C,D} to avoid position effects. The six strategies are: (1) **S1 Baseline Control** (*True* $+ 1\times$HIGH$+2\times$LOW), a mixed–pressure reference; (2) **S2 Popular Trap** (*True* $+ 3\times$HIGH), which maximizes popularity pressure against the correct option; (3) **S3 Popularity Gradient** (*True* + three distractors in strictly decreasing pop: HIGH$\rightarrow$MED$\rightarrow$LOW), to expose systematic preference along a popularity ranking; (4) **S4 Direct Popularity Contest** (*True* $+ 1\times$*extremely high*—the top–ranked candidate in $\mathcal{C}(q)$—$+ 2\times$MED; if $a^\star$ is top–ranked, use the next–most–popular as "extremely high"), a head–to–head stress test against the single most popular alternative; (5) **S5 Reverse Popularity Control** (*True* $+ 3\times$LOW), which minimizes popularity pressure; and (6) **S6 None of the Above** (the **correct option** is "None of the above"; distractors are $1\times$ highest–popularity candidate and $2\times$LOW), probing uncertainty handling when the truth is absent.

3 INVESTIGATION ON POPULARITY BIAS

3.1 EXPERIMENTAL SETUP

Models Our study focuses on modern decoder-only LLMs across diverse model families and parameter scales. We experiment with *23 LLMs from popular families*: Llama-3.2-1B/3B (Grattafiori et al., 2024), Llama-3.1-8B (Grattafiori et al., 2024), Llama-3-8B (Grattafiori et al., 2024), Llama-2-7B/13B (Touvron et al., 2023), Qwen2.5-0.5B/1.5B/3B/7B (Qwen et al., 2024), gemma-3-1B/4b/12B (Team et al., 2025), gemma-2-2B/9B/27B (Team et al., 2024), falcon-7B (Almazrouei et al., 2023), Mistral-7B-v0.2/v0.3 (Jiang et al., 2023), Phi-1_5/2/4 (Abdin et al., 2024), and Zephyr-7B (Tunstall et al., 2023). All models are open-source and accessible via HuggingFace, allowing us to extract output probabilities for option ID tokens `A/B/C/D`. Our evaluation protocol follows standard MCQ evaluation frameworks. We access output probabilities of option ID tokens `A/B/C/D` and select the option with maximal probability as the model prediction. All experiments use 0-shot evaluation to eliminate biases from in-context examples. We evaluate each model across all six strategies and four datasets, generating comprehensive popularity bias profiles. See Appendix B and C for detailed evaluation procedures and prompting formats.

3.2 MEASUREMENT OF POPULARITY BIAS

We define *popularity bias* as a model's systematic inclination to select more popular entities over factually correct yet less popular ones in multiple-choice settings. Each option o is annotated with a scalar popularity score $\mathrm{Pop}(o) \in [0, 1]$ (Wikipedia popularity in our data).

3.2.1 BIAS METRICS

Relative popularity surplus. For an item with a ground-truth option $o^\star$ and a model-selected option $\hat{o}$, we compute the *relative popularity surplus*

$$\mathrm{RPS} \;=\; \frac{\mathrm{Pop}(\hat{o}) - \mathrm{Pop}(o^\star)}{\max\{\mathrm{Pop}(\hat{o}),\, \mathrm{Pop}(o^\star)\} + \varepsilon}, \tag{1}$$

[2] If a stratum lacks sufficient items, we borrow from the nearest stratum while preserving rank order.

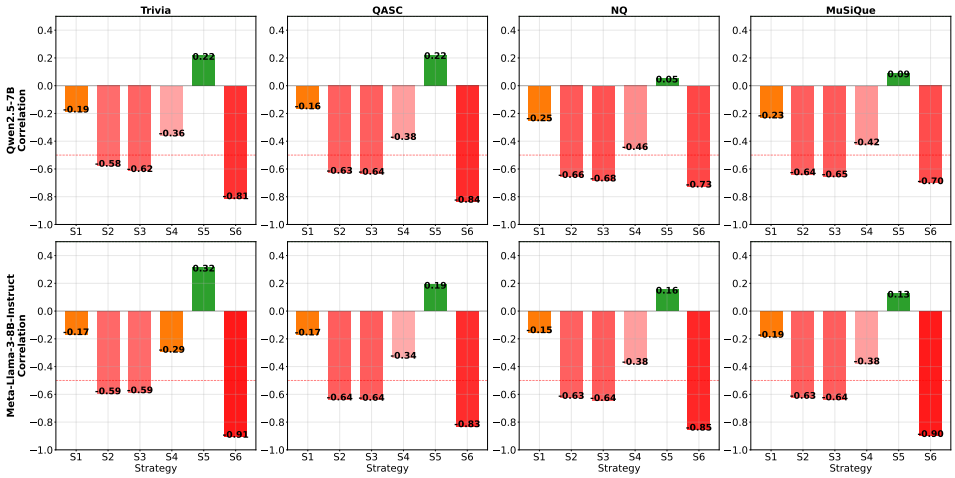

Figure 3: Popularity bias across strategies and datasets for representative models `Llama-3-8B` and `Qwen2.5-7B`. Y-axis shows correlation (Spearman between correctness and relative popularity surplus). Negative values indicate popularity bias (selecting popular options correlates with being wrong), while positive values show appropriate popularity-accuracy alignment. S6 (None of the Above) exhibits strongest bias, while S5 (Reverse Control) validates experimental design with positive correlations. Complete results for all 23 models are provided in Appendix E.

with $\varepsilon = 10^{-6}$ to avoid division by zero. $\mathrm{RPS} > 0$ means the chosen option is more popular than the correct option; $\mathrm{RPS} < 0$, the opposite.

Correlation. We quantify bias as the rank correlation between correctness and surplus: $\mathrm{Corr} = \mathrm{Spearman}\big(\mathbf{1}\{\hat{o} = o^\star\}, \mathrm{RPS}\big)$. Negative values indicate popularity bias; positive values indicate alignment when popularity and correctness coincide (e.g., S5).

Popularity Gap (PopGap). $\mathrm{PopGap} = \mathbb{E}\big[\mathrm{Pop}(\hat{o}) - \mathrm{Pop}(o^\star)\big]$. Values > 0 reflect systematic *overshooting* toward popular options.

High-Popularity Selection Rate (HPSR). Within each item (with K options), rank by $\mathrm{Pop}(\cdot)$ and mark a selection as high-pop if its rank is in the top half: $\mathrm{HPSR} = \mathbb{E}\big[\mathbf{1}\{\mathrm{rank}_{\mathrm{pop}}(\hat{o}) \le \lfloor K/2 \rfloor\}\big]$. Higher HPSR $\Rightarrow$ stronger pull toward common entities.

3.2.2 Performance and Calibration Metrics

Beyond measuring bias directly, we evaluate how popularity pressure impacts model performance and calibration. The primary measure of task success is **Accuracy**, defined as the standard proportion of correct answers, $\mathrm{Acc} = \mathbb{E}[\mathbf{1}\{\hat{o} = o^\star\}]$. We therefore measure **Confidence**, the model's average predicted probability of its chosen option, $\mathrm{Conf} = \mathbb{E}[\max_i p_i]$.

To assess how well the confidence aligns with correctness, we compute **Alignment** (Ni et al., 2025), a proxy measure for calibration defined as $\mathrm{Align} = \mathbb{E}[1 - |\mathbf{1}\{\hat{o} = o^\star\} - \max_i p_i|]$. A well-aligned model is confident when it is correct and uncertain when it is wrong, maximizing this score. Finally, to directly test our hypothesis that familiarity drives the bias, we introduce a diagnostic metric: **Overconfidence by Popularity Bucket**. Here, we partition predictions based on the popularity of the selected answer (Low, Med, High) and report the average gap between confidence and accuracy in each bucket, $\mathrm{OverConf}(b) = \mathbb{E}[\max_i p_i - \mathbf{1}\{\hat{o} = o^\star\} \mid b]$. This metric allows us to precisely quantify the extent to which entity popularity leads to unwarranted overconfidence (used in Fig. 4).

3.3 Key Observations

We conduct an extensive evaluation with 23 LLMs across four benchmarks to understand the effects of popularity-distractor bias. We show a representative subset of results on the MuSiQue dataset in Table 2 for a brief presentation and provide the full results in Appendix E. We draw the following main observations:

Popularity pressure produces systematic negative correlations. Across models and datasets, strategies that push toward common entities (S2–S4, S6) yield negative correlations, with S6 ("None of the Above") being the most extreme. For **Llama-3.1-8B**, S6 correlations span -0.80 (MuSiQue) to -0.81 (TriviaQA); for **Qwen2.5-7B**, they range from -0.70 (MuSiQue) to -0.84 (QASC), consistent with Fig. 3. Averaged over *all* 23 models, the S6 means are -0.668 (MuSiQue), -0.668 (NQ), -0.715 (QASC), and -0.704 (TriviaQA) (see Tables 17b, 14b, 11b, 8b). S2/S3 also show strong negatives: dataset means cluster around -0.56 to -0.61 (MuSiQue/NQ/QASC/TriviaQA; see Tables 15b, 16a, 12b, 13a, 9b, 10a,6b, 7a). Consistent with these correlation patterns, Table 2 reports MuSiQue accuracies for four representative models across S1–S6. Under S2/S3, **Llama-3-8B** and **Qwen2.5-7B** show small gains over S1 (49.6→51.0/51.2 and 49.5→51.0/52.3), whereas accuracy collapses under S6 (39. and 8.2, respectively). **Gemma-2-9B** follows the same trend (32.1→10.1 on S6), while **Gemma-2-27B** is an outlier that improves under S2/S3 and attains 56.1% on S6, suggesting stronger handling of "None of the Above."

Table 2: Accuracy (%) on the multi-hop **MuSiQue dataset** for representative models across our six strategies.

Strategy	Meta-Llama-3-8B	Qwen2.5-7B	Gemma-2-27B	Gemma-2-9B
S1: Baseline	49.6	49.5	39.8	32.1
S2: Pop Trap	51.0	51.0	41.1	31.0
S3: Gradient	51.2	52.3	42.1	33.9
S4: Direct	48.5	49.9	39.5	30.9
S5: Reverse	50.9	52.4	37.6	30.1
S6: None	39.6	8.2	56.1	10.1

When popularity aligns with truth, the sign flips and accuracy peaks. Reverse control (S5) reliably yields *positive* correlations: dataset-level means are $+0.123$ (MuSiQue), $+0.124$ (NQ), $+0.204$ (QASC), $+0.263$ (TriviaQA). Accuracy also rises: for **Llama-3.1-8B**, S1→S5 gives $+7.3$ points on TriviaQA (64.9%→72.2%) and $+3.7$ on NQ (58.5%→62.2%); **Qwen2.5-7B** similarly improves to 60.5–67.2% on NQ/TriviaQA (Tables 9a, 11a, 15a, 17a). Notably, POPGAP turns slightly negative in S5 (typically -0.03 to -0.08), confirming the model follows correctness when popularity helps rather than hurts.

S6 exposes a failure to abstain under uncertainty. Under S6, models rarely choose "None of the Above" even when it is correct, preferring the most popular named entity. For **Qwen2.5-7B**, S6 accuracies are only 8–16% (MuSiQue: 8.2%, NQ: 8.8%, QASC: 14.3%, TriviaQA: 16.1%) while high-pop selection rates jump to 58–64% and POPGAP inflates to $+0.32$–$+0.41$ across the same datasets. Alignment simultaneously degrades (e.g., **Mistral-7B-v0.2** on TriviaQA: confidence 0.94 with alignment $\approx$ 0.12), revealing confident-but-wrong behavior in ambiguous cases (Tables 8b, 11b, 14b, 17b).

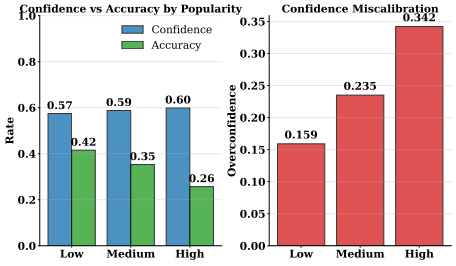

Figure 4: **Confidence vs. accuracy by popularity.** Popularity buckets follow the script: *Low* ≤ 0.3, *Medium* $(0.3, 0.7]$, *High* > 0.7.

Overall, Fig. 3 together with Tables (15b, 16a, 12b, 13a, 9b, 10a, 6b, 7a) suggest that: (i) popularity pressure consistently steers models toward common-but-wrong options; (ii) the effect peaks when the correct action is to abstain (S6); and (iii) aligning popularity with truth (S5) reverses the sign and lifts accuracy, showing the negative correlations in S2/S3/S6 reflect a genuine, systematic bias rather than limited capability.

3.4 WHAT CONTRIBUTES POPULARITY BIAS?

Confidence miscalibration linked to entity popularity. Models become *more confident* as the popularity of the selected entity increases, even while becoming *less accurate*. As shown In Figure 4, where results are aggregated across models, confidence rises **0.57**→**0.59**→**0.60** across *Low*→*Medium*→*High* popularity, while accuracy falls **0.42**→**0.35**→**0.26**. The resulting overconfidence (confidence−accuracy) therefore grows sharply: **+0.159** (Low) → **+0.235** (Medium) → **+0.342** (High), a **+0.183** absolute increase. In other words, moving from low to high-pop entities, models are *3% more confident* yet *16% less accurate*, quantifying a strong familiarity⇒certainty shortcut.

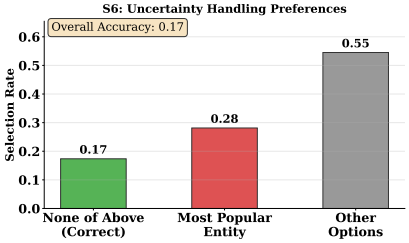

Figure 5: **S6 uncertainty handling.** Bars show selection rates; the inset reports overall accuracy.

Inadequate uncertainty handling in ambiguous contexts. When uncertainty is required (S6: *None of the Above*), models overwhelmingly choose named entities instead of acknowledging un-

certainty. Concretely, the correct *None of the Above* option is selected only **17%** of the time, while the single *most popular* named entity is chosen **28%** of the time (a **+11%** gap; **1.65**× as often) as shown in Figure 5. The remaining **55%** of selections go to other (less popular) named options. Because the gold label in S6 is the none-of-above choice, overall accuracy is just **17%**, showing that a learned preference for popular names overwhelms appropriate uncertainty expression in this setting.

Family-Level Popularity Bias We group models by their family and summarize popularity-distractor bias under **Strategy S2 (Popular Trap)** using the correlation between correctness and relative popularity surplus (lower is better; more negative indicates stronger bias). As shown in Figure 6, *all* families exhibit clear negative correlations under S2, indicating a systematic pull toward common but incorrect options when popularity pressure is high. These family-level aggregates highlight that the effect is widespread rather than model-specific.

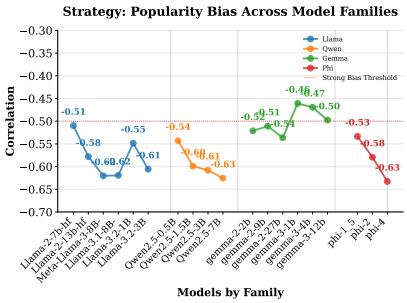

Figure 6: **Strategy S2 (Popular Trap): family-level popularity bias.** Negative values indicate stronger popularity–distractor bias.

4 Debiasing Method

4.1 Problem Setup

We consider an MCQ instance with options $o_1, \ldots, o_K$ ($K \in \{4\}$). A base model outputs a probability vector $\mathbf{p} = (p_1, \ldots, p_K)$ and each option is annotated with a scalar popularity score $\mathrm{Pop}(o_i) \in [0, 1]$ (Section 3.2). Our objective is to transform $\mathbf{p}$ into a debiased distribution $\tilde{\mathbf{p}}$ that reduces the tendency to over-select common entities while retaining signal that aligns fame and truth (S5).

Notation. We write $\mathrm{rank}_{\mathrm{pop}}(o_i)$ for the within-item popularity rank (1 = most popular), $\bar{\mathrm{Pop}}$ and $\mathrm{Var}(\mathrm{Pop})$ for the mean and variance of per-item popularities, respectively. Throughout the paper, ε denotes a small constant for numerical stability, and $\mathrm{softmax}$ performs elementwise exponentiation and normalization.

4.2 Popularity as a Fame Prior

Following the decomposition view used for selection biases in MCQs (e.g., position/token bias) (Zheng et al., 2023), we treat fame as a *probability prior* that skews the observed distribution:

$$p_i \;\propto\; \underbrace{\pi_i}_{\text{fame prior, } \uparrow \text{ in Pop}} \cdot \underbrace{\phi_i}_{\text{fame-free belief}} \qquad \Rightarrow \qquad \phi_i \;\propto\; \frac{p_i}{\pi_i}. \tag{2}$$

Our method estimates π from *observable, label-free* patterns in the current item, divides it out, then tempers confidence to prevent overcorrection. Unlike permutation averaging, it operates in a single pass over options.

4.3 PopDebias (Popularity-Prior)

We estimate a nonnegative *bias strength* b from features that are available at inference time:

$$\mathbf{f} \;=\; \big[\, \mathrm{Var}(\mathrm{Pop}),\; \max_i \mathrm{Pop}(o_i),\; \mathrm{Pop}(o_{\hat{\imath}}),\; \mathrm{rank}_{\mathrm{pop}}(o_{\hat{\imath}}),\; \max_i p_i \,\big],$$

where $\hat{\imath} = \arg\max_i p_i$ is the model's top prediction. Intuitively, bias should rise when the model is very confident on a very common option and the popularity distribution has low variance (i.e., distractors cluster in fame).

Bias estimation. On a small calibration split, we fit a lightweight regressor f_θ to predict the true *popularity gap* between the model's choice and the correct answer:

$$\theta \;\leftarrow\; \arg\min_\theta \sum_{(q,x)\in\mathcal{D}_{\mathrm{cal}}} \Big\| f_\theta(\mathbf{f}(q,x)) - \big(\mathrm{Pop}(o_{\hat{\imath}}) - \mathrm{Pop}(o^\star)\big) \Big\|^2. \tag{3}$$

At test time (no labels), we set $b = f_\theta(\mathbf{f})$. If a regressor is not available, we use a bounded heuristic scaled by a calibration constant $\bar{b}$:

$$b \;\propto\; \frac{\mathrm{Pop}(o_{\hat{\imath}})}{\bar{\mathrm{Pop}} + \varepsilon} \;+\; \max_i p_i \;+\; \frac{1}{\mathrm{Var}(\mathrm{Pop}) + \varepsilon}, \qquad b \leftarrow \bar{b}\,\mathrm{clip}(b;\, 0, \infty). \tag{4}$$

Constructing the fame prior and debiasing. Given b, we build a within-item prior that increases with popularity and apply a single-pass reweighting with *adaptive* confidence tempering:

$$w_i \;=\; 1-b\cdot\frac{\mathrm{Pop}(o_i)}{\max_j \mathrm{Pop}(o_j)+\varepsilon}, \qquad \gamma(b,c,\psi) \;=\; \mathrm{clip}\big(1-\beta\, b\, c\, \psi,\; 0,\; 1\big), \qquad \widetilde{p}_i \;\propto\; \Big(\gamma\, p_i + \tfrac{1-\gamma}{K}\Big)\, w_i, \tag{5}$$

followed by normalization $\widetilde{\mathbf{p}} \leftarrow \widetilde{\mathbf{p}}/\sum_j \widetilde{p}_j$. Here w_i implements π_i^{-1} up to a bounded linearization that avoids extreme division in small-option regimes. The tempering coefficient γ depends on (i) the estimated bias b, (ii) model confidence $c = \max_i p_i$, and (iii) a *popularity pressure* factor ψ defined from the entropy of the normalized popularity profile $r_i \propto \mathrm{Pop}(o_i)$:

$$\psi \;=\; 1 - \frac{H(r)}{\log K}, \qquad H(r) = -\sum_{i=1}^{K} r_i \log(r_i+\varepsilon). \tag{6}$$

ψ increases when one or two options dominate in fame (low entropy), indicating a sharper "popular trap." The scalar $\beta > 0$ controls the overall tempering strength (we set β on a small calibration split). When b is small or the popularity profile is flat, $\gamma \approx 1$ and $\widetilde{\mathbf{p}} \approx \mathbf{p}$.

We use a small calibration ratio (default 10%) to fit f_θ, set $\bar{b}$ in Eq. 4, and pick β for Eq. 5. Inference is label-free and adds $O(K)$ per-item overhead (feature extraction and one closed-form update). Unlike permutation-averaging debiasing for position/token biases (Zheng et al., 2023), our method performs a single forward pass and estimates a within-item *fame prior* from observable popularity patterns. No option permutations are used at inference. The overall procedure of PopDebais is summarized as Algorithm 1.

The method is $O(K)$ per item and adds a tiny linear regressor trained once on $\mathcal{D}_{\mathrm{cal}}$ (dozens–hundreds of samples suffice in practice). We clip w_i to $[0,1]$, use $\varepsilon = 10^{-10}$, and constrain $\gamma \in [0,1]$ via the $\mathrm{clip}(\cdot)$ in Eq. 5. Statistical tests (e.g., McNemar) and bias metrics follow in Section 3.2. For intuition about how b, c, and ψ shape γ and the final reweighting, we refer the reader to the step-by-step example in Appendix D.

Algorithm 1 POPDEBIAS: Implementation-oriented pseudocode (single pass)

Require: base model $\mathcal{M}$, items $\mathcal{D}$, (optional) regressor f_θ, hyperparams $\beta > 0, \bar{b} > 0, \varepsilon = 10^{-10}$, caps $b_{\max}$, $\gamma \in [0,1]$
Ensure: debiased predictions $\mathcal{Y}$

(One-time calibration) Fit f_θ on a small split to predict popularity gap; choose β and set heuristic scale $\bar{b}$.
for item $(q,x) \in \mathcal{D}$ **do**
  $\mathbf{p} \leftarrow \mathcal{M}(q,x)$; get per-option $\mathrm{Pop}(o_i)$; $K \leftarrow |\{o_i\}|$; $c \leftarrow \max_i p_i$; $\hat{\imath} \leftarrow \arg\max_i p_i$
  Bias strength b**:**
    if f_θ exists **then** $b \leftarrow \mathrm{clip}(f_\theta(\mathbf{f}), 0, b_{\max})$
    else $b \leftarrow \bar{b} \cdot \mathrm{clip}(\text{heuristic from Eq. 4}, 0, b_{\max})$
  Early-exit check (cheap): if $b \cdot c < \tau$ **then** output $\arg\max_i p_i$ and **continue** ▷ τ small, e.g., 0.02
  Popularity pressure ψ**:** normalize $r_i \leftarrow \dfrac{\mathrm{Pop}(o_i)}{\sum_j \mathrm{Pop}(o_j)+\varepsilon}$; compute $H(r)$ and ψ via Eq. 6
  Tempering: $\gamma \leftarrow \mathrm{clip}(1-\beta\, b\, c\, \psi, 0, 1)$
  Within-item prior weights: $w_i \leftarrow \mathrm{clip}\Big(1 - b\,\dfrac{\mathrm{Pop}(o_i)}{\max_j \mathrm{Pop}(o_j)+\varepsilon},\; 0,\; 1\Big)$
  Degeneracy guard: if $\sum_i w_i = 0$ **then** set $w_i \leftarrow \varepsilon$ for all i
  Single-pass update: $\widetilde{p}_i \propto \big(\gamma\, p_i + \frac{1-\gamma}{K}\big)\, w_i$; normalize $\widetilde{\mathbf{p}}$
  Add $\arg\max_i \widetilde{p}_i$ to $\mathcal{Y}$
end for
return $\mathcal{Y}$

5 EXPERIMENT RESULTS

5.1 MAIN RESULTS

We evaluate **PopDebias** on 23 open-source LMs across four MCQ benchmarks. Averaged over all datasets and strategies, accuracy improves from **37.3%** to **43.3%** (**+6.0** points), while the mean absolute correlation $|\rho|$ decreases from **0.408** to **0.388**. Accuracy improves on **23/23** models and $|\rho|$ is reduced on **19/23** models. To show the effect in a head-to-head setting, Table 3 reports the results on **S4 (Direct Contest)** across all datasets: PopDebias consistently raises accuracy and lowers both HPSR and $|\rho|$, with strong gains even on smaller/base models. For example, **Falcon-7B** on NQ

Table 3: **Comprehensive Debiasing Results on Strategy S4 (Direct Contest) Across All Models and Datasets.** We report Accuracy (%), Correlation (ρ), and High-Pop Selection Rate (HPSR, %) in a "Before → After" format. Improvements are in **bold**. Asterisks (*) denote statistically significant accuracy improvements ($p < 0.05$, McNemar's test).

	NQ			QASC			MuSiQue			TriviaQA		
Model	**Acc (%)**	**Corr (ρ)**	**HPSR (%)**	**Acc (%)**	**Corr (ρ)**	**HPSR (%)**	**Acc (%)**	**Corr (ρ)**	**HPSR (%)**	**Acc (%)**	**Corr (ρ)**	**HPSR (%)**
Llama-3-8B	56.1 → **57.2***	-0.38 → **-0.35**	38.5 → **36.9**	52.1 → **52.6**	-0.34 → **-0.32**	48.5 → **47.5**	48.5 → **49.7***	-0.38 → **-0.34**	42.8 → **40.5**	64.8 → 65.0	-0.29 → **-0.27**	43.2 → **42.6**
Llama-3.1-8B	57.6 → **59.5***	-0.43 → **-0.35**	39.3 → **35.9**	52.0 → **53.0**	-0.33 → **-0.28**	47.8 → **45.0**	48.9 → **50.1***	-0.33 → **-0.22**	41.4 → **35.2**	64.8 → 65.2	-0.32 → **-0.30**	43.4 → **42.7**
Llama-3.2-3B	47.5 → **51.5***	-0.37 → **-0.25**	39.8 → **32.9**	45.2 → **47.0***	-0.35 → **-0.28**	50.5 → **45.5**	42.1 → **44.7***	-0.36 → **-0.18**	45.8 → **34.4**	51.5 → **53.0***	-0.32 → **-0.22**	46.2 → **42.0**
Llama-3.2-1B	35.0 → **41.6***	-0.37 → **-0.22**	46.4 → **33.2**	32.9 → **35.9***	-0.28 → **-0.19**	49.1 → **40.9**	35.9 → **37.6***	-0.36 → **-0.18**	48.2 → **35.4**	32.5 → **35.5***	-0.27 → **-0.16**	49.0 → **40.3**
Llama-2-13B	46.9 → **51.3***	-0.38 → **-0.21**	41.1 → **31.5**	37.5 → **39.9***	-0.28 → **-0.18**	47.9 → **41.0**	41.2 → **45.0***	-0.39 → **-0.19**	47.2 → **34.8**	50.7 → **52.5***	-0.29 → **-0.17**	45.0 → **40.0**
Llama-2-7B	29.6 → **35.0***	-0.30 → **-0.18**	44.1 → **32.1**	30.2 → **35.3***	-0.26 → **-0.05**	48.9 → **31.0**	30.1 → **36.0***	-0.30 → **-0.13**	46.6 → **31.0**	34.1 → **37.8***	-0.24 → **-0.07**	45.6 → **34.0**
Qwen2.5-7B	56.1 → **58.8***	-0.46 → **-0.38**	41.6 → **37.6**	55.9 → **57.2***	-0.38 → **-0.33**	48.2 → **45.6**	49.9 → **51.4***	-0.42 → **-0.32**	45.8 → **40.0**	60.8 → **61.5**	-0.36 → **-0.30**	45.8 → **43.9**
Qwen2.5-3B	50.7 → **54.5***	-0.42 → **-0.33**	42.0 → **36.5**	52.8 → **54.4***	-0.38 → **-0.32**	49.5 → **46.3**	46.4 → **49.0***	-0.37 → **-0.20**	44.8 → **35.2**	55.0 → **55.9**	-0.32 → **-0.23**	45.3 → **41.5**
Qwen2.5-1.5B	46.1 → **49.0***	-0.36 → **-0.27**	40.5 → **34.9**	48.2 → **49.9***	-0.34 → **-0.28**	49.2 → **45.4**	42.1 → **45.0***	-0.35 → **-0.22**	45.2 → **36.0**	44.6 → **46.0***	-0.29 → **-0.18**	46.2 → **41.1**
Qwen2.5-0.5B	35.6 → **41.3***	-0.35 → **-0.22**	43.6 → **33.1**	38.3 → **39.8***	-0.28 → **-0.21**	48.3 → **42.7**	34.9 → **39.5***	-0.34 → **-0.17**	46.0 → **32.6**	30.9 → **34.8***	-0.22 → **-0.09**	45.8 → **36.5**
Gemma-2-27B	39.8 → **40.6**	-0.33 → **-0.31**	41.5 → **39.6**	28.8 → 29.1	-0.23 → **-0.22**	45.8 → **44.6**	39.5 → **40.1**	-0.32 → **-0.30**	43.4 → **41.6**	46.0 → 46.2	-0.22 → **-0.21**	42.5 → **41.6**
Gemma-2-9B	32.6 → 33.0	-0.30 → **-0.29**	42.1 → **40.6**	28.8 → 28.9	-0.25 → **-0.23**	47.5 → **46.2**	30.9 → **31.4**	-0.28 → **-0.24**	43.3 → **40.4**	34.0 → 34.4	-0.23 → **-0.21**	44.8 → **43.7**
Gemma-2-2B	31.6 → **35.4***	-0.34 → **-0.26**	45.6 → **37.5**	30.4 → **32.2***	-0.24 → **-0.18**	47.1 → **41.7**	31.9 → **33.4***	-0.35 → **-0.28**	46.7 → **40.5**	31.3 → **32.4**	-0.21 → **-0.17**	44.6 → **41.5**
Mistral-7B-v0.3	57.0 → **57.8**	-0.42 → **-0.38**	39.4 → **37.4**	44.8 → **45.4**	-0.33 → **-0.29**	49.4 → **47.0**	47.0 → **48.4***	-0.36 → **-0.29**	43.9 → **39.5**	59.1 → 59.5	-0.33 → **-0.31**	44.1 → **43.2**
Zephyr-7B	52.5 → **53.6***	-0.41 → **-0.38**	40.7 → **39.0**	42.0 → 42.5	-0.31 → **-0.29**	49.0 → **47.5**	45.4 → 46.0	-0.35 → **-0.31**	44.1 → **41.6**	56.9 → 57.4	-0.34 → **-0.31**	46.0 → **45.0**
Falcon-7B	25.1 → **32.9***	-0.30 → **-0.14**	46.2 → **30.0**	25.0 → **30.1***	-0.23 → **-0.06**	48.8 → **32.0**	25.6 → **31.8***	-0.28 → **-0.12**	47.1 → **30.2**	27.6 → **31.8***	-0.18 → **-0.01**	44.2 → **30.1**

rises **25.1%** → **32.9%** (**+7.8** pp) while cutting HPSR **46.2%** → **30.0%**; **Llama-2-7B** on QASC improves **30.2%** → **35.3%** and largely neutralizes the correlation ($-0.26 \rightarrow \mathbf{-0.05}$). *We select β via a brief hyperparameter scan and use $\beta = 10$ by default; see App. G for full cross-dataset sweeps (Figs. 10–12) and analysis.*

Table 4: **Debiasing Effectiveness Summary on Adversarial Strategies.** Average accuracy gain (ΔAcc, pp), average correlation improvement ($\Delta\rho$), and the fraction of models with statistically significant improvements (Sig. Rate) for each strategy and dataset.

	S2: Pop Trap			S3: Gradient			S4: Direct			S6: None		
Dataset	**ΔAcc**	**$\Delta\rho$**	**Sig.**	**ΔAcc**	**$\Delta\rho$**	**Sig.**	**ΔAcc**	**$\Delta\rho$**	**Sig.**	**ΔAcc**	**$\Delta\rho$**	**Sig.**
NQ	+7.4	+0.08	22/23	+8.7	+0.06	23/23	+3.0	+0.10	19/23	+16.3	+0.03	21/23
QASC	+7.8	+0.06	22/23	+8.2	+0.05	23/23	+2.4	+0.07	16/23	+14.8	+0.01	21/23
MuSiQue	+9.4	+0.11	23/23	+10.3	+0.09	23/23	+2.9	+0.10	16/23	+15.5	+0.02	22/23
TriviaQA	+4.7	+0.07	19/23	+5.0	+0.05	20/23	+1.6	+0.08	11/23	+16.0	+0.01	21/23

Complementing this, Table 4 aggregates average gains on the adversarial strategies (S2/S3/S4/S6): the largest accuracy lifts appear under strong popularity pressure (S2/S3), S6 shows large recoveries when "None of the Above" is present, and S4 delivers consistent improvements with many statistically significant wins. Overall, the pattern is clear—**PopDebias** curbs popularity chasing and converts it into measurable accuracy gains across models and datasets. Full per-model, per-strategy results are reported in the Appendix for TriviaQA (Tables 18–23), QASC (Tables 24–29), NQ (Tables 30–35), and MuSiQue (Tables 36–41).

5.2 Prompting Strategies

To assess whether prompting (§C) alone mitigates popularity bias—and how it compares to post-hoc debiasing—we run an ablation on **MuSiQue, S2 (Popular Trap)** using three prompting conditions (**Standard**, **Bias Warning**, **Chain-of-Thought**) and then apply two debiasers: **PrideDebias** (Zheng et al., 2023) and our **PopDebias**. As shown in Table 5, prompting by itself produces only modest and inconsistent accuracy changes and leaves both PopGap and ρ largely unaffected. **PrideDebias** offers limited benefit and can even decrease accuracy in some cases (e.g., **Llama-3-8B**: 51.0 → 49.9; **Qwen-2.5-7B**: 51.0 → 50.5 under *Standard*). In contrast, **PopDebias** consistently improves accuracy and reduces High-Pop and PopGap, while moving ρ toward zero across all prompts and models: under *Standard*, **Qwen-2.5-7B** increases to **61.5**; under *Bias Warning*, **Llama-3-8B** reaches **57.8**; and under *CoT*, **Gemma-2-27B** reaches **53.6**. These results indicate that **PopDebias** provides a more reliable post-hoc correction than permutation-based prior estimation and complements prompting strategies without requiring retraining.

Table 5: Prompt Analysis on S2 (Popular Trap) Strategy on MuSiQue. Comparison of standard prompting, bias warnings, and chain-of-thought reasoning with and without PopDebias application.

Method	**Model**	**Acc.**	**High-Pop**	**PopGap**	**Corr (ρ)**
Standard	Llama-3-8B	51.0	35.5	0.158	-0.625
	+PrideDebias	49.9	36.0	0.163	-0.622
	+PopDebias	**55.7**	**30.0**	**0.108**	**-0.583**
	Qwen-2.5-7B	51.0	36.2	0.174	-0.639
	+PrideDebias	50.5	36.3	0.177	-0.640
	+PopDebias	**61.5**	**23.7**	**0.065**	**-0.536**
	Gemma-2-9B	31.0	44.4	0.233	-0.502
	+PrideDebias	32.8	44.8	0.233	-0.529
	+PopDebias	**37.2**	**36.8**	**0.164**	**-0.486**
	Gemma-2-27B	41.1	38.5	0.193	-0.561
	+PrideDebias	42.2	38.3	0.190	-0.573
	+PopDebias	**47.0**	**31.6**	**0.129**	**-0.534**
Bias Warning	Llama-3-8B	52.1	33.9	0.151	-0.618
	+PrideDebias	51.5	34.3	0.152	-0.619
	+PopDebias	**57.8**	**27.1**	**0.091**	**-0.557**
	Qwen-2.5-7B	49.2	38.0	0.191	-0.639
	+PrideDebias	50.2	37.8	0.185	-0.456
	+PopDebias	**59.6**	**25.1**	**0.076**	**-0.535**
	Gemma-2-9B	39.2	39.3	0.204	-0.541
	+PrideDebias	40.2	39.1	0.200	-0.546
	+PopDebias	**45.4**	**32.0**	**0.138**	**-0.517**
	Gemma-2-27B	45.9	35.9	0.176	-0.577
	+PrideDebias	47.8	35.4	0.171	-0.595
	+PopDebias	**51.4**	**29.1**	**0.115**	**-0.538**
CoT	Llama-3-8B	50.6	35.4	0.160	-0.613
	+PrideDebias	50.5	35.6	0.158	-0.404
	+PopDebias	**57.0**	**27.9**	**0.093**	**-0.561**
	Qwen-2.5-7B	51.3	36.0	0.172	-0.638
	+PrideDebias	51.6	35.8	0.170	-0.636
	+PopDebias	**61.6**	**24.7**	**0.072**	**-0.558**
	Gemma-2-9B	35.7	41.9	0.216	-0.530
	+PrideDebias	37.6	41.5	0.213	-0.538
	+PopDebias	**42.2**	**34.5**	**0.147**	**-0.514**
	Gemma-2-27B	49.0	36.2	0.169	-0.611
	+PrideDebias	48.8	36.3	0.170	-0.610
	+PopDebias	**53.6**	**30.4**	**0.117**	**-0.573**

5.3 Adaptive Tempering: Selecting β

Our adaptive tempering uses $\gamma(b, c, \psi) = \mathrm{clip}\big(1 - \beta\, b\, c\, \psi, 0, 1\big)$ to shrink confidence when predictions are both confident and made under strong popularity pressure. Figure 7 (TriviaQA) illustrates the consistent trend we observe across datasets:

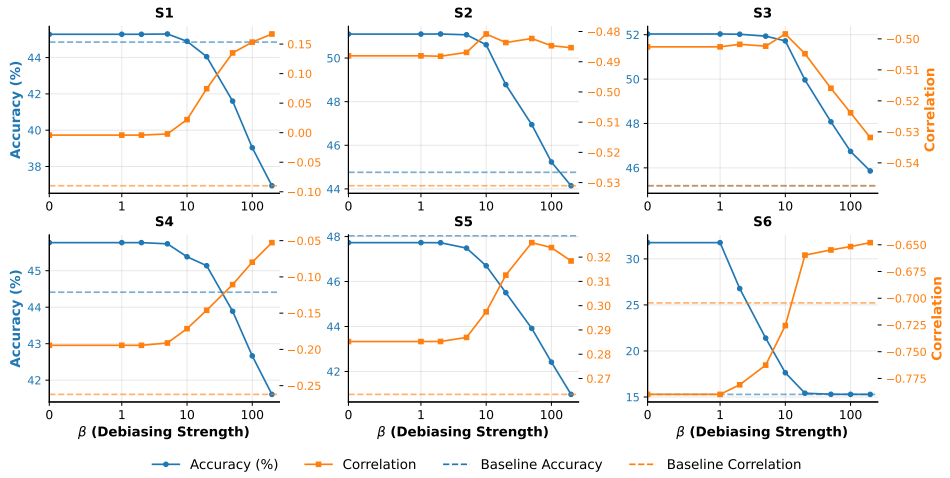

Figure 7: **TriviaQA β sweep (S1–S6).** Moderate tempering ($\beta \approx 10$–20) lowers $|\mathrm{Corr}|$ and improves/preserves accuracy under popularity pressure (S2/S3/S6). Very large β can over-temper when popularity aligns with truth (S5).

under S2/S3/S6, increasing β from 0 to **10–20** reliably *reduces* $|\mathrm{Corr}|$ and *improves or preserves* accuracy; S1 is largely flat; and in the reverse control S5, very large β (e.g., ≥ 50) can slightly shave accuracy by over-tempering a sometimes helpful popularity signal. Based on these curves, we use β=10 by default and reserve larger values only for stress settings. Full cross-dataset sweeps are provided in Appendix G (NQ: Fig. 10; MuSiQue: Fig. 11; QASC: Fig. 12).

6 Related Work

A growing body of work links LLM performance to the head–tail frequency of knowledge. Kandpal et al. (2023) show that models learn head (popular) facts much more readily than tail facts and that scale disproportionately benefits head knowledge. Razeghi et al. (2022) further demonstrate that pre-training duplication frequency strongly predicts downstream success, reinforcing the view that distributional popularity shapes what models recall. Our study asks a complementary question in MCQs: when popularity conflicts with truth at *inference time*, do models over-weight popular options?. Formatting choices in MCQs can systematically sway model predictions. Pezeshkpour & Hruschka (2023); Zheng et al. (2023) document robust *position bias* in multiple-choice settings, where merely reordering options shifts LLM choices, suggesting that non-semantic cues (beyond knowledge) can drive errors. We add a distinct, orthogonal factor—*option popularity*—and show that it reliably induces accuracy drops and negative correctness–popularity associations under adversarial strategies. Work on LLM calibration and familiarity effects observes that models can be confidently wrong, especially on tail facts; popularity/frequency often correlates with both accuracy and confidence (Kandpal et al., 2023; Razeghi et al., 2022). Our results connect this to MCQ behavior: popularity acts like a label-free prior that inflates confidence on well-known entities, motivating a prior-division correction.

7 Conclusion

In this work, we identify and mitigate a critical popularity bias in Large Language Models, where they systematically favor well-known but incorrect distractors over obscure correct answers in multiple-choice tasks. To causally demonstrate this vulnerability, To causally demonstrate this vulnerability, we developed **PopMCQ** with *six strategies for MCQs* that manipulate option popularity while holding the gold label fixed, revealing that the bias largely stems from confidence miscalibration linked to entity familiarity. We then introduced **PopDebias**, a lightweight, training-free, inference-time correction that neutralizes this effect by dividing out a label-free popularity prior from the model's probability distribution. Extensive experiments on 23 open-source models showed that PopDebias is highly effective, consistently improving accuracy (an average gain of +6.0 points) and reducing bias across four diverse QA datasets, thereby presenting a robust and practical solution to a fundamental shortcut learned by modern LLMs.

REFERENCES

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

Kenya Abe, Kunihiro Takeoka, Makoto P Kato, and Masafumi Oyamada. Llm-based query expansion fails for unfamiliar and ambiguous queries. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 3035–3039, 2025.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. The falcon series of open language models. *arXiv preprint arXiv:2311.16867*, 2023.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Gerd Gigerenzer, Peter M Todd, the ABC Research Group, et al. *Simple heuristics that make us smart*. Oxford University Press, 2000.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, et al. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36: 62991–63010, 2023.

Dongsheng Jiang, Yuchen Liu, Songlin Liu, Jin'e Zhao, Hao Zhang, Zhen Gao, Xiaopeng Zhang, Jin Li, and Hongkai Xiong. From clip to dino: Visual encoders shout in multi-modal large language models. *arXiv preprint arXiv:2310.08825*, 2023.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147/.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. In *International conference on machine learning*, pp. 15696–15707. PMLR, 2023.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. Qasc: A dataset for question answering via sentence composition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 8082–8090, 2020.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466, 2019.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.546. URL https://aclanthology.org/2023.acl-long.546/.

Jamshid Mozafari, Abdelrahman Abdallah, Bhawna Piryani, and Adam Jatowt. Wrong answers can also be useful: Plausibleqa-a large-scale qa dataset with answer plausibility scores. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 3832–3842, 2025a.

Jamshid Mozafari, Bhawna Piryani, Abdelrahman Abdallah, and Adam Jatowt. Hinteval: A comprehensive framework for hint generation and evaluation for questions. *arXiv preprint arXiv:2502.00857*, 2025b.

Shiyu Ni, Keping Bi, Jiafeng Guo, and Xueqi Cheng. How knowledge popularity influences and enhances llm knowledge boundary perception. *arXiv preprint arXiv:2505.17537*, 2025.

Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions, 2023. URL https://arxiv.org/abs/2308.11483.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2024.

Yasaman Razeghi, Robert L. Logan IV, Matt Gardner, and Sameer Singh. Impact of pretraining term frequencies on few-shot reasoning, 2022. URL https://arxiv.org/abs/2202.07206.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*, 2018.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Musique: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro Von Werra, Clémentine Fourrier, Nathan Habib, et al. Zephyr: Direct distillation of lm alignment. *arXiv preprint arXiv:2310.16944*, 2023.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. *arXiv preprint arXiv:2309.03882*, 2023.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*, 2023.

A Use of Large Language Models

Large Language Models were used solely for grammatical corrections and sentence structure improvements in the manuscript. All research contributions, methodology, experimental design, and findings are entirely the work of the human authors.

B Prompt for Candidate Answer Generation

We use the prompt design by Mozafari et al. (2025a) in Fig. 8 to synthesize a pool of *plausible but incorrect* candidate answers for each MCQ item. The instruction explicitly asks the model to ignore the ground-truth label and to produce 20 unique alternatives, each accompanied by a plausibility score and a short justification. The JSON schema standardizes parsing and downstream filtering.

Design choices and safeguards. (i) *Leakage guard:* The first line forces the generator to assume it does not know the gold answer, and we later filter out near-duplicates or paraphrases of the gold label. (ii) *Calibrated plausibility:* A numeric score (0–100) captures the generator's internal belief about how reasonable each candidate is in context; this serves as weak supervision for ranking or weighting. (iii) *Diversity:* Requiring 20 unique candidates encourages coverage of common confusions (synonyms, close entities, type-consistent distractors). (iv) *Justifications:* One-sentence rationales allow us to detect speculative or off-topic candidates and can be re-used as features.

After generation, we (a) validate JSON; (b) canonicalize strings (case-folding, punctuation/whitespace normalization); (c) drop duplicates and any candidate that matches the ground truth or its aliases; (d) optionally keep the top-m candidates by plausibility (e.g., $m \in \{5, 10\}$) or apply a minimum score threshold (e.g., ≥ 40); and (e) log failures for re-generation. The resulting candidate set provides (1) hard distractors for evaluation and (2) weak labels (scores) for reweighting or ranking objectives in our debiasing pipeline.

Synthetic training sample

Assume you are unaware that the answer to "`<question>`" is "`<ground_truth>`".

Generate 20 unique candidate answers, ensuring "`<ground_truth>`" is excluded.

For each, provide a plausibility score (0–100) evaluating how reasonable it is in context, and a brief justification.

Output as JSON:

```
[
    {"Candidate": "<answer>", "Plausibility": <score>, "Justification": "<reason>"}
]
```

Figure 8: The prompt used for generating candidate answers, with placeholders for the question and ground-truth answer.

C Prompts Used in Prompting Ablations

We compare three prompting conditions for MCQ inference: **P1** (standard MCQ), **P2** (explicit bias warning that instructs the model to ignore popularity), and **P3** (lightweight chain-of-thought that scaffolds option-by-option deliberation). Figure 9 shows the exact system messages and templates.

Design intent. P1 establishes a strong baseline under the default instruction. P2 tests whether a minimal, targeted instruction can attenuate "fame" selection bias without changing the task format. P3 examines whether structured deliberation reduces anchoring on well-known entities by forcing evidence-based comparison across options.

Execution details. Unless otherwise noted, we use deterministic decoding (temperature $= 0$) and reset context between items. Options and wording are identical across prompting conditions; only the instruction template varies. We enforce a single-token final answer (`A/B/C/D`) by appending `Answer:` and post-processing non-conforming outputs via a simple regex. No external tools, retrieval, or few-shot demonstrations are used in these ablations.

Evaluation. We score only the final choice and disregard any intermediate rationale text (for P3). For bias diagnostics, we compute PopGap, RPS, and HPSR on the selected option (Section 3.2). We report accuracy and bias metrics per prompt.

P1: Standard MCQ Prompt

System instruction:

```
You are a helpful Assistant who is designed to answer MCQ questions.
```

Template:

```
Question: {question}
Options:
A. {option_A}
B. {option_B}
C. {option_C}
D. {option_D}
Answer:
```

P2: Bias Warning Prompt

System instruction:

```
You are a helpful Assistant who is designed to answer MCQ questions.
Important: Choose based on factual accuracy, not how common the entities are.
Popular options may be incorrect distractors.
```

Template:

```
Question: {question}
Options:
A. {option_A}
B. {option_B}
C. {option_C}
D. {option_D}
Think carefully about factual accuracy rather than popularity when choosing your answer.
Answer:
```

P3: Chain-of-Thought Prompt

System instruction:

```
You are a helpful Assistant who is designed to answer MCQ questions.
Think step by step through each option before choosing your answer.
```

Template:

```
Question: {question}
Options:
A. {option_A}
B. {option_B}
C. {option_C}
D. {option_D}
Let's think through each option carefully:
1. Consider the factual accuracy of each option
2. Avoid choosing based on how well-known the entities are
3. Select the most factually correct answer
Answer:
```

Figure 9: Prompts used for standard prompting, explicit bias warning, and chain-of-thought (CoT).

D STEP-BY-STEP NUMERICAL EXAMPLE OF POPDEBIAS

Consider a 4-option item (correct answer is **C**):

Option	A	B	**C (truth)**	D
Popularity $\mathrm{Pop}(o_i)$	0.92	0.60	**0.35**	0.40
Base prob. p_i	0.40	0.25	**0.20**	0.15

The base model (argmax) selects A (popular distractor). We now apply POPULARITY-PRIOR.

1. **Features (label-free).**
$$\bar{\mathrm{Pop}} = \frac{0.92 + 0.60 + 0.35 + 0.40}{4} = 0.5675, \quad \mathrm{Var}(\mathrm{Pop}) = 0.0502,$$
$$\hat{\imath} = \arg\max_i p_i = \text{A}, \quad \max_i p_i = c = 0.40, \quad \mathrm{Pop}(o_i) = 0.92, \quad \max_j \mathrm{Pop}(o_j) = 0.92.$$

2. **Bias strength b (heuristic, scaled).** Using the same indicators as our implementation:
$$\underbrace{\frac{\mathrm{Pop}(o_i)}{\bar{\mathrm{Pop}}}}_{=1.621}, \quad \underbrace{\max_i p_i}_{=0.40}, \quad \underbrace{\frac{1}{\mathrm{Var}(\mathrm{Pop}) + \varepsilon}}_{=19.94},$$
their mean is 7.320. With a calibration scale $avg\_bias\_strength = 0.1$,
$$b \;=\; 0.1 \times 7.320 \;=\; 0.732.$$

3. **Popularity prior weights w_i (within-item).**
$$w_i \;=\; 1 \;-\; b \cdot \frac{\mathrm{Pop}(o_i)}{\max_j \mathrm{Pop}(o_j) + \varepsilon}.$$
Numerically:
$$w_A = 1 - 0.732 \cdot 1.000 = 0.268, \quad w_B = 1 - 0.732 \cdot 0.652 = 0.523,$$
$$w_C = 1 - 0.732 \cdot 0.380 = 0.722, \quad w_D = 1 - 0.732 \cdot 0.435 = 0.682.$$

4. **Adaptive confidence tempering.** We use the adaptive schedule $\gamma = \mathrm{clip}\big(1 - \beta\, b\, c\, \psi,\, 0,\, 1\big)$ with $\beta = 1$ and ψ computed from Eq. 6 (the model picked the most popular option):
$$\gamma \;=\; 1 - (1) \cdot (0.732) \cdot (0.40) \cdot (1) \;=\; 0.707.$$
The uniform share is $\frac{1-\gamma}{K} = \frac{0.293}{4} = 0.0733$.

5. **Single-pass reweighting and normalization.** Form the tempered mixture and apply w_i:
$$\widetilde{p}_i^{(\mathrm{unnorm})} \;=\; \big(\gamma\, p_i + \tfrac{1-\gamma}{K}\big)\, w_i.$$
Numbers (before normalization):
$$A : (0.707 \cdot 0.40 + 0.0733) \cdot 0.268 \;=\; 0.0954,$$
$$B : (0.707 \cdot 0.25 + 0.0733) \cdot 0.523 \;=\; 0.1308,$$
$$C : (0.707 \cdot 0.20 + 0.0733) \cdot 0.722 \;=\; 0.1550,$$
$$D : (0.707 \cdot 0.15 + 0.0733) \cdot 0.682 \;=\; 0.1223.$$
Normalize ($Z = \sum_i \widetilde{p}_i^{(\mathrm{unnorm})} = 0.5035$):
$$\widetilde{\mathbf{p}} = (0.189,\; 0.260,\; \mathbf{0.308},\; 0.243),$$
so the argmax flips from A to **C** (the true answer).

6. **Effect on bias metrics (diagnostic).**
   - *PopGap* (selected − truth): before $0.92 - 0.35 = 0.57$; after $0.35 - 0.35 = 0.00$.
   - *RPS* $\big((\mathrm{Pop}_{\text{sel}} - \mathrm{Pop}_{\text{truth}})/\max\big)$: before $0.57/0.92 = 0.620$; after 0.
   - *HPSR*: before = high-pop (A, rank 1); after = low-pop (C, rank 4).
   - Confidence (max prob.): $0.40 \rightarrow 0.308$ (adaptive tempering when bias is high).

This instance exhibits the classic "popular trap": high confidence on the most common distractor with low variance in option popularity. The within-item prior w_i downweights fame proportionally, and the adaptive γ tempers overconfidence, allowing the model to recover the less-popular but correct option.

MuSiQue: S1 Baseline Control
Models: 23, Avg correlation: -0.150

Model	Accuracy (%)	HPSR (%)	PopGap	Corr	Confidence	Alignment
Qwen2.5-7B	49.5	51.2	+0.070	-0.230	0.554	0.573
Llama-3.2-1B	35.1	51.5	+0.078	-0.219	0.400	0.550
Llama-3.2-3B	42.0	50.4	+0.069	-0.212	0.452	0.554
Qwen2.5-3B	45.0	48.8	+0.066	-0.190	0.484	0.567
Meta-Llama-3-8B-Inst	49.6	48.9	+0.053	-0.186	0.750	0.558
gemma-2-2b	33.6	48.1	+0.060	-0.174	0.569	0.494
phi-1_5	30.6	49.0	+0.081	-0.171	0.400	0.555
zephyr-7b-beta	44.0	48.0	+0.057	-0.167	0.761	0.525
Qwen2.5-1.5B	42.3	48.5	+0.064	-0.167	0.471	0.557
gemma-3-4b	28.6	48.6	+0.063	-0.155	0.792	0.385
Llama-2-13b	41.2	48.0	+0.059	-0.145	0.436	0.550
Qwen2.5-0.5B	35.6	46.0	+0.055	-0.138	0.423	0.549
gemma-3-1b	26.2	48.5	+0.063	-0.136	0.643	0.433
gemma-3-12b	27.3	48.0	+0.069	-0.134	0.715	0.407
gemma-2-27b	39.8	46.0	+0.040	-0.132	0.777	0.463
Llama-2-7b	31.8	47.1	+0.052	-0.131	0.355	0.560
gemma-2-9b	32.1	45.5	+0.049	-0.118	0.799	0.403
Mistral-7B-Inst-v0.3	47.2	46.3	+0.037	-0.117	0.707	0.554
Mistral-7B-Inst-v0.2	42.8	45.8	+0.043	-0.117	0.880	0.477
Llama-3.1-8B-Inst	48.8	45.1	+0.034	-0.115	0.588	0.580
falcon-7b-Inst	25.4	46.5	+0.053	-0.114	0.395	0.554
phi-4	53.0	46.5	+0.035	-0.111	0.700	0.591
phi-2	42.1	43.9	+0.022	-0.081	0.505	0.539

(a) MuSiQue: S1 Baseline Control results across all models

MuSiQue: S2 Popular Trap
Models: 23, Avg correlation: -0.569

Model	Accuracy (%)	HPSR (%)	PopGap	Corr	Confidence	Alignment
Qwen2.5-7B	51.0	36.2	+0.174	-0.639	0.567	0.573
Qwen2.5-3B	47.4	38.6	+0.189	-0.629	0.489	0.564
Meta-Llama-3-8B-Inst	51.0	35.5	+0.158	-0.625	0.762	0.570
phi-4	55.8	32.8	+0.139	-0.616	0.708	0.601
Llama-3.1-8B-Inst	50.6	35.5	+0.162	-0.615	0.601	0.583
zephyr-7b-beta	47.7	37.9	+0.182	-0.615	0.768	0.549
Llama-3.2-3B	43.1	41.0	+0.207	-0.613	0.459	0.549
Qwen2.5-1.5B	44.3	39.8	+0.198	-0.611	0.480	0.553
Llama-2-13b	41.5	41.1	+0.206	-0.597	0.434	0.550
Mistral-7B-Inst-v0.3	48.8	35.1	+0.172	-0.588	0.712	0.561
Qwen2.5-0.5B	36.0	44.6	+0.226	-0.580	0.430	0.548
phi-2	46.0	37.3	+0.165	-0.575	0.524	0.556
Llama-3.2-1B	35.9	44.5	+0.234	-0.573	0.396	0.549
Mistral-7B-Inst-v0.2	47.7	34.8	+0.171	-0.569	0.883	0.515
gemma-2-27b	41.1	38.5	+0.193	-0.561	0.775	0.476
phi-1_5	32.4	46.0	+0.245	-0.555	0.406	0.548
gemma-2-2b	33.0	45.0	+0.233	-0.542	0.566	0.491
Llama-2-7b	29.1	47.0	+0.246	-0.516	0.354	0.569
gemma-3-12b	26.9	49.8	+0.260	-0.512	0.713	0.402
gemma-2-9b	31.0	44.4	+0.233	-0.502	0.793	0.396
falcon-7b-Inst	26.1	47.4	+0.253	-0.492	0.395	0.551
gemma-3-1b	25.5	48.7	+0.257	-0.490	0.637	0.427
gemma-3-4b	26.0	46.1	+0.251	-0.481	0.789	0.363

(b) MuSiQue: S2 Popular Trap results across all models

Table 6: MuSiQue — S1 Baseline Control (left) and S2 Popular Trap (right)

MuSiQue: S3 Popularity Gradient
Models: 23, Avg correlation: -0.593

Model	Accuracy (%)	HPSR (%)	PopGap	Corr	Confidence	Alignment
Qwen2.5-7B	52.3	36.5	+0.175	-0.654	0.568	0.572
phi-4	55.8	35.5	+0.145	-0.641	0.717	0.606
Llama-3.1-8B-Inst	51.0	39.8	+0.173	-0.639	0.597	0.585
Qwen2.5-3B	48.4	37.9	+0.186	-0.637	0.492	0.565
Meta-Llama-3-8B-Inst	51.2	35.4	+0.163	-0.637	0.762	0.571
gemma-2-27b	42.1	48.0	+0.231	-0.628	0.779	0.482
phi-2	47.5	40.8	+0.176	-0.612	0.525	0.559
Mistral-7B-Inst-v0.3	49.9	37.1	+0.169	-0.612	0.709	0.570
Llama-3.2-3B	43.7	40.3	+0.202	-0.608	0.460	0.548
zephyr-7b-beta	47.7	34.5	+0.170	-0.602	0.767	0.549
Llama-2-13b	40.6	46.2	+0.219	-0.599	0.432	0.549
Qwen2.5-1.5B	43.0	40.5	+0.209	-0.598	0.482	0.555
gemma-2-9b	33.9	55.1	+0.272	-0.596	0.801	0.412
Llama-3.2-1B	36.1	45.9	+0.240	-0.593	0.397	0.546
Mistral-7B-Inst-v0.2	45.7	38.9	+0.177	-0.593	0.882	0.507
gemma-3-4b	27.8	62.9	+0.317	-0.575	0.793	0.373
Llama-2-7b	30.6	58.3	+0.279	-0.573	0.354	0.561
phi-1_5	29.8	56.8	+0.285	-0.566	0.404	0.552
gemma-2-2b	31.1	50.2	+0.257	-0.562	0.556	0.495
Qwen2.5-0.5B	37.6	32.5	+0.196	-0.551	0.427	0.545
gemma-3-1b	25.2	59.2	+0.303	-0.536	0.644	0.430
falcon-7b-Inst	25.1	64.3	+0.312	-0.536	0.395	0.554
gemma-3-12b	25.6	49.1	+0.249	-0.484	0.719	0.391

(a) MuSiQue: S3 Popularity Gradient results across all models

MuSiQue: S4 Direct Contest
Models: 23, Avg correlation: -0.335

Model	Accuracy (%)	HPSR (%)	PopGap	Corr	Confidence	Alignment
Qwen2.5-7B	49.9	45.8	+0.100	-0.423	0.556	0.572
Llama-2-13b	41.2	47.2	+0.104	-0.392	0.432	0.543
Meta-Llama-3-8B-Inst	48.5	42.8	+0.080	-0.377	0.747	0.566
Qwen2.5-3B	46.4	44.8	+0.098	-0.374	0.486	0.562
Llama-3.2-3B	42.1	45.8	+0.108	-0.364	0.457	0.551
Mistral-7B-Inst-v0.3	47.0	43.9	+0.086	-0.363	0.708	0.553
Llama-3.2-1B	35.9	48.2	+0.124	-0.360	0.397	0.549
Qwen2.5-1.5B	42.1	45.2	+0.098	-0.354	0.472	0.553
zephyr-7b-beta	45.4	44.1	+0.087	-0.352	0.766	0.525
gemma-2-2b	31.9	46.7	+0.117	-0.350	0.566	0.494
phi-4	52.8	41.0	+0.072	-0.343	0.706	0.597
Qwen2.5-0.5B	34.9	46.0	+0.120	-0.341	0.424	0.550
Llama-3.1-8B-Inst	48.9	41.4	+0.069	-0.332	0.590	0.578
phi-1_5	30.6	47.9	+0.125	-0.327	0.403	0.549
Mistral-7B-Inst-v0.2	43.0	42.8	+0.076	-0.322	0.875	0.479
gemma-2-27b	39.5	43.4	+0.092	-0.316	0.777	0.468
gemma-3-12b	26.4	48.1	+0.131	-0.313	0.715	0.395
Llama-2-7b	30.1	46.6	+0.108	-0.298	0.355	0.565
gemma-3-4b	28.1	46.0	+0.113	-0.298	0.788	0.383
gemma-3-1b	24.7	47.1	+0.129	-0.292	0.641	0.429
falcon-7b-Inst	25.6	47.1	+0.133	-0.282	0.395	0.553
gemma-2-9b	30.9	43.3	+0.103	-0.277	0.796	0.388
phi-2	43.0	39.0	+0.061	-0.265	0.507	0.547

(b) MuSiQue: S4 Direct Contest results across all models

Table 7: MuSiQue — S3 Popularity Gradient (left) and S4 Direct Contest (right)

E Results and Model Trends

How to read the tables. Each table reports **Accuracy**, **High-Popularity Selection Rate** (HPSR), **PopGap** (selected–truth popularity difference), **Correlation** (rank correlation between correctness and relative popularity surplus), mean **Confidence** of the chosen option, and the calibration proxy **Alignment**. The caption header lists the number of models and the dataset-level mean of Correlation for that strategy. Larger Accuracy/Alignment is better; more negative Correlation indicates stronger popularity bias; PopGap> 0 means systematic overshooting toward popular options; HPSR close to 50% is neutral under four-option items.

E.1 MuSiQue (Tables 6a–8b)

Table 6a (S1 Baseline; mean Correlation -0.150). Baseline bias is mild but present: most models show small positive PopGap ($\sim$0.03–0.08). `phi-4` attains the top accuracy (53.0%), while strong 7–8B models (`Llama-3.1-8B`, `Qwen2.5-7B`, `Meta-Llama-3-8B`) cluster around 49–50%. HPSR hovers near 50% for capable models, indicating no pronounced "always-pick-fame" behavior in the baseline.

Table 6b (S2 Popular Trap; mean -0.569). Pressure toward common distractors sharply increases bias: all models show PopGap> 0 (often 0.16–0.26) and Correlation near -0.6. Accuracies remain moderate for stronger models (e.g., `phi-4` 55.8%, `Llama-3.1-8B` 50.6%) but drop together with alignment for smaller ones.

MuSiQue: S5 Reverse Control
Models: 23, Avg correlation: 0.123

Model	Accuracy (%)	HPSR (%)	PopGap	Corr	Confidence	Alignment
Llama-3.2-3B	44.7	49.7	-0.045	0.070	0.449	0.562
Llama-3.2-1B	37.3	51.2	-0.043	0.073	0.394	0.555
Qwen2.5-7B	52.4	49.6	-0.030	0.091	0.543	0.591
gemma-3-12b	25.1	45.8	-0.075	0.104	0.717	0.395
Qwen2.5-1.5B	42.2	48.4	-0.048	0.105	0.463	0.572
gemma-2-2b	31.5	48.0	-0.068	0.109	0.560	0.500
phi-1_5	33.4	46.7	-0.064	0.115	0.399	0.548
gemma-3-1b	24.8	45.4	-0.073	0.115	0.640	0.430
Llama-2-13b	41.3	48.0	-0.057	0.116	0.426	0.552
Qwen2.5-3B	48.3	50.0	-0.043	0.120	0.475	0.581
Mistral-7B-Inst-v0.3	46.4	49.0	-0.059	0.124	0.691	0.569
Qwen2.5-0.5B	35.0	46.5	-0.059	0.125	0.419	0.554
gemma-3-4b	27.9	44.5	-0.077	0.127	0.783	0.380
Llama-2-7b	29.8	45.6	-0.072	0.127	0.353	0.565
Meta-Llama-3-8B-Inst	50.9	48.6	-0.043	0.128	0.735	0.584
Mistral-7B-Inst-v0.2	43.9	47.5	-0.067	0.135	0.868	0.489
falcon-7b-Inst	26.6	43.0	-0.081	0.137	0.395	0.552
gemma-2-27b	37.6	46.7	-0.065	0.138	0.761	0.450
phi-4	53.3	48.2	-0.043	0.138	0.692	0.607
zephyr-7b-beta	45.1	48.0	-0.059	0.139	0.749	0.547
gemma-2-9b	30.1	45.0	-0.073	0.145	0.795	0.393
Llama-3.1-8B-Inst	48.8	46.4	-0.051	0.151	0.579	0.587
phi-2	39.7	43.5	-0.075	0.188	0.488	0.548

(a) MuSiQue: S5 Reverse Control results across all models

MuSiQue: S6 None of the Above
Models: 23, Avg correlation: -0.668

Model	Accuracy (%)	HPSR (%)	PopGap	Corr	Confidence	Alignment
gemma-2-2b	60.3	25.0	+0.133	-0.934	0.570	0.512
gemma-2-27b	56.1	26.9	+0.149	-0.916	0.754	0.543
Meta-Llama-3-8B-Inst	39.6	36.9	+0.199	-0.901	0.670	0.453
gemma-3-12b	29.5	43.8	+0.236	-0.857	0.722	0.426
Mistral-7B-Inst-v0.3	22.8	47.4	+0.241	-0.833	0.669	0.384
Mistral-7B-Inst-v0.2	22.3	46.8	+0.228	-0.825	0.869	0.293
phi-4	22.0	47.1	+0.241	-0.813	0.617	0.435
gemma-3-1b	19.6	50.7	+0.278	-0.809	0.593	0.401
phi-1_5	20.7	51.7	+0.286	-0.807	0.388	0.552
Llama-3.1-8B-Inst	17.9	51.0	+0.262	-0.805	0.501	0.491
zephyr-7b-beta	13.8	54.5	+0.287	-0.760	0.721	0.318
Llama-2-13b	11.4	56.1	+0.292	-0.752	0.397	0.575
Qwen2.5-3B	11.6	56.0	+0.305	-0.739	0.415	0.555
Llama-2-7b	10.3	57.2	+0.283	-0.718	0.326	0.634
gemma-2-9b	10.1	53.6	+0.290	-0.710	0.760	0.265
Qwen2.5-7B	8.2	58.4	+0.325	-0.703	0.501	0.492
gemma-3-4b	9.8	55.8	+0.315	-0.703	0.739	0.278
phi-2	8.9	52.9	+0.238	-0.693	0.460	0.520
Qwen2.5-1.5B	2.1	61.5	+0.331	-0.424	0.448	0.546
Qwen2.5-0.5B	1.9	58.1	+0.300	-0.391	0.411	0.583
Llama-3.2-3B	0.2	63.5	+0.350	-0.145	0.437	0.562
falcon-7b-Inst	0.1	62.2	+0.330	-0.126	0.389	0.610
Llama-3.2-1B	0.0	65.2	+0.382	0.000	0.444	0.556

(b) MuSiQue: S6 None of the Above results across all models

Table 8: MuSiQue — S5 Reverse Control (left) and S6 None of the Above (right)

NQ: S1 Baseline Control
Models: 25, Avg correlation: -0.166

Model	Accuracy (%)	HPSR (%)	PopGap	Corr	Confidence	Alignment
Qwen2.5-7B	55.9	49.7	+0.064	-0.250	0.682	0.620
Llama-3.2-1B	36.5	52.2	+0.097	-0.237	0.416	0.558
Qwen2.5-3B	51.2	49.5	+0.066	-0.213	0.593	0.601
Llama-3.2-3B	48.5	48.3	+0.057	-0.211	0.527	0.572
zephyr-7b-beta	53.5	48.4	+0.054	-0.206	0.858	0.582
Mistral-7B-Inst-v0.3	55.5	46.5	+0.048	-0.184	0.835	0.603
Qwen2.5-0.5B	38.3	49.1	+0.080	-0.182	0.461	0.547
Llama-2-13b	49.0	47.0	+0.050	-0.181	0.471	0.563
phi-4	55.4	48.0	+0.054	-0.181	0.784	0.614
Qwen2.5-1.5B	46.2	48.1	+0.063	-0.180	0.574	0.581
Llama-3.1-8B-Inst	58.5	47.4	+0.037	-0.178	0.702	0.628
Mistral-7B-Inst-v0.2	53.5	46.9	+0.043	-0.171	0.935	0.558
phi-1_5	31.4	48.1	+0.074	-0.165	0.408	0.550
gemma-2-2b	33.1	46.5	+0.072	-0.154	0.570	0.496
Meta-Llama-3-8B-Inst	56.0	45.9	+0.034	-0.153	0.812	0.620
phi-2	42.9	46.6	+0.053	-0.152	0.530	0.551
gemma-3-4b	28.1	46.2	+0.066	-0.143	0.776	0.381
gemma-3-12b	27.9	45.5	+0.067	-0.139	0.734	0.399
gemma-2-9b	34.8	45.8	+0.064	-0.135	0.832	0.420
Llama-2-7b	31.5	45.2	+0.053	-0.128	0.433	0.536
gemma-3-1b	24.1	44.9	+0.060	-0.111	0.665	0.409
gemma-2-27b	40.0	43.5	+0.040	-0.097	0.850	0.451
falcon-7b-Inst	24.1	43.0	+0.056	-0.083	0.387	0.561

(a) NQ: S1 Baseline Control results across all models

NQ: S2 Popular Trap
Models: 23, Avg correlation: -0.584

Model	Accuracy (%)	HPSR (%)	PopGap	Corr	Confidence	Alignment
Qwen2.5-7B	57.0	29.9	+0.146	-0.658	0.690	0.627
phi-4	59.0	28.6	+0.128	-0.646	0.790	0.632
Mistral-7B-Inst-v0.3	56.8	29.5	+0.144	-0.645	0.832	0.617
zephyr-7b-beta	55.5	30.9	+0.147	-0.642	0.859	0.590
Llama-3.1-8B-Inst	61.2	26.5	+0.115	-0.641	0.708	0.646
Qwen2.5-1.5B	47.0	36.2	+0.189	-0.632	0.587	0.575
Mistral-7B-Inst-v0.2	56.4	28.7	+0.135	-0.626	0.934	0.587
Meta-Llama-3-8B-Inst	58.5	27.7	+0.118	-0.626	0.820	0.633
Qwen2.5-3B	53.4	30.9	+0.157	-0.625	0.600	0.603
phi-2	45.5	37.0	+0.187	-0.624	0.548	0.559
Llama-3.2-3B	48.4	33.7	+0.168	-0.609	0.532	0.575
Llama-2-13b	48.5	33.2	+0.160	-0.601	0.468	0.562
Llama-3.2-1B	34.4	43.5	+0.237	-0.578	0.419	0.552
gemma-2-27b	39.7	38.7	+0.199	-0.577	0.856	0.448
Qwen2.5-0.5B	34.4	42.7	+0.232	-0.566	0.465	0.556
gemma-2-2b	33.5	42.1	+0.230	-0.550	0.569	0.500
gemma-2-9b	33.6	42.8	+0.226	-0.549	0.831	0.406
phi-1_5	31.9	43.8	+0.230	-0.546	0.408	0.551
Llama-2-7b	29.8	44.8	+0.233	-0.528	0.436	0.538
gemma-3-12b	27.5	45.9	+0.245	-0.517	0.742	0.398
falcon-7b-Inst	26.6	46.2	+0.243	-0.492	0.387	0.554
gemma-3-4b	24.5	47.4	+0.257	-0.485	0.782	0.356
gemma-3-1b	22.9	48.9	+0.264	-0.478	0.666	0.409

(b) NQ: S2 Popular Trap results across all models

Table 9: NQ — S1 Baseline Control (left) and S2 Popular Trap (right)

Table 7a (S3 Popularity Gradient; mean -0.593). Ordering candidates by popularity further amplifies the pattern: high-capacity models retain $\sim$50–56% accuracy yet exhibit correlations ≈ -0.64, while small models both over-select popular options (HPSR $> 55\%$) and lose accuracy.

Table 7b (S4 Direct Contest; mean -0.335). Directly pitting truth vs. a popular rival yields intermediate bias: POPGAP remains positive but smaller than S2/S3; Correlation weakens to ~ -0.33. Accuracy spreads mirror S1/S2, showing that a single common foil is already sufficient to tilt choices for many models.

Table 8a (S5 Reverse Control; mean $+0.123$). When popularity aligns with truth, the sign flips: Correlation becomes positive for *all* models, and POPGAP turns slightly negative. Accuracy often ticks up relative to S1 (e.g., `Qwen2.5-7B` 52.4%), confirming that models can leverage popularity signal *when it helps*.

Table 8b (S6 None of the Above; mean -0.668). The strongest bias under abstention: correlations are extremely negative (down to -0.93) and POPGAP is large and positive. Notably, some models still post high accuracies here (e.g., `gemma-2-2b-it` 60.3%), yet *still* display very negative correlation—evidence that, conditional on popularity differences, errors are tightly coupled to picking the common option.

NQ: S3 Popularity Gradient
Models: 23, Avg correlation: -0.608

Model	Accuracy (%)	HPSR (%)	PopGap	Corr	Confidence	Alignment
Qwen2.5-7B	56.7	31.9	+0.157	-0.684	0.689	0.627
phi-4	59.1	30.3	+0.131	-0.666	0.790	0.629
Llama-3.1-8B-Inst	61.2	28.9	+0.125	-0.654	0.710	0.647
Llama-2-13b	47.5	38.8	+0.181	-0.653	0.469	0.563
Qwen2.5-3B	53.2	31.7	+0.161	-0.649	0.602	0.605
Mistral-7B-Inst-v0.3	57.7	29.8	+0.141	-0.648	0.845	0.610
gemma-2-27b	42.0	44.9	+0.218	-0.642	0.851	0.465
Meta-Llama-3-8B-Inst	60.2	27.5	+0.125	-0.642	0.823	0.640
Llama-3.2-3B	49.9	34.8	+0.166	-0.637	0.530	0.568
Mistral-7B-Inst-v0.2	56.7	30.2	+0.140	-0.635	0.938	0.590
Qwen2.5-1.5B	46.6	35.9	+0.188	-0.627	0.588	0.574
zephyr-7b-beta	55.5	30.4	+0.145	-0.621	0.867	0.589
phi-2	44.2	38.0	+0.186	-0.604	0.549	0.559
Llama-3.2-1B	37.0	44.1	+0.234	-0.600	0.418	0.548
gemma-2-9b	33.9	50.1	+0.245	-0.597	0.836	0.405
gemma-2-2b	33.1	48.0	+0.248	-0.578	0.567	0.503
Llama-2-7b	29.9	56.1	+0.268	-0.578	0.437	0.539
phi-1_5	33.1	49.0	+0.243	-0.575	0.408	0.544
Qwen2.5-0.5B	36.4	37.4	+0.220	-0.570	0.468	0.550
gemma-3-4b	26.1	57.4	+0.290	-0.547	0.785	0.359
gemma-3-1b	25.6	55.1	+0.289	-0.529	0.670	0.415
falcon-7b-Inst	23.8	59.1	+0.284	-0.520	0.387	0.561
gemma-3-12b	27.9	46.9	+0.245	-0.517	0.737	0.396

(a) NQ: S3 Popularity Gradient results across all models

NQ: S4 Direct Contest
Models: 23, Avg correlation: -0.359

Model	Accuracy (%)	HPSR (%)	PopGap	Corr	Confidence	Alignment
Qwen2.5-7B	56.1	41.6	+0.096	-0.457	0.682	0.621
Llama-3.1-8B-Inst	57.6	39.3	+0.076	-0.426	0.704	0.628
phi-4	56.1	39.5	+0.087	-0.422	0.780	0.617
Mistral-7B-Inst-v0.3	57.0	39.4	+0.077	-0.420	0.836	0.604
Qwen2.5-3B	50.7	42.0	+0.100	-0.418	0.591	0.602
zephyr-7b-beta	52.5	40.7	+0.089	-0.409	0.859	0.571
Meta-Llama-3-8B-Inst	56.1	38.5	+0.072	-0.381	0.820	0.611
Llama-2-13b	46.9	41.1	+0.092	-0.377	0.466	0.561
Llama-3.2-1B	35.0	46.4	+0.144	-0.374	0.419	0.555
Mistral-7B-Inst-v0.2	53.9	38.2	+0.076	-0.371	0.931	0.562
Llama-3.2-3B	47.5	39.8	+0.092	-0.371	0.533	0.565
Qwen2.5-1.5B	46.1	40.5	+0.105	-0.356	0.578	0.568
Qwen2.5-0.5B	35.6	43.6	+0.130	-0.354	0.463	0.557
gemma-2-2b	31.6	45.6	+0.132	-0.345	0.573	0.491
gemma-2-27b	39.8	41.5	+0.096	-0.335	0.850	0.448
phi-1_5	31.1	45.4	+0.132	-0.332	0.409	0.548
gemma-3-12b	27.3	45.0	+0.135	-0.321	0.743	0.390
phi-2	42.3	40.2	+0.099	-0.318	0.538	0.544
Llama-2-7b	29.6	44.1	+0.122	-0.305	0.433	0.544
gemma-2-9b	32.6	42.1	+0.112	-0.305	0.831	0.403
falcon-7b-Inst	25.1	46.2	+0.143	-0.303	0.387	0.556
gemma-3-1b	24.6	44.6	+0.128	-0.284	0.664	0.412
gemma-3-4b	27.0	44.0	+0.127	-0.284	0.778	0.376

(b) NQ: S4 Direct Contest results across all models

Table 10: NQ — S3 Popularity Gradient (left) and S4 Direct Contest (right)

NQ: S5 Reverse Control
Models: 23, Avg correlation: 0.124

Model	Accuracy (%)	HPSR (%)	PopGap	Corr	Confidence	Alignment
Qwen2.5-7B	60.5	51.1	-0.008	0.053	0.684	0.655
Llama-3.2-1B	39.0	51.6	-0.025	0.063	0.414	0.567
Qwen2.5-1.5B	47.1	49.5	-0.020	0.077	0.571	0.596
Qwen2.5-3B	55.0	49.8	-0.019	0.079	0.590	0.626
Qwen2.5-0.5B	36.0	50.0	-0.028	0.083	0.458	0.572
Llama-3.2-3B	51.5	49.1	-0.029	0.104	0.525	0.592
phi-4	60.5	49.8	-0.024	0.107	0.784	0.650
Llama-3.1-8B-Inst	62.2	48.2	-0.028	0.121	0.695	0.661
gemma-3-12b	28.6	46.4	-0.057	0.124	0.741	0.401
zephyr-7b-beta	56.3	49.0	-0.031	0.124	0.850	0.605
phi-1_5	31.2	45.8	-0.052	0.128	0.405	0.557
gemma-2-2b	33.1	46.7	-0.050	0.129	0.567	0.495
phi-2	43.2	48.0	-0.034	0.132	0.520	0.560
Mistral-7B-Inst-v0.3	57.9	48.6	-0.034	0.134	0.827	0.624
Mistral-7B-Inst-v0.2	55.0	48.0	-0.036	0.142	0.931	0.575
gemma-2-9b	35.6	45.3	-0.050	0.144	0.834	0.425
gemma-3-4b	26.6	44.9	-0.067	0.146	0.774	0.381
gemma-3-1b	24.9	43.5	-0.070	0.149	0.665	0.424
Llama-2-7b	29.9	44.1	-0.066	0.156	0.430	0.541
Meta-Llama-3-8B-Inst	61.2	48.4	-0.029	0.156	0.808	0.658
falcon-7b-Inst	25.6	42.4	-0.069	0.157	0.385	0.558
gemma-2-27b	39.5	44.5	-0.059	0.176	0.844	0.449
Llama-2-13b	50.2	46.7	-0.046	0.179	0.460	0.567

(a) NQ: S5 Reverse Control results across all models

NQ: S6 None of the Above
Models: 23, Avg correlation: -0.668

Model	Accuracy (%)	HPSR (%)	PopGap	Corr	Confidence	Alignment
gemma-2-2b	56.8	28.6	+0.184	-0.929	0.548	0.495
phi-4	26.8	49.6	+0.316	-0.871	0.727	0.410
gemma-3-12b	26.1	49.0	+0.311	-0.864	0.722	0.406
Meta-Llama-3-8B-Inst	26.6	48.6	+0.292	-0.855	0.722	0.381
phi-1_5	22.5	52.6	+0.347	-0.853	0.392	0.544
gemma-3-1b	23.4	50.0	+0.309	-0.849	0.603	0.414
gemma-2-27b	25.0	47.9	+0.294	-0.839	0.778	0.335
Mistral-7B-Inst-v0.3	18.6	54.8	+0.347	-0.833	0.801	0.294
Llama-3.1-8B-Inst	17.8	54.5	+0.337	-0.812	0.601	0.404
Llama-2-13b	12.8	59.4	+0.374	-0.789	0.404	0.562
zephyr-7b-beta	12.9	58.5	+0.370	-0.777	0.822	0.234
gemma-3-4b	12.3	56.7	+0.363	-0.768	0.715	0.307
Qwen2.5-1.5B	9.3	61.0	+0.384	-0.731	0.521	0.472
Qwen2.5-3B	9.9	62.6	+0.416	-0.729	0.512	0.477
Qwen2.5-7B	8.8	64.0	+0.410	-0.728	0.611	0.400
Mistral-7B-Inst-v0.2	8.0	62.1	+0.369	-0.689	0.928	0.129
phi-2	6.0	59.2	+0.360	-0.645	0.490	0.497
gemma-2-9b	5.0	61.2	+0.403	-0.583	0.807	0.200
Llama-2-7b	4.9	62.4	+0.386	-0.578	0.382	0.600
Qwen2.5-0.5B	1.9	62.3	+0.412	-0.420	0.464	0.531
falcon-7b-Inst	0.1	64.8	+0.388	-0.122	0.386	0.614
Llama-3.2-3B	0.1	66.3	+0.433	-0.104	0.500	0.499
Llama-3.2-1B	0.0	69.1	+0.465	0.000	0.450	0.550

(b) NQ: S6 None of the Above results across all models

Table 11: NQ — S5 Reverse Control (left) and S6 None of the Above (right)

E.2 Natural Questions (Tables 9a–11b)

Table 9a (S1; mean -0.166). Baseline bias is again mild: small positive PopGap (0.03–0.08) and correlation around -0.18. Top accuracies are in the high 50s–low 60s (e.g., `Llama-3.1-8B` 58.5%).

Table 9b (S2; mean -0.584). Popularity pressure induces strong negatives across the board (correlation ~ -0.63), with PopGap typically 0.13–0.24. Large models (`Llama-3.1-8B`, `phi-4`) keep accuracies $\geq$59–61%, while small ones degrade notably.

Table 10a (S3; mean -0.608). Gradient ordering pushes bias slightly beyond S2 (means ≈ -0.65), and HPSR rises for weaker models. Accuracy ordering among families is similar to S2.

Table 10b (S4; mean -0.359). Direct contests reduce but do not remove bias: correlation is ~ -0.30 to -0.46, with moderate positive PopGap; accuracy remains respectable for strong models.

Table 11a (S5; mean $+0.124$). All models become positively correlated; PopGap is slightly negative. Several models improve over S1 (e.g., `Llama-3.1-8B` 62.2%, `phi-4` 60.5%).

Table 11b (S6; mean -0.668). Abstention is again the worst: extremely negative correlations (to -0.93) and large PopGap (≥ 0.29). Even when accuracy is non-trivial for a few models (e.g., `gemma-2-2b-it` 56.8%), correlation remains very negative, indicating a strong fame-over-truth failure mode.

QASC: S1 Baseline Control
Models: 23, Avg correlation: -0.109

Model	Accuracy (%)	HPSR (%)	PopGap	Corr	Confidence	Alignment
phi-4	53.9	56.6	+0.047	-0.186	0.814	0.576
Llama-3.1-8B-Inst	52.0	56.6	+0.044	-0.182	0.744	0.577
Meta-Llama-3-8B-Inst	52.9	56.3	+0.046	-0.169	0.863	0.577
Llama-3.2-3B	44.6	56.2	+0.047	-0.165	0.626	0.535
Qwen2.5-7B	58.1	56.6	+0.042	-0.164	0.781	0.610
Qwen2.5-3B	53.4	56.0	+0.041	-0.160	0.701	0.584
Qwen2.5-1.5B	49.0	55.4	+0.038	-0.158	0.686	0.554
Mistral-7B-Inst-v0.3	47.0	55.4	+0.040	-0.157	0.838	0.519
zephyr-7b-beta	43.2	54.0	+0.038	-0.138	0.886	0.477
Mistral-7B-Inst-v0.2	47.6	54.8	+0.032	-0.136	0.942	0.497
Llama-3.2-1B	32.2	53.0	+0.047	-0.121	0.519	0.512
phi-2	45.0	53.0	+0.028	-0.102	0.648	0.539
Llama-2-13b	37.0	51.5	+0.039	-0.095	0.520	0.526
Llama-2-7b	30.8	50.2	+0.031	-0.094	0.365	0.562
Qwen2.5-0.5B	35.6	50.8	+0.022	-0.083	0.601	0.503
gemma-3-12b	23.9	48.9	+0.037	-0.068	0.757	0.364
falcon-7b-Inst	23.4	47.5	+0.032	-0.065	0.376	0.567
gemma-2-2b	29.6	48.5	+0.018	-0.052	0.641	0.445
phi-1_5	34.2	49.0	+0.016	-0.050	0.555	0.510
gemma-2-9b	28.9	48.2	+0.025	-0.049	0.875	0.348
gemma-3-4b	24.4	48.1	+0.026	-0.046	0.793	0.342
gemma-3-1b	24.9	47.8	+0.026	-0.041	0.669	0.411
gemma-2-27b	28.4	47.2	+0.015	-0.032	0.879	0.340

(a) QASC: S1 Baseline Control results across all models

QASC: S2 Popular Trap
Models: 23, Avg correlation: -0.556

Model	Accuracy (%)	HPSR (%)	PopGap	Corr	Confidence	Alignment
Meta-Llama-3-8B-Inst	53.1	35.4	+0.190	-0.637	0.869	0.564
phi-4	52.5	35.5	+0.190	-0.635	0.816	0.571
Qwen2.5-7B	57.9	32.1	+0.166	-0.628	0.785	0.611
Qwen2.5-3B	53.1	34.9	+0.182	-0.620	0.707	0.581
Llama-3.1-8B-Inst	53.4	34.9	+0.188	-0.619	0.753	0.578
Llama-3.2-3B	46.8	38.5	+0.213	-0.614	0.632	0.536
Qwen2.5-1.5B	49.2	36.0	+0.196	-0.613	0.690	0.561
Mistral-7B-Inst-v0.3	46.4	38.2	+0.216	-0.602	0.830	0.521
phi-2	48.6	36.2	+0.200	-0.597	0.652	0.544
Mistral-7B-Inst-v0.2	47.5	36.8	+0.204	-0.590	0.939	0.500
zephyr-7b-beta	44.4	39.2	+0.219	-0.586	0.882	0.483
phi-1_5	37.5	41.7	+0.239	-0.550	0.558	0.509
Qwen2.5-0.5B	37.0	41.6	+0.240	-0.549	0.614	0.510
Llama-3.2-1B	32.8	45.9	+0.272	-0.548	0.518	0.517
Llama-2-13b	37.0	42.6	+0.245	-0.547	0.517	0.531
Llama-2-7b	31.1	45.9	+0.259	-0.513	0.369	0.563
gemma-2-2b	28.9	46.9	+0.281	-0.500	0.634	0.447
falcon-7b-Inst	26.9	47.4	+0.274	-0.488	0.377	0.557
gemma-2-27b	29.0	45.8	+0.271	-0.487	0.875	0.348
gemma-2-9b	27.4	46.7	+0.280	-0.484	0.882	0.334
gemma-3-12b	24.9	49.4	+0.292	-0.482	0.754	0.377
gemma-3-4b	23.2	49.2	+0.298	-0.455	0.796	0.330
gemma-3-1b	22.6	49.5	+0.299	-0.449	0.663	0.399

(b) QASC: S2 Popular Trap results across all models

Table 12: QASC — S1 Baseline Control (left) and S2 Popular Trap (right)

QASC: S3 Popularity Gradient
Models: 23, Avg correlation: -0.574

Model	Accuracy (%)	HPSR (%)	PopGap	Corr	Confidence	Alignment
Meta-Llama-3-8B-Inst	51.9	35.9	+0.198	-0.640	0.865	0.564
Qwen2.5-7B	58.0	32.1	+0.167	-0.635	0.789	0.611
phi-4	53.0	35.6	+0.189	-0.628	0.816	0.570
Llama-3.1-8B-Inst	52.0	36.5	+0.195	-0.626	0.755	0.582
Qwen2.5-3B	53.5	34.5	+0.181	-0.619	0.714	0.580
Qwen2.5-1.5B	49.0	36.2	+0.198	-0.616	0.693	0.559
Llama-3.2-3B	46.9	38.5	+0.218	-0.614	0.635	0.540
phi-2	48.4	37.2	+0.202	-0.610	0.657	0.545
Mistral-7B-Inst-v0.2	48.6	37.1	+0.205	-0.602	0.944	0.507
Mistral-7B-Inst-v0.3	47.5	37.5	+0.210	-0.600	0.841	0.529
zephyr-7b-beta	45.3	37.5	+0.216	-0.578	0.888	0.489
Llama-2-13b	38.6	45.7	+0.246	-0.575	0.516	0.535
gemma-2-27b	30.4	56.1	+0.302	-0.560	0.881	0.354
phi-1_5	36.5	44.9	+0.250	-0.555	0.562	0.510
Llama-3.2-1B	33.1	44.6	+0.273	-0.552	0.515	0.524
Qwen2.5-0.5B	39.0	38.4	+0.230	-0.550	0.610	0.506
gemma-2-9b	29.1	56.5	+0.311	-0.550	0.889	0.341
Llama-2-7b	29.1	55.3	+0.295	-0.546	0.371	0.568
gemma-2-2b	30.9	51.0	+0.282	-0.542	0.639	0.450
gemma-3-4b	24.9	60.0	+0.332	-0.526	0.793	0.344
gemma-3-1b	25.8	57.1	+0.320	-0.519	0.663	0.413
falcon-7b-Inst	22.9	59.5	+0.314	-0.489	0.378	0.565
gemma-3-12b	24.5	49.7	+0.295	-0.481	0.755	0.367

(a) QASC: S3 Popularity Gradient results across all models

QASC: S4 Direct Contest
Models: 23, Avg correlation: -0.292

Model	Accuracy (%)	HPSR (%)	PopGap	Corr	Confidence	Alignment
Qwen2.5-7B	55.9	48.2	+0.096	-0.384	0.782	0.604
Qwen2.5-3B	52.8	49.5	+0.101	-0.380	0.699	0.573
Llama-3.2-3B	45.2	50.5	+0.117	-0.352	0.633	0.534
phi-4	52.5	47.8	+0.097	-0.345	0.815	0.564
Qwen2.5-1.5B	48.2	49.2	+0.105	-0.337	0.690	0.555
Meta-Llama-3-8B-Inst	52.1	48.5	+0.101	-0.336	0.866	0.565
Llama-3.1-8B-Inst	52.0	47.8	+0.100	-0.331	0.756	0.573
Mistral-7B-Inst-v0.3	44.8	49.4	+0.112	-0.328	0.833	0.508
phi-2	46.5	46.9	+0.096	-0.309	0.656	0.538
zephyr-7b-beta	42.0	49.0	+0.110	-0.309	0.882	0.470
Mistral-7B-Inst-v0.2	45.6	46.8	+0.100	-0.303	0.939	0.481
Llama-3.2-1B	32.9	49.1	+0.127	-0.285	0.519	0.516
Qwen2.5-0.5B	38.3	48.3	+0.109	-0.282	0.602	0.508
Llama-2-13b	37.5	47.9	+0.114	-0.276	0.521	0.529
Llama-2-7b	30.2	48.9	+0.128	-0.264	0.368	0.563
phi-1_5	36.9	45.4	+0.107	-0.262	0.559	0.509
gemma-2-9b	28.8	47.5	+0.123	-0.246	0.882	0.340
gemma-3-12b	23.8	49.7	+0.145	-0.244	0.760	0.368
gemma-2-2b	30.4	47.1	+0.114	-0.240	0.635	0.450
gemma-2-27b	28.8	45.8	+0.113	-0.233	0.880	0.340
gemma-3-4b	24.2	46.4	+0.122	-0.230	0.801	0.341
falcon-7b-Inst	25.0	48.8	+0.136	-0.228	0.377	0.561
gemma-3-1b	24.9	45.4	+0.121	-0.220	0.669	0.411

(b) QASC: S4 Direct Contest results across all models

Table 13: QASC — S3 Popularity Gradient (left) and S4 Direct Contest (right)

E.3 QASC (TABLES 12A–14B)

Table 12a (S1; mean -0.109). Baseline has the *weakest bias* (mean $\rho = -0.109$): accuracy low–mid 50%, POPGAP $\sim$0.02–0.05, HPSR $\approx$55

Table 12b (S2; mean -0.556). With popular distractors, bias rises to ~ -0.60 and POPGAP reaches 0.19–0.30. Accuracy remains mid-40s to high-50s for capable models, but calibration worsens for smaller ones.

Table 13a (S3; mean -0.574). Popularity ordering produces similarly strong negatives as S2; weaker families show pronounced HPSR inflation ($\geq 56\%$) and large POPGAP (≥ 0.30).

Table 13b (S4; mean -0.292). As in other datasets, direct contests attenuate bias but keep correlation negative. POPGAP remains positive ($\sim$0.10), consistent with a persistent pull toward fame.

Table 14a (S5; mean $+0.204$). The largest positive mean among datasets: all models flip sign; POPGAP is moderately negative. Strong and mid-size models gain accuracy relative to S1.

Table 14b (S6; mean -0.715). This is the most severe dataset–strategy combination on average: correlations near -0.84 for many models. Accuracies are generally low (single to low-double digits for most), HPSR high ($\geq 56\%$), and POPGAP large (≥ 0.34), reflecting over-selection of common entities when the correct action is abstention.

QASC: S5 Reverse Control
Models: 23, Avg correlation: 0.204

Model	Accuracy (%)	HPSR (%)	PopGap	Corr	Confidence	Alignment
zephyr-7b-beta	46.0	53.8	-0.089	0.181	0.883	0.506
gemma-3-4b	23.4	47.5	-0.142	0.185	0.791	0.339
phi-4	55.8	55.0	-0.071	0.186	0.814	0.601
gemma-2-2b	30.4	50.0	-0.121	0.188	0.637	0.448
Llama-2-7b	31.8	51.6	-0.123	0.188	0.365	0.564
Llama-3.2-3B	47.6	54.2	-0.088	0.188	0.619	0.552
gemma-3-1b	23.2	47.3	-0.146	0.190	0.673	0.402
Meta-Llama-3-8B-Inst	55.5	55.1	-0.070	0.192	0.865	0.595
Mistral-7B-Inst-v0.3	48.3	53.2	-0.089	0.195	0.836	0.537
gemma-3-12b	25.6	48.5	-0.133	0.197	0.751	0.374
gemma-2-9b	29.4	48.5	-0.132	0.200	0.879	0.349
Llama-3.2-1B	33.8	50.0	-0.126	0.205	0.511	0.520
gemma-2-27b	27.1	48.7	-0.140	0.206	0.881	0.329
falcon-7b-Inst	25.8	46.4	-0.137	0.206	0.376	0.561
Llama-3.1-8B-Inst	54.4	54.4	-0.074	0.206	0.745	0.602
Mistral-7B-Inst-v0.2	48.1	52.8	-0.096	0.208	0.937	0.504
phi-1_5	37.0	50.0	-0.121	0.215	0.544	0.525
Qwen2.5-1.5B	52.0	53.9	-0.082	0.216	0.683	0.575
Qwen2.5-7B	60.5	54.6	-0.066	0.223	0.782	0.636
Qwen2.5-0.5B	36.8	49.7	-0.121	0.224	0.596	0.515
phi-2	47.4	51.3	-0.099	0.225	0.640	0.561
Qwen2.5-3B	57.0	54.0	-0.072	0.230	0.698	0.609
Llama-2-13b	41.1	51.4	-0.111	0.233	0.514	0.535

(a) QASC: S5 Reverse Control results across all models

QASC: S6 None of the Above
Models: 23, Avg correlation: -0.715

Model	Accuracy (%)	HPSR (%)	PopGap	Corr	Confidence	Alignment
gemma-3-12b	37.4	41.2	+0.235	-0.925	0.773	0.467
gemma-2-2b	38.1	40.0	+0.224	-0.924	0.566	0.442
phi-1_5	22.4	50.6	+0.290	-0.861	0.507	0.459
Qwen2.5-3B	15.1	56.4	+0.343	-0.840	0.583	0.409
gemma-2-27b	16.4	53.6	+0.304	-0.839	0.834	0.244
Qwen2.5-7B	14.3	57.3	+0.362	-0.837	0.678	0.342
Meta-Llama-3-8B-Inst	16.8	54.1	+0.328	-0.830	0.780	0.285
gemma-3-1b	16.2	54.1	+0.301	-0.828	0.634	0.375
phi-4	12.2	58.7	+0.363	-0.809	0.784	0.269
Mistral-7B-Inst-v0.3	12.1	55.8	+0.342	-0.808	0.826	0.227
Llama-2-7b	13.2	56.4	+0.311	-0.802	0.340	0.611
zephyr-7b-beta	8.8	60.0	+0.354	-0.771	0.876	0.161
phi-2	8.0	58.6	+0.319	-0.734	0.595	0.393
Llama-3.1-8B-Inst	6.4	61.3	+0.363	-0.695	0.679	0.316
gemma-3-4b	6.8	60.2	+0.349	-0.690	0.711	0.298
Qwen2.5-1.5B	4.8	63.3	+0.373	-0.666	0.626	0.372
Llama-2-13b	4.8	63.3	+0.376	-0.664	0.466	0.521
Mistral-7B-Inst-v0.2	4.8	60.4	+0.359	-0.642	0.949	0.085
Qwen2.5-0.5B	5.1	62.0	+0.349	-0.640	0.579	0.413
gemma-2-9b	2.5	63.0	+0.368	-0.524	0.850	0.153
falcon-7b-Inst	2.1	62.8	+0.344	-0.456	0.362	0.631
Llama-3.2-3B	1.5	65.5	+0.386	-0.423	0.581	0.417
Llama-3.2-1B	0.3	64.0	+0.360	-0.229	0.549	0.450

(b) QASC: S6 None of the Above results across all models

Table 14: QASC — S5 Reverse Control (left) and S6 None of the Above (right)

TriviaQA: S1 Baseline Control
Models: 23, Avg Correlation: -0.090

Model	Accuracy (%)	HPSR (%)	PopGap	Corr	Confidence	Alignment
Qwen2.5-7B	60.4	55.6	+0.055	-0.187	0.787	0.646
Meta-Llama-3-8B-Inst	64.6	55.1	+0.040	-0.167	0.905	0.667
Llama-3.1-8B-Inst	64.9	55.1	+0.044	-0.165	0.813	0.678
Mistral-7B-Inst-v0.3	60.9	55.1	+0.048	-0.150	0.866	0.636
zephyr-7b-beta	59.2	54.8	+0.039	-0.148	0.896	0.611
Llama-3.2-3B	52.6	53.9	+0.052	-0.143	0.631	0.583
phi-4	66.8	55.4	+0.041	-0.135	0.872	0.687
Llama-3.2-1B	33.7	53.4	+0.076	-0.131	0.473	0.532
Llama-2-13b	52.9	53.8	+0.048	-0.130	0.570	0.569
Qwen2.5-3B	54.2	53.5	+0.047	-0.104	0.687	0.607
Qwen2.5-1.5B	44.4	52.6	+0.054	-0.095	0.644	0.563
Mistral-7B-Inst-v0.2	58.2	53.0	+0.039	-0.092	0.947	0.601
Llama-2-7b	34.2	50.5	+0.046	-0.085	0.390	0.552
gemma-3-12b	30.2	48.5	+0.041	-0.070	0.752	0.408
Qwen2.5-0.5B	29.2	48.8	+0.038	-0.054	0.494	0.533
gemma-2-27b	46.2	49.4	+0.030	-0.051	0.856	0.507
gemma-2-2b	32.7	48.4	+0.040	-0.050	0.623	0.474
gemma-3-4b	28.4	46.5	+0.037	-0.044	0.813	0.371
gemma-3-1b	24.2	46.1	+0.035	-0.039	0.688	0.395
phi-1_5	29.5	47.0	+0.033	-0.039	0.429	0.544
gemma-2-9b	35.0	47.7	+0.033	-0.031	0.850	0.405
falcon-7b-Inst	26.5	45.1	+0.023	-0.015	0.390	0.552
phi-2	42.7	45.4	+0.009	0.058	0.585	0.537

(a) TriviaQA: S1 Baseline Control results across all models

TriviaQA: S2 Popular Trap
Models: 23, Avg correlation: -0.531

Model	Accuracy (%)	HPSR (%)	PopGap	Corr	Confidence	Alignment
phi-4	65.2	27.3	+0.122	-0.633	0.873	0.680
Llama-3.1-8B-Inst	65.1	25.3	+0.108	-0.602	0.821	0.678
Meta-Llama-3-8B-Inst	64.2	25.2	+0.106	-0.593	0.910	0.667
Mistral-7B-Inst-v0.3	59.5	28.8	+0.127	-0.589	0.867	0.628
Llama-3.2-3B	50.3	33.7	+0.159	-0.587	0.634	0.579
Mistral-7B-Inst-v0.2	58.7	28.4	+0.126	-0.582	0.956	0.605
zephyr-7b-beta	57.9	29.1	+0.128	-0.580	0.897	0.607
Qwen2.5-7B	60.3	27.5	+0.124	-0.577	0.795	0.647
Llama-2-13b	52.6	31.8	+0.149	-0.568	0.574	0.563
Qwen2.5-3B	54.8	29.9	+0.140	-0.559	0.695	0.602
Qwen2.5-1.5B	44.8	35.3	+0.176	-0.540	0.646	0.561
phi-2	49.6	31.2	+0.145	-0.522	0.602	0.547
gemma-2-27b	46.0	33.1	+0.155	-0.521	0.862	0.502
gemma-2-9b	36.0	39.8	+0.201	-0.509	0.859	0.411
Llama-3.2-1B	31.9	43.5	+0.220	-0.495	0.469	0.538
gemma-2-2b	32.8	41.9	+0.209	-0.493	0.624	0.479
phi-1_5	31.6	41.9	+0.220	-0.483	0.441	0.535
Llama-2-7b	32.0	41.3	+0.214	-0.482	0.391	0.553
gemma-3-12b	29.6	43.5	+0.225	-0.479	0.749	0.409
Qwen2.5-0.5B	29.8	43.8	+0.226	-0.476	0.499	0.521
falcon-7b-Inst	27.3	44.2	+0.234	-0.460	0.391	0.550
gemma-3-4b	26.8	45.0	+0.236	-0.456	0.813	0.357
gemma-3-1b	22.8	46.5	+0.242	-0.427	0.694	0.383

(b) TriviaQA: S2 Popular Trap results across all models

Table 15: TriviaQA — S1 Baseline Control (left) and S2 Popular Trap (right)

E.4 TRIVIAQA (TABLES 15A–17B)

Table 15a (S1; mean -0.090). Baseline bias is small (the mildest across all datasets), with top accuracies in the mid–high 60s (`phi-4` 66.8%, `Llama-3.1-8B` 64.9%, `Meta-Llama-3-8B` 64.6%). POPGAP is modest (0.03–0.06) and HPSR $\approx 54\%$ for strong models.

Table 15b (S2; mean -0.531). Popular traps drive correlation to ≈ -0.58; POPGAP increases to 0.12–0.23. Strong models keep accuracies around 60–65%; small ones deteriorate.

Table 16a (S3; mean -0.547). Popularity gradients sustain strong negatives (≈ -0.60) and further inflate HPSR for weaker models. Accuracy ordering across families mirrors S2.

Table 16b (S4; mean -0.262). Direct contests are gentler (means ≈ -0.26); POPGAP narrows ($\sim$0.06–0.10), yet remains positive for virtually all models.

Table 17a (S5; mean $+0.263$). All models become positively correlated; several achieve their best accuracies here (e.g., `Llama-3.1-8B` 72.2%, `Meta-Llama-3-8B` 72.0%, `phi-4` 73.4%). POPGAP switches sign (slightly negative), confirming alignment of popularity and correctness.

Table 17b (S6; mean -0.704). Abstention again yields the largest negatives (down to -0.91) and large positive POPGAP (≥ 0.30). Accuracies for many models collapse to single digits with high HPSR ($\geq 58\%$), indicating a strong tendency to *avoid* the correct "None of the Above" in favor of the most common entity.

TriviaQA: S3 Popularity Gradient
Models: 23, Avg correlation: -0.547

Model	Accuracy (%)	HPSR (%)	PopGap	Corr	Confidence	Alignment
Qwen2.5-7B	60.6	29.0	+0.139	-0.615	0.791	0.645
Llama-3.1-8B-Inst	64.6	27.1	+0.112	-0.608	0.818	0.674
phi-4	67.4	24.4	+0.102	-0.598	0.876	0.687
Mistral-7B-Inst-v0.3	59.9	29.1	+0.127	-0.597	0.870	0.632
zephyr-7b-beta	58.6	29.6	+0.130	-0.594	0.893	0.613
Llama-3.2-3B	51.2	32.8	+0.156	-0.591	0.633	0.581
Meta-Llama-3-8B-Inst	67.9	23.0	+0.097	-0.586	0.913	0.701
gemma-2-27b	47.6	38.5	+0.176	-0.584	0.872	0.506
Qwen2.5-1.5B	45.9	36.6	+0.178	-0.570	0.650	0.559
Mistral-7B-Inst-v0.2	58.7	28.5	+0.125	-0.570	0.956	0.602
Llama-2-13b	51.1	33.8	+0.151	-0.569	0.575	0.556
Qwen2.5-3B	55.5	29.6	+0.141	-0.563	0.696	0.606
phi-2	48.9	34.4	+0.160	-0.545	0.609	0.548
gemma-2-9b	35.6	46.5	+0.228	-0.544	0.857	0.401
Llama-2-7b	31.3	49.4	+0.245	-0.519	0.396	0.550
gemma-3-4b	27.5	54.8	+0.272	-0.519	0.816	0.351
phi-1_5	32.5	46.7	+0.228	-0.519	0.443	0.533
gemma-2-2b	33.4	45.0	+0.220	-0.507	0.626	0.476
Llama-3.2-1B	32.3	43.6	+0.228	-0.500	0.465	0.540
gemma-3-1b	24.5	55.0	+0.275	-0.486	0.695	0.390
falcon-7b-Inst	24.8	55.4	+0.271	-0.477	0.392	0.555
gemma-3-12b	29.1	45.0	+0.229	-0.477	0.752	0.405
Qwen2.5-0.5B	30.8	36.9	+0.206	-0.452	0.497	0.524

(a) TriviaQA: S3 Popularity Gradient results across all models

TriviaQA: S4 Direct Contest
Models: 23, Avg correlation: -0.262

Model	Accuracy (%)	HPSR (%)	PopGap	Corr	Confidence	Alignment
phi-4	67.4	44.2	+0.066	-0.363	0.873	0.691
Qwen2.5-7B	60.8	45.8	+0.083	-0.359	0.784	0.648
zephyr-7b-beta	56.9	46.0	+0.082	-0.337	0.893	0.601
Mistral-7B-Inst-v0.3	59.1	44.1	+0.077	-0.332	0.858	0.629
Qwen2.5-3B	55.0	45.3	+0.091	-0.325	0.686	0.604
Llama-3.1-8B-Inst	64.8	43.4	+0.060	-0.324	0.819	0.674
Llama-3.2-3B	51.5	46.2	+0.090	-0.322	0.630	0.579
Meta-Llama-3-8B-Inst	64.8	43.2	+0.060	-0.294	0.907	0.669
Qwen2.5-1.5B	44.6	46.2	+0.107	-0.290	0.648	0.557
Llama-2-13b	50.7	45.0	+0.086	-0.288	0.567	0.570
Mistral-7B-Inst-v0.2	57.5	42.5	+0.069	-0.274	0.953	0.595
Llama-3.2-1B	32.5	49.0	+0.130	-0.268	0.471	0.541
Llama-2-7b	34.1	45.6	+0.100	-0.241	0.392	0.552
gemma-2-9b	34.0	44.8	+0.101	-0.226	0.850	0.403
gemma-2-27b	46.0	42.5	+0.074	-0.225	0.862	0.495
Qwen2.5-0.5B	30.9	45.8	+0.107	-0.217	0.498	0.525
phi-1_5	30.1	44.5	+0.103	-0.209	0.436	0.540
gemma-2-2b	31.3	44.6	+0.103	-0.207	0.622	0.478
gemma-3-12b	28.5	44.5	+0.099	-0.198	0.753	0.406
gemma-3-4b	27.0	44.6	+0.104	-0.190	0.812	0.358
falcon-7b-Inst	27.6	44.2	+0.102	-0.184	0.392	0.550
phi-2	43.4	41.5	+0.073	-0.182	0.592	0.538
gemma-3-1b	22.9	44.5	+0.109	-0.171	0.691	0.393

(b) TriviaQA: S4 Direct Contest results across all models

Table 16: TriviaQA — S3 Popularity Gradient (left) and S4 Direct Contest (right)

TriviaQA: S5 Reverse Control
Models: 23, Avg correlation: 0.263

Model	Accuracy (%)	HPSR (%)	PopGap	Corr	Confidence	Alignment
Qwen2.5-7B	67.2	55.4	-0.034	0.217	0.773	0.706
Llama-3.2-1B	37.5	50.2	-0.082	0.219	0.456	0.554
gemma-3-1b	23.2	44.5	-0.117	0.220	0.685	0.398
falcon-7b-Inst	25.4	44.4	-0.118	0.228	0.390	0.555
gemma-3-4b	27.9	45.7	-0.109	0.236	0.806	0.367
gemma-3-12b	30.0	47.3	-0.103	0.239	0.750	0.410
Qwen2.5-1.5B	50.2	51.2	-0.068	0.247	0.620	0.612
Qwen2.5-0.5B	32.2	47.9	-0.103	0.251	0.472	0.550
zephyr-7b-beta	62.0	53.5	-0.053	0.252	0.886	0.655
gemma-2-2b	33.7	47.4	-0.103	0.256	0.615	0.485
Llama-2-7b	36.4	48.6	-0.097	0.257	0.382	0.553
gemma-2-9b	36.0	47.9	-0.096	0.257	0.855	0.409
phi-1_5	31.3	44.9	-0.107	0.267	0.422	0.548
Qwen2.5-3B	60.5	53.1	-0.054	0.268	0.670	0.657
Llama-2-13b	56.9	51.5	-0.065	0.275	0.555	0.593
Llama-3.2-3B	58.5	51.9	-0.061	0.277	0.616	0.629
gemma-2-27b	47.4	49.8	-0.081	0.278	0.857	0.517
Mistral-7B-Inst-v0.3	64.8	53.0	-0.049	0.278	0.857	0.684
phi-4	73.4	55.5	-0.032	0.281	0.866	0.750
Llama-3.1-8B-Inst	72.2	54.1	-0.038	0.302	0.803	0.733
Mistral-7B-Inst-v0.2	61.5	52.2	-0.062	0.305	0.944	0.636
Meta-Llama-3-8B-Inst	72.0	53.9	-0.042	0.315	0.898	0.741
phi-2	44.8	46.8	-0.103	0.335	0.562	0.569

(a) TriviaQA: S5 Reverse Control results across all models

TriviaQA: S6 None of the Above
Models: 23, Avg correlation: -0.704

Model	Accuracy (%)	HPSR (%)	PopGap	Corr	Confidence	Alignment
gemma-2-2b	43.0	39.1	+0.257	-0.913	0.592	0.451
Meta-Llama-3-8B-Inst	34.2	45.5	+0.313	-0.906	0.788	0.392
phi-4	31.9	46.6	+0.301	-0.899	0.795	0.415
gemma-2-27b	29.4	46.9	+0.283	-0.893	0.799	0.348
gemma-3-12b	28.1	48.5	+0.317	-0.879	0.729	0.418
Mistral-7B-Inst-v0.3	24.3	53.8	+0.365	-0.870	0.793	0.322
zephyr-7b-beta	17.1	57.8	+0.383	-0.819	0.847	0.248
phi-1_5	19.7	54.5	+0.352	-0.819	0.416	0.534
Qwen2.5-3B	15.6	59.3	+0.407	-0.818	0.549	0.445
Qwen2.5-7B	16.1	59.0	+0.407	-0.811	0.655	0.376
Llama-3.1-8B-Inst	13.4	60.0	+0.403	-0.797	0.647	0.362
gemma-3-1b	14.4	55.8	+0.353	-0.781	0.640	0.364
Llama-2-13b	13.2	60.5	+0.425	-0.774	0.476	0.497
Llama-2-7b	11.3	59.7	+0.404	-0.752	0.350	0.609
gemma-3-4b	10.0	59.2	+0.388	-0.715	0.759	0.261
Mistral-7B-Inst-v0.2	7.6	62.5	+0.401	-0.704	0.937	0.119
Qwen2.5-1.5B	5.9	64.3	+0.435	-0.648	0.569	0.432
gemma-2-9b	5.8	61.4	+0.390	-0.638	0.818	0.195
phi-2	6.3	58.0	+0.344	-0.625	0.537	0.452
Llama-3.2-3B	1.8	68.7	+0.468	-0.428	0.547	0.451
Qwen2.5-0.5B	1.8	65.1	+0.409	-0.401	0.483	0.511
falcon-7b-Inst	0.5	65.1	+0.394	-0.235	0.386	0.613
Llama-3.2-1B	0.1	71.4	+0.490	-0.077	0.501	0.499

(b) TriviaQA: S6 None of the Above results across all models

Table 17: TriviaQA — S5 Reverse Control (left) and S6 None of the Above (right)

Cross-table summary. Across all datasets: (i) S1 shows mild bias; (ii) S2/S3 induce strong negative correlations and larger PopGap; (iii) S4 is intermediate; (iv) S5 flips the sign and slightly improves accuracy; and (v) S6 is consistently the most adverse setting, with very negative correlations and high HPSR. These consistent patterns, together with the per-model spreads, suggest a shared mechanism that privileges entity familiarity over factuality under popularity pressure, and a pronounced failure to abstain when the correct option is "None of the Above".

F Detailed Results for DeBAIS Models

How to read these tables. Green "After" values denote improvements relative to "Before"; red denotes regressions. For **HPSR** (popularity susceptibility), **lower is better**. For **PopGap**, **absolute values closer to** 0 **are better** (reduces popularity skew). **Corr** is the correlation with popularity signals; values **closer to** 0 **are better**. **Alignment** is better when higher. **Confidence** is descriptive (we do not optimize it directly). For the Spearman correlations with popularity (**Spearman Conf Pop** and **Spearman Align Pop**), **magnitudes closer to** 0 **are better** (we seek minimal monotonic dependence on popularity).

F.1 TriviaQA

Table 18: Comprehensive debiasing results for **TriviaQA: S1 Baseline Control**. An improvement in an 'After' cell is colored green; a regression is colored red.

Model	Accuracy (%)		HPSR (%)		PopGap		Corr		Confidence		Alignment		Spearman Conf Pop		Spearman Align Pop	
	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After
Llama-3.2-3B	52.6	53.1	53.9	51.0	0.052	0.028	-0.143	-0.060	0.631	0.574	0.583	0.581	0.117	-0.029	-0.140	-0.149
Mistral-7B-Instructruct-v0.3	60.9	61.1	55.1	54.5	0.048	0.042	-0.150	-0.125	0.866	0.798	0.636	0.627	0.057	-0.329	-0.158	-0.282
gemma-3-12b-it	30.2	30.1	48.5	46.3	0.041	0.021	-0.070	-0.035	0.752	0.665	0.408	0.448	0.044	-0.278	-0.098	0.050
phi-2	42.7	43.0	45.4	40.2	0.009	-0.033	0.058	0.157	0.585	0.536	0.537	0.544	0.101	0.018	-0.101	-0.099
Meta-Llama-3-8B-Instructruct	64.6	64.6	55.1	55.0	0.040	0.039	-0.167	-0.159	0.905	0.837	0.667	0.654	0.058	-0.444	-0.159	-0.353
Qwen2.5-7B	60.4	60.8	55.6	53.9	0.055	0.040	-0.187	-0.132	0.787	0.724	0.646	0.639	0.054	-0.180	-0.158	-0.244
Qwen2.5-1.5B	44.4	44.8	52.6	47.9	0.054	0.013	-0.095	0.010	0.644	0.562	0.563	0.571	0.140	-0.015	-0.160	-0.153
Mistral-7B-Instructruct-v0.2	58.2	58.1	53.0	52.6	0.039	0.035	-0.092	-0.082	0.947	0.870	0.601	0.595	0.036	-0.597	-0.132	-0.249
gemma-2-2b-it	32.7	33.4	48.4	43.4	0.040	-0.007	-0.050	0.030	0.623	0.553	0.474	0.501	0.075	-0.105	-0.063	0.046
gemma-2-27b-it	46.2	46.2	49.4	48.6	0.030	0.024	-0.051	-0.032	0.856	0.774	0.507	0.523	-0.023	-0.452	-0.100	-0.086
falcon-7b-instruct	26.5	27.8	45.1	28.4	0.023	-0.118	-0.015	0.232	0.390	0.366	0.552	0.559	-0.006	-0.254	0.035	0.135
Llama-3.1-8B-Instructruct	64.9	65.5	55.1	54.4	0.044	0.035	-0.165	-0.134	0.813	0.753	0.678	0.665	0.047	-0.239	-0.148	-0.286
phi-4	66.8	67.0	55.4	54.4	0.041	0.034	-0.135	-0.102	0.872	0.826	0.687	0.680	0.081	-0.254	-0.106	-0.285
gemma-2-9b-it	35.0	35.1	47.7	45.8	0.033	0.015	-0.031	0.003	0.850	0.750	0.405	0.444	0.116	-0.355	-0.157	-0.004
zephyr-7b-beta	59.2	59.2	54.8	54.5	0.039	0.037	-0.148	-0.135	0.896	0.833	0.611	0.602	0.042	-0.408	-0.140	-0.253
Llama-2-7b-hf	34.2	35.4	50.5	35.0	0.046	-0.084	-0.085	0.181	0.390	0.360	0.552	0.557	0.063	-0.007	-0.011	-0.000
Qwen2.5-3B	54.2	54.8	53.5	51.3	0.047	0.024	-0.104	-0.032	0.687	0.631	0.607	0.603	0.092	-0.047	-0.159	-0.188
Llama-3.2-1B	33.7	34.4	53.4	43.2	0.076	-0.008	-0.131	0.053	0.473	0.420	0.532	0.550	0.133	0.074	-0.045	-0.010
Llama-2-13b-hf	52.9	53.2	53.8	48.4	0.048	0.002	-0.130	0.013	0.570	0.520	0.569	0.567	0.040	-0.001	-0.098	-0.103
gemma-3-4b-it	28.4	28.8	46.5	44.6	0.037	0.020	-0.044	-0.016	0.813	0.714	0.371	0.418	0.028	-0.376	-0.062	0.118
phi-1_5	29.5	30.1	47.0	34.7	0.033	-0.074	-0.039	0.152	0.429	0.395	0.544	0.558	0.077	0.002	0.008	0.021
Qwen2.5-0.5B	29.2	30.6	48.8	38.3	0.038	-0.052	-0.054	0.108	0.494	0.443	0.533	0.548	0.087	0.036	-0.042	-0.025
gemma-3-1b-it	24.2	24.8	46.1	42.2	0.035	-0.001	-0.039	0.011	0.688	0.605	0.395	0.444	0.079	-0.181	-0.088	0.069

Table 19: Comprehensive debiasing results for **TriviaQA: S2 Popular Trap**. An improvement in an 'After' cell is colored green; a regression is colored red.

Model	Accuracy (%)		HPSR (%)		PopGap		Corr		Confidence		Alignment		Spearman Conf Pop		Spearman Align Pop	
	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After
Llama-3.2-3B	50.3	56.4	33.7	26.2	0.159	0.095	-0.587	-0.524	0.634	0.590	0.579	0.587	-0.095	-0.033	-0.232	-0.071
Mistral-7B-Instructruct-v0.3	59.5	61.8	28.8	26.6	0.127	0.107	-0.589	-0.571	0.867	0.790	0.628	0.656	-0.113	-0.285	-0.378	-0.375
gemma-3-12b-it	29.6	34.5	43.5	37.1	0.225	0.171	-0.479	-0.463	0.749	0.639	0.409	0.490	-0.032	-0.097	-0.289	-0.160
phi-2	49.6	55.7	31.2	24.0	0.145	0.082	-0.522	-0.452	0.602	0.556	0.547	0.558	-0.070	-0.030	-0.259	-0.086
Meta-Llama-3-8B-Instructruct	64.2	65.0	25.2	24.2	0.106	0.098	-0.593	-0.582	0.910	0.840	0.667	0.689	-0.128	-0.346	-0.344	-0.409
Qwen2.5-7B	60.3	64.3	27.5	23.4	0.124	0.087	-0.577	-0.525	0.795	0.734	0.647	0.662	-0.135	-0.155	-0.345	-0.258
Qwen2.5-1.5B	44.8	51.9	35.3	26.9	0.176	0.102	-0.540	-0.478	0.646	0.587	0.561	0.578	-0.063	-0.015	-0.272	-0.075
Mistral-7B-Instructruct-v0.2	58.7	59.5	28.4	27.5	0.126	0.119	-0.582	-0.577	0.956	0.862	0.605	0.645	-0.121	-0.421	-0.386	-0.437
gemma-2-2b-it	32.8	40.1	41.9	32.7	0.209	0.128	-0.493	-0.459	0.624	0.554	0.479	0.519	-0.005	0.010	-0.237	-0.055
gemma-2-27b-it	46.0	47.5	33.1	31.2	0.155	0.141	-0.521	-0.512	0.862	0.770	0.502	0.554	-0.106	-0.281	-0.402	-0.375
falcon-7b-instruct	27.3	42.5	44.2	25.9	0.234	0.078	-0.460	-0.396	0.391	0.373	0.550	0.532	0.035	-0.217	0.305	0.432
Llama-3.1-8B-Instructruct	65.1	67.2	25.3	22.9	0.108	0.085	-0.602	-0.563	0.821	0.766	0.678	0.688	-0.129	-0.206	-0.328	-0.296
phi-4	65.2	67.3	27.3	25.1	0.122	0.103	-0.633	-0.615	0.873	0.813	0.680	0.699	-0.169	-0.283	-0.350	-0.366
gemma-2-9b-it	36.0	37.9	39.8	37.2	0.201	0.178	-0.509	-0.497	0.859	0.736	0.411	0.498	0.023	-0.225	-0.393	-0.315
zephyr-7b-beta	57.9	59.5	29.1	27.4	0.128	0.112	-0.580	-0.560	0.897	0.814	0.607	0.643	-0.106	-0.330	-0.383	-0.409
Llama-2-7b-hf	32.0	49.3	41.3	21.6	0.214	0.050	-0.482	-0.359	0.391	0.377	0.553	0.521	-0.007	-0.151	0.250	0.366
Qwen2.5-3B	54.8	60.8	29.9	23.3	0.140	0.082	-0.559	-0.497	0.695	0.643	0.602	0.611	-0.118	-0.067	-0.292	-0.142
Llama-3.2-1B	31.9	44.0	43.5	28.1	0.220	0.091	-0.495	-0.427	0.469	0.429	0.538	0.536	-0.042	-0.068	0.053	0.257
Llama-2-13b-hf	52.6	59.5	31.8	23.1	0.149	0.074	-0.568	-0.465	0.574	0.534	0.563	0.565	-0.091	-0.046	-0.160	0.020
gemma-3-4b-it	26.8	30.9	45.0	40.1	0.236	0.191	-0.456	-0.454	0.813	0.677	0.357	0.465	-0.006	-0.136	-0.299	-0.169
phi-1_5	31.6	45.6	41.9	26.1	0.220	0.080	-0.483	-0.409	0.441	0.409	0.535	0.527	0.054	-0.044	0.149	0.322
Qwen2.5-0.5B	29.8	42.7	43.8	28.7	0.226	0.093	-0.476	-0.415	0.499	0.455	0.521	0.530	0.003	-0.023	0.007	0.214
gemma-3-1b-it	22.8	31.5	46.5	36.4	0.242	0.148	-0.427	-0.425	0.694	0.579	0.383	0.464	0.093	0.107	-0.278	-0.056

S1 Baseline Control (Table 18). Across models, accuracy changes are modest (typically within ± 1 point), while HPSR generally drops, indicating reduced popularity susceptibility even when the task does not adversarially emphasize popularity. PopGap and Corr move toward zero for most models, showing less systematic preference for popular answers. Alignment is largely stable with small fluctuations. Spearman correlations often drift toward zero but can be noisy for smaller models.

S2 Popular Trap (Table 19). This adversarial setting produces the clearest gains: accuracy increases are sizable (often $+4$–12 points, especially for smaller/older models), and HPSR drops strongly. PopGap shrinks toward zero and Corr becomes less extreme, indicating the debias step counters the trap. Alignment often improves or holds steady. Spearman correlations move closer to zero for many models, though some well-aligned large models show mixed Spearman movement (typical when gains concentrate on hard/popular distractors).

Table 20: Comprehensive debiasing results for **TriviaQA: S3 Popularity Gradient**. An improvement in an 'After' cell is colored green; a regression is colored red.

Model	Accuracy (%)		HPSR (%)		PopGap		Corr		Confidence		Alignment		Spearman Conf Pop		Spearman Align Pop	
	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After
Llama-3.2-3B	51.2	58.4	32.8	24.9	0.156	0.087	-0.591	-0.520	0.633	0.592	0.581	0.587	-0.089	-0.028	-0.236	-0.059
Mistral-7B-Instructruct-v0.3	59.9	62.0	29.1	26.9	0.127	0.107	-0.597	-0.574	0.870	0.794	0.632	0.661	-0.119	-0.266	-0.392	-0.372
gemma-3-12b-it	29.1	34.6	45.0	38.5	0.229	0.170	-0.477	-0.468	0.752	0.630	0.405	0.490	-0.046	-0.083	-0.274	-0.122
phi-2	48.9	56.6	34.4	25.1	0.160	0.079	-0.545	-0.452	0.609	0.556	0.548	0.563	-0.088	-0.047	-0.293	-0.088
Meta-Llama-3-8B-Instructruct	67.9	70.7	23.0	20.0	0.097	0.070	-0.586	-0.535	0.913	0.820	0.701	0.716	-0.118	-0.194	-0.355	-0.303
Qwen2.5-7B	60.6	64.0	29.0	25.0	0.139	0.103	-0.615	-0.572	0.791	0.724	0.645	0.666	-0.148	-0.157	-0.363	-0.242
Qwen2.5-1.5B	45.9	53.4	36.6	28.1	0.178	0.104	-0.570	-0.507	0.650	0.589	0.559	0.577	-0.074	-0.038	-0.265	-0.074
Mistral-7B-Instructruct-v0.2	58.7	59.0	28.5	28.0	0.125	0.121	-0.570	-0.565	0.956	0.861	0.602	0.648	-0.121	-0.441	-0.408	-0.453
gemma-2-2b-it	33.4	42.0	45.0	34.9	0.220	0.131	-0.507	-0.472	0.626	0.550	0.476	0.520	-0.030	-0.015	-0.286	-0.070
gemma-2-27b-it	47.6	49.0	38.5	36.9	0.176	0.162	-0.584	-0.573	0.872	0.757	0.506	0.580	-0.074	-0.347	-0.497	-0.489
falcon-7b-instruct	24.8	44.5	55.4	33.2	0.271	0.082	-0.477	-0.422	0.392	0.372	0.555	0.532	0.047	-0.206	0.345	0.464
Llama-3.1-8B-Instructruct	64.6	68.3	27.1	22.7	0.112	0.071	-0.608	-0.540	0.818	0.755	0.674	0.686	-0.135	-0.158	-0.337	-0.251
phi-4	67.4	68.8	24.4	23.1	0.102	0.088	-0.598	-0.571	0.876	0.813	0.687	0.703	-0.091	-0.149	-0.297	-0.268
gemma-2-9b-it	35.6	38.2	46.5	43.4	0.228	0.199	-0.544	-0.536	0.857	0.711	0.401	0.512	0.071	-0.197	-0.491	-0.405
zephyr-7b-beta	58.6	60.4	29.6	27.7	0.130	0.114	-0.594	-0.574	0.893	0.806	0.613	0.649	-0.079	-0.303	-0.385	-0.400
Llama-2-7b-hf	31.3	51.6	49.4	27.2	0.245	0.060	-0.519	-0.421	0.396	0.383	0.550	0.514	-0.001	-0.141	0.284	0.374
Qwen2.5-3B	55.5	61.0	29.6	23.6	0.141	0.087	-0.563	-0.502	0.696	0.644	0.606	0.617	-0.124	-0.097	-0.295	-0.152
Llama-3.2-1B	32.3	46.8	43.6	26.8	0.228	0.085	-0.500	-0.424	0.465	0.428	0.540	0.528	-0.050	-0.041	0.092	0.298
Llama-2-13b-hf	51.1	57.8	33.8	26.3	0.151	0.082	-0.569	-0.496	0.575	0.536	0.556	0.561	-0.066	-0.047	-0.165	0.012
gemma-3-4b-it	27.5	30.9	54.8	49.9	0.272	0.231	-0.519	-0.514	0.816	0.658	0.351	0.480	0.038	-0.114	-0.439	-0.332
phi-1_5	32.5	45.8	46.7	30.5	0.228	0.090	-0.519	-0.449	0.443	0.406	0.533	0.530	0.037	-0.049	0.169	0.350
Qwen2.5-0.5B	30.8	42.0	36.9	23.4	0.206	0.087	-0.452	-0.384	0.497	0.453	0.524	0.533	-0.006	-0.032	0.022	0.225
gemma-3-1b-it	24.5	30.9	55.0	46.4	0.275	0.200	-0.486	-0.488	0.695	0.565	0.390	0.489	0.126	0.095	-0.358	-0.175

Table 21: Comprehensive debiasing results for **TriviaQA: S4 Direct Contest**. An improvement in an 'After' cell is colored green; a regression is colored red.

Model	Accuracy (%)		HPSR (%)		PopGap		Corr		Confidence		Alignment		Spearman Conf Pop		Spearman Align Pop	
	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After
Llama-3.2-3B	51.5	53.0	46.2	42.0	0.090	0.053	-0.322	-0.220	0.630	0.577	0.579	0.583	0.051	-0.016	-0.140	-0.124
Mistral-7B-Instructruct-v0.3	59.1	59.5	44.1	43.2	0.077	0.067	-0.332	-0.309	0.858	0.783	0.629	0.630	0.004	-0.328	-0.223	-0.325
gemma-3-12b-it	28.5	28.9	44.5	42.5	0.099	0.083	-0.198	-0.172	0.753	0.666	0.406	0.455	0.002	-0.302	-0.157	-0.005
phi-2	43.4	44.9	41.5	37.4	0.073	0.040	-0.182	-0.107	0.592	0.540	0.538	0.548	0.062	-0.038	-0.166	-0.099
Meta-Llama-3-8B-Instructruct	64.8	65.0	43.2	42.6	0.060	0.055	-0.294	-0.275	0.907	0.849	0.669	0.670	0.014	-0.411	-0.204	-0.372
Qwen2.5-7B	60.8	61.5	45.8	43.9	0.083	0.067	-0.359	-0.298	0.784	0.731	0.648	0.649	-0.018	-0.148	-0.237	-0.261
Qwen2.5-1.5B	44.6	46.0	46.2	41.1	0.107	0.061	-0.290	-0.182	0.648	0.579	0.557	0.571	0.040	-0.052	-0.187	-0.128
Mistral-7B-Instructruct-v0.2	57.5	57.6	42.5	42.2	0.069	0.067	-0.274	-0.268	0.953	0.878	0.595	0.598	-0.028	-0.616	-0.225	-0.337
gemma-2-2b-it	31.3	32.4	44.6	41.5	0.103	0.075	-0.207	-0.168	0.622	0.559	0.478	0.507	0.001	-0.165	-0.117	-0.012
gemma-2-27b-it	46.0	46.2	42.5	41.6	0.074	0.067	-0.225	-0.208	0.862	0.786	0.495	0.519	-0.030	-0.417	-0.212	-0.198
falcon-7b-instruct	27.6	31.8	44.2	30.1	0.102	-0.012	-0.184	-0.012	0.392	0.366	0.550	0.552	-0.005	-0.228	0.151	0.240
Llama-3.1-8B-Instructruct	64.8	65.2	43.4	42.7	0.060	0.055	-0.324	-0.300	0.819	0.774	0.674	0.670	0.008	-0.216	-0.206	-0.307
phi-4	67.4	67.8	44.2	43.6	0.066	0.061	-0.363	-0.342	0.873	0.824	0.691	0.690	0.003	-0.262	-0.191	-0.328
gemma-2-9b-it	34.0	34.4	44.8	43.7	0.101	0.091	-0.226	-0.213	0.850	0.747	0.403	0.455	0.044	-0.404	-0.217	-0.078
zephyr-7b-beta	56.9	57.4	46.0	45.0	0.082	0.072	-0.337	-0.313	0.893	0.816	0.601	0.610	-0.020	-0.439	-0.210	-0.302
Llama-2-7b-hf	34.1	37.8	45.6	34.0	0.100	0.002	-0.241	-0.065	0.392	0.369	0.552	0.549	-0.001	-0.030	0.128	0.181
Qwen2.5-3B	55.0	55.9	45.3	41.5	0.091	0.060	-0.325	-0.227	0.686	0.636	0.604	0.609	-0.016	-0.093	-0.207	-0.189
Llama-3.2-1B	32.5	35.5	49.0	40.3	0.130	0.052	-0.268	-0.158	0.471	0.422	0.541	0.553	0.075	0.020	0.004	0.104
Llama-2-13b-hf	50.7	52.5	45.0	40.0	0.086	0.044	-0.288	-0.172	0.567	0.522	0.570	0.572	-0.017	-0.067	-0.106	-0.067
gemma-3-4b-it	27.0	27.6	44.6	43.5	0.104	0.092	-0.190	-0.177	0.812	0.713	0.358	0.420	0.042	-0.357	-0.177	0.022
phi-1_5	30.1	33.5	44.5	32.6	0.103	0.003	-0.209	-0.047	0.436	0.400	0.540	0.549	0.072	0.026	0.083	0.152
Qwen2.5-0.5B	30.9	34.8	45.8	36.5	0.107	0.025	-0.217	-0.092	0.498	0.440	0.525	0.538	0.098	0.035	-0.011	0.092
gemma-3-1b-it	22.9	23.8	44.5	41.2	0.109	0.079	-0.171	-0.143	0.691	0.607	0.393	0.448	0.067	-0.157	-0.143	0.013

Table 22: Comprehensive debiasing results for **TriviaQA: S5 Reverse Control**. An improvement in an 'After' cell is colored green; a regression is colored red.

Model	Accuracy (%)		HPSR (%)		PopGap		Corr		Confidence		Alignment		Spearman Conf Pop		Spearman Align Pop	
	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After
Llama-3.2-3B	58.5	58.6	51.9	51.5	-0.061	-0.066	0.277	0.292	0.616	0.594	0.629	0.612	0.170	0.112	0.125	0.074
Mistral-7B-Instructruct-v0.3	64.8	64.8	53.0	52.7	-0.049	-0.052	0.278	0.288	0.857	0.833	0.684	0.669	0.107	-0.053	0.183	0.100
gemma-3-12b-it	30.0	29.2	47.3	46.3	-0.103	-0.114	0.239	0.248	0.750	0.708	0.410	0.411	0.054	-0.116	0.100	0.112
phi-2	44.8	44.1	46.8	44.6	-0.103	-0.124	0.335	0.368	0.562	0.539	0.569	0.562	0.155	0.138	0.072	0.053
Meta-Llama-3-8B-Instructruct	72.0	72.0	53.9	53.9	-0.042	-0.042	0.315	0.315	0.898	0.879	0.741	0.728	0.173	0.053	0.201	0.134
Qwen2.5-7B	67.2	67.2	55.4	55.0	-0.034	-0.038	0.217	0.229	0.773	0.750	0.706	0.689	0.152	0.050	0.162	0.084
Qwen2.5-1.5B	50.2	50.0	51.2	50.5	-0.068	-0.075	0.247	0.263	0.620	0.595	0.612	0.600	0.190	0.131	0.121	0.089
Mistral-7B-Instructruct-v0.2	61.5	61.5	52.2	52.2	-0.062	-0.062	0.305	0.305	0.944	0.922	0.636	0.626	0.083	-0.118	0.174	0.123
gemma-2-2b-it	33.7	33.5	47.4	46.2	-0.103	-0.117	0.256	0.275	0.615	0.588	0.485	0.483	0.097	0.012	0.077	0.061
gemma-2-27b-it	47.4	47.4	49.8	49.5	-0.081	-0.085	0.278	0.285	0.857	0.823	0.517	0.507	0.047	-0.128	0.143	0.128
falcon-7b-instruct	25.4	23.9	44.4	38.3	-0.118	-0.175	0.228	0.289	0.390	0.381	0.555	0.562	-0.022	-0.099	-0.100	0.015
Llama-3.1-8B-Instructruct	72.2	72.2	54.1	54.0	-0.038	-0.039	0.302	0.307	0.803	0.784	0.733	0.719	0.149	0.058	0.160	0.092
phi-4	73.4	73.4	55.5	55.5	-0.032	-0.032	0.281	0.281	0.866	0.852	0.750	0.739	0.198	0.074	0.213	0.114
gemma-2-9b-it	36.0	35.9	47.9	47.7	-0.096	-0.098	0.257	0.260	0.855	0.813	0.409	0.409	0.073	-0.183	0.083	0.142
zephyr-7b-beta	62.0	62.1	53.5	53.5	-0.053	-0.053	0.252	0.254	0.886	0.860	0.655	0.640	0.101	-0.144	0.176	0.071
Llama-2-7b-hf	36.4	35.0	48.6	42.9	-0.097	-0.152	0.257	0.338	0.382	0.372	0.553	0.556	0.018	-0.014	-0.110	-0.055
Qwen2.5-3B	60.5	60.6	53.1	52.4	-0.054	-0.061	0.268	0.289	0.670	0.648	0.657	0.641	0.150	0.100	0.147	0.109
Llama-3.2-1B	37.5	37.0	50.2	47.6	-0.082	-0.108	0.219	0.259	0.456	0.437	0.554	0.550	0.082	0.074	-0.012	-0.041
Llama-2-13b-hf	56.9	56.7	51.5	50.6	-0.065	-0.075	0.275	0.301	0.555	0.537	0.593	0.583	0.107	0.084	0.034	0.006
gemma-3-4b-it	27.9	27.8	45.7	45.0	-0.109	-0.115	0.236	0.244	0.806	0.774	0.367	0.370	-0.018	-0.164	0.133	0.177
phi-1_5	31.3	30.3	44.9	39.4	-0.107	-0.159	0.267	0.335	0.422	0.407	0.548	0.551	0.058	0.002	-0.064	-0.029
Qwen2.5-0.5B	32.2	31.8	47.9	44.0	-0.103	-0.141	0.251	0.303	0.472	0.450	0.550	0.547	0.118	0.067	-0.048	-0.052
gemma-3-1b-it	23.2	23.1	44.5	43.5	-0.117	-0.127	0.220	0.230	0.685	0.654	0.398	0.407	0.046	-0.047	0.066	0.091

S3 Popularity Gradient (Table 20). With graded popularity pressure, patterns mirror S2: accuracy rises, HPSR declines, PopGap compresses, and Corr approaches zero. Improvements are broad-based across sizes, with smaller models seeing the largest relative gains. Alignment tends to improve slightly; confidence shifts are modest.

S4 Direct Contest (Table 21). Head-to-head contests modestly raise accuracy for most models while consistently lowering HPSR and pulling PopGap/Corr toward zero. Alignment remains stable

Table 23: Comprehensive debiasing results for **TriviaQA: S6 None of the Above**. An improvement in an 'After' cell is colored green; a regression is colored red.

Model	Accuracy (%)		HPSR (%)		PopGap		Corr		Confidence		Alignment		Spearman Conf Pop		Spearman Align Pop	
	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After
Llama-3.2-3B	1.8	17.1	68.7	37.1	0.468	0.059	-0.428	-0.723	0.547	0.435	0.451	0.519	0.197	0.041	-0.172	0.363
Mistral-7B-Instructruct-v0.3	24.3	41.0	53.8	28.0	0.365	0.051	-0.870	-0.847	0.793	0.606	0.322	0.457	0.259	0.041	-0.574	-0.235
gemma-3-12b-it	28.1	42.3	48.5	25.8	0.317	0.051	-0.879	-0.855	0.729	0.610	0.418	0.503	-0.025	-0.110	-0.570	-0.281
phi-2	6.3	21.9	58.0	33.6	0.344	0.051	-0.625	-0.760	0.537	0.461	0.452	0.493	0.180	0.103	-0.067	0.307
Meta-Llama-3-8B-Instructruct	34.2	54.9	45.5	20.4	0.313	0.043	-0.906	-0.891	0.788	0.635	0.392	0.490	0.189	0.012	-0.679	-0.350
Qwen2.5-7B	16.1	37.0	59.0	28.8	0.407	0.049	-0.811	-0.833	0.655	0.510	0.376	0.465	0.228	0.100	-0.377	0.114
Qwen2.5-1.5B	5.9	18.9	64.3	35.8	0.435	0.060	-0.648	-0.741	0.569	0.449	0.432	0.512	0.220	0.044	-0.217	0.305
Mistral-7B-Instructruct-v0.2	7.6	22.6	62.5	37.5	0.401	0.084	-0.704	-0.787	0.937	0.669	0.119	0.319	0.217	0.013	-0.309	0.021
gemma-2-2b-it	43.0	56.9	39.1	20.1	0.257	0.036	-0.913	-0.882	0.592	0.526	0.451	0.493	0.168	-0.014	-0.377	-0.053
gemma-2-27b-it	29.4	44.0	46.9	26.6	0.283	0.052	-0.893	-0.883	0.799	0.656	0.348	0.447	0.205	0.040	-0.603	-0.363
falcon-7b-instruct	0.5	8.9	65.1	38.5	0.394	0.054	-0.235	-0.587	0.386	0.383	0.613	0.598	0.009	-0.214	0.005	0.378
Llama-3.1-8B-Instructruct	13.4	31.9	60.0	31.4	0.403	0.054	-0.797	-0.823	0.647	0.502	0.362	0.462	0.230	0.072	-0.267	0.168
phi-4	31.9	48.6	46.6	24.8	0.301	0.044	-0.899	-0.880	0.795	0.651	0.415	0.512	0.033	-0.113	-0.682	-0.368
gemma-2-9b-it	5.8	20.3	61.4	35.9	0.390	0.074	-0.638	-0.758	0.818	0.613	0.195	0.346	0.173	0.156	-0.216	0.038
zephyr-7b-beta	17.1	37.6	57.8	29.9	0.383	0.060	-0.819	-0.845	0.847	0.635	0.248	0.390	0.217	0.120	-0.459	-0.188
Llama-2-7b-hf	11.3	27.6	59.7	29.8	0.404	0.037	-0.752	-0.774	0.350	0.349	0.609	0.579	0.235	-0.150	0.157	0.575
Qwen2.5-3B	15.6	39.6	59.3	26.9	0.407	0.041	-0.818	-0.835	0.549	0.450	0.445	0.490	0.231	0.056	-0.174	0.359
Llama-3.2-1B	0.1	7.7	71.4	41.1	0.490	0.069	-0.077	-0.541	0.501	0.417	0.499	0.561	0.135	-0.071	-0.133	0.298
Llama-2-13b-hf	13.2	32.4	60.5	28.2	0.425	0.046	-0.774	-0.798	0.476	0.406	0.497	0.525	0.204	-0.058	0.061	0.518
gemma-3-4b-it	10.0	28.7	59.2	32.0	0.388	0.066	-0.715	-0.791	0.759	0.579	0.261	0.381	0.164	0.190	-0.240	0.016
phi-1_5	19.7	43.2	54.5	24.4	0.352	0.032	-0.819	-0.840	0.416	0.395	0.534	0.507	0.161	0.000	0.326	0.622
Qwen2.5-0.5B	1.8	15.2	65.1	36.6	0.409	0.051	-0.401	-0.680	0.483	0.418	0.511	0.547	0.148	-0.013	-0.088	0.338
gemma-3-1b-it	14.4	32.4	55.8	29.8	0.353	0.055	-0.781	-0.814	0.640	0.528	0.364	0.433	0.194	0.178	-0.208	0.080

or improves a bit. Spearman terms generally move toward zero but remain variable across architectures.

S5 Reverse Control (Table 22). As intended, this control leaves accuracy nearly flat and avoids large swings elsewhere. Small drifts appear (slightly more negative PopGap/Corr for some models), but alignment and accuracy are effectively preserved, confirming the debias does not overfit to the original direction of popularity.

S6 None of the Above (Table 23). This setting shows large accuracy jumps (often double-digit) and strong HPSR reductions, indicating better rejection of popular but incorrect options. PopGap compresses sharply. Corr and Spearman terms sometimes fluctuate (especially on smaller models) as models become more conservative; alignment generally improves or holds.

F.2 QASC

Table 24: Comprehensive debiasing results for **QASC: S1 Baseline Control**. An improvement in an 'After' cell is colored green; a regression is colored red.

Model	Accuracy (%)		HPSR (%)		PopGap		Corr		Confidence		Alignment		Spearman Conf Pop		Spearman Align Pop	
	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After
Llama-3.2-3B	44.6	45.5	56.2	51.9	0.047	0.010	-0.165	-0.083	0.626	0.564	0.535	0.540	0.011	-0.154	-0.059	-0.023
Mistral-7B-Instructruct-v0.3	47.0	47.0	55.4	54.8	0.040	0.034	-0.157	-0.143	0.838	0.748	0.519	0.522	-0.041	-0.494	-0.052	-0.044
gemma-3-12b-it	23.9	24.4	48.9	47.0	0.037	0.019	-0.068	-0.047	0.757	0.664	0.364	0.413	0.016	-0.335	-0.044	0.150
phi-2	45.0	45.4	53.0	51.0	0.028	0.009	-0.102	-0.059	0.648	0.582	0.539	0.540	0.033	-0.212	-0.060	-0.030
Meta-Llama-3-8B-Instructruct	52.9	52.9	56.3	55.6	0.046	0.040	-0.169	-0.152	0.863	0.775	0.577	0.574	0.000	-0.458	-0.082	-0.157
Qwen2.5-7B	58.1	58.5	56.6	55.4	0.042	0.030	-0.164	-0.128	0.781	0.708	0.610	0.599	0.021	-0.281	-0.101	-0.164
Qwen2.5-1.5B	49.0	49.8	55.4	52.9	0.038	0.016	-0.158	-0.096	0.686	0.624	0.554	0.555	0.034	-0.157	-0.086	-0.069
Mistral-7B-Instructruct-v0.2	47.6	47.6	54.8	54.7	0.032	0.031	-0.136	-0.134	0.942	0.847	0.497	0.505	-0.011	-0.644	-0.070	-0.070
gemma-2-2b-it	29.6	30.6	48.5	43.8	0.018	-0.023	-0.052	0.018	0.641	0.578	0.445	0.470	0.033	-0.149	-0.044	0.050
gemma-2-27b-it	28.4	28.7	47.2	46.2	0.015	0.006	-0.032	-0.021	0.879	0.786	0.340	0.380	-0.057	-0.537	-0.036	0.155
falcon-7b-instruct	23.4	25.3	47.5	32.8	0.032	-0.099	-0.065	0.123	0.376	0.353	0.567	0.571	-0.026	-0.198	0.039	0.085
Llama-3.1-8B-Instructruct	52.0	52.0	56.6	53.6	0.044	0.024	-0.182	-0.113	0.744	0.671	0.577	0.576	0.023	-0.212	-0.054	-0.058
phi-4	53.9	54.2	56.6	55.5	0.047	0.036	-0.186	-0.157	0.814	0.730	0.576	0.570	-0.015	-0.379	-0.099	-0.149
gemma-2-9b-it	28.9	29.2	48.2	46.9	0.025	0.011	-0.049	-0.027	0.875	0.779	0.348	0.388	0.011	-0.459	-0.061	0.127
zephyr-7b-beta	43.2	43.5	54.0	53.1	0.038	0.029	-0.138	-0.120	0.886	0.789	0.477	0.492	0.015	-0.493	-0.068	-0.021
Llama-2-7b-hf	30.8	31.1	50.2	35.6	0.031	-0.086	-0.094	0.122	0.365	0.347	0.562	0.568	-0.002	-0.068	0.028	0.059
Qwen2.5-3B	53.4	54.1	56.0	53.2	0.041	0.018	-0.160	-0.095	0.701	0.641	0.584	0.583	0.022	-0.150	-0.092	-0.095
Llama-3.2-1B	32.2	33.3	53.0	46.1	0.047	-0.014	-0.121	-0.016	0.519	0.466	0.512	0.527	0.045	-0.098	-0.008	0.035
Llama-2-13b-hf	37.0	38.7	51.5	45.5	0.039	-0.014	-0.095	0.003	0.520	0.475	0.526	0.531	-0.022	-0.110	0.038	0.068
gemma-3-4b-it	24.4	24.3	48.1	45.8	0.026	0.006	-0.046	-0.020	0.793	0.709	0.342	0.388	-0.020	-0.377	-0.035	0.159
phi-1_5	34.2	34.9	49.0	42.9	0.016	-0.035	-0.050	0.057	0.555	0.511	0.510	0.522	0.044	-0.046	-0.028	0.010
Qwen2.5-0.5B	35.6	35.9	50.8	48.2	0.022	0.000	-0.083	-0.041	0.601	0.551	0.503	0.516	0.031	-0.145	-0.047	0.015
gemma-3-1b-it	24.9	25.1	47.8	42.7	0.026	-0.019	-0.041	0.019	0.669	0.603	0.411	0.445	0.008	-0.189	-0.037	0.075

S1 Baseline Control (Table 24). Accuracy shifts are small and positive on average; HPSR falls modestly. PopGap and Corr trend toward zero, signaling less popularity-driven bias even without adversarial pressure. Alignment is largely stable with minor improvements. Spearman correlations move closer to zero for many models, with some variance by size.

S2 Popular Trap (Table 25). Substantial accuracy gains (commonly +6–12 points) and consistent HPSR drops indicate robust mitigation under trap conditions. PopGap shrinks and Corr moves toward zero, reducing systematic attraction to popular distractors. Alignment typically improves;

Table 25: Comprehensive debiasing results for **QASC: S2 Popular Trap**. An improvement in an 'After' cell is colored green; a regression is colored red.

Model	Accuracy (%)		HPSR (%)		PopGap		Corr		Confidence		Alignment		Spearman Conf Pop		Spearman Align Pop	
	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After
Llama-3.2-3B	46.8	58.1	38.5	26.6	0.213	0.106	-0.614	-0.537	0.632	0.565	0.536	0.561	-0.106	-0.039	-0.269	-0.023
Mistral-7B-Instructruct-v0.3	46.4	54.1	38.2	30.2	0.216	0.141	-0.602	-0.571	0.830	0.707	0.521	0.592	-0.174	-0.154	-0.421	-0.263
gemma-3-12b-it	24.9	35.9	49.4	38.0	0.292	0.186	-0.482	-0.494	0.754	0.621	0.377	0.475	0.034	0.083	-0.356	-0.183
phi-2	48.6	60.3	36.2	25.2	0.200	0.094	-0.597	-0.523	0.652	0.587	0.544	0.567	-0.097	-0.035	-0.310	-0.075
Meta-Llama-3-8B-Instructruct	53.1	59.2	35.4	29.4	0.190	0.132	-0.637	-0.605	0.869	0.754	0.564	0.622	-0.140	-0.155	-0.402	-0.299
Qwen2.5-7B	57.9	65.3	32.1	24.8	0.166	0.099	-0.628	-0.567	0.785	0.703	0.611	0.643	-0.162	-0.131	-0.398	-0.228
Qwen2.5-1.5B	49.2	59.5	36.0	25.8	0.196	0.102	-0.613	-0.545	0.690	0.621	0.561	0.590	-0.154	-0.086	-0.344	-0.117
Mistral-7B-Instructruct-v0.2	47.5	53.0	36.8	30.9	0.204	0.150	-0.590	-0.563	0.939	0.793	0.500	0.590	-0.170	-0.270	-0.451	-0.376
gemma-2-2b-it	28.9	43.0	46.9	33.0	0.281	0.152	-0.500	-0.482	0.634	0.543	0.447	0.500	-0.018	0.058	-0.256	-0.021
gemma-2-27b-it	29.0	36.8	45.8	38.1	0.271	0.195	-0.487	-0.488	0.875	0.710	0.348	0.471	-0.085	-0.076	-0.350	-0.256
falcon-7b-instruct	26.9	48.9	47.4	25.3	0.274	0.081	-0.488	-0.414	0.377	0.370	0.557	0.523	0.008	-0.319	0.375	0.497
Llama-3.1-8B-Instructruct	53.4	61.9	34.9	26.2	0.188	0.107	-0.619	-0.556	0.753	0.671	0.578	0.615	-0.143	-0.093	-0.361	-0.169
phi-4	52.5	60.0	35.5	28.4	0.190	0.122	-0.635	-0.594	0.816	0.716	0.571	0.623	-0.153	-0.137	-0.412	-0.258
gemma-2-9b-it	27.4	34.9	46.7	38.8	0.280	0.206	-0.484	-0.489	0.882	0.709	0.334	0.464	-0.035	-0.055	-0.382	-0.253
zephyr-7b-beta	44.4	51.0	39.2	31.9	0.219	0.153	-0.586	-0.559	0.882	0.740	0.483	0.571	-0.132	-0.161	-0.430	-0.313
Llama-2-7b-hf	31.1	53.5	45.9	23.3	0.259	0.061	-0.513	-0.404	0.369	0.368	0.563	0.519	-0.104	-0.342	0.440	0.498
Qwen2.5-3B	53.1	61.8	34.9	26.4	0.182	0.103	-0.620	-0.558	0.707	0.640	0.581	0.608	-0.176	-0.110	-0.359	-0.148
Llama-3.2-1B	32.8	50.2	45.9	28.2	0.272	0.115	-0.548	-0.506	0.518	0.464	0.517	0.523	-0.068	-0.029	-0.060	0.240
Llama-2-13b-hf	37.0	54.1	42.6	25.8	0.245	0.094	-0.547	-0.476	0.517	0.477	0.531	0.532	-0.090	-0.096	-0.035	0.209
gemma-3-4b-it	23.2	33.8	49.2	38.5	0.298	0.196	-0.455	-0.472	0.796	0.635	0.330	0.449	0.017	0.082	-0.343	-0.166
phi-1_5	37.5	52.0	41.7	26.8	0.239	0.108	-0.550	-0.490	0.558	0.502	0.509	0.527	-0.011	0.001	-0.125	0.113
Qwen2.5-0.5B	37.0	50.1	41.6	28.4	0.240	0.119	-0.549	-0.499	0.614	0.542	0.510	0.540	-0.097	-0.023	-0.259	0.021
gemma-3-1b-it	22.6	36.1	49.5	35.6	0.299	0.173	-0.449	-0.460	0.663	0.557	0.399	0.473	0.072	0.129	-0.297	-0.063

Table 26: Comprehensive debiasing results for **QASC: S3 Popularity Gradient**. An improvement in an 'After' cell is colored green; a regression is colored red.

Model	Accuracy (%)		HPSR (%)		PopGap		Corr		Confidence		Alignment		Spearman Conf Pop		Spearman Align Pop	
	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After
Llama-3.2-3B	46.9	59.3	38.5	26.1	0.218	0.105	-0.614	-0.541	0.635	0.569	0.540	0.563	-0.121	-0.045	-0.304	-0.050
Mistral-7B-Instructruct-v0.3	47.5	54.8	37.5	30.0	0.210	0.140	-0.600	-0.568	0.841	0.722	0.529	0.598	-0.134	-0.143	-0.445	-0.293
gemma-3-12b-it	24.5	35.8	49.7	38.2	0.295	0.188	-0.481	-0.497	0.755	0.616	0.367	0.465	0.023	0.108	-0.322	-0.126
phi-2	48.4	59.0	37.2	26.6	0.202	0.105	-0.610	-0.547	0.657	0.589	0.545	0.571	-0.127	-0.066	-0.329	-0.084
Meta-Llama-3-8B-Instructruct	51.9	58.8	35.9	29.3	0.198	0.133	-0.640	-0.607	0.865	0.749	0.564	0.626	-0.193	-0.191	-0.413	-0.308
Qwen2.5-7B	58.0	64.9	32.1	25.0	0.167	0.103	-0.635	-0.580	0.789	0.707	0.611	0.647	-0.169	-0.152	-0.401	-0.246
Qwen2.5-1.5B	49.0	59.0	36.2	26.4	0.198	0.107	-0.616	-0.556	0.693	0.620	0.559	0.590	-0.160	-0.090	-0.338	-0.107
Mistral-7B-Instructruct-v0.2	48.6	53.1	37.1	32.5	0.205	0.160	-0.602	-0.581	0.944	0.799	0.507	0.597	-0.114	-0.301	-0.481	-0.414
gemma-2-2b-it	30.9	44.5	51.0	37.0	0.282	0.157	-0.542	-0.532	0.639	0.539	0.450	0.508	0.067	0.120	-0.327	-0.033
gemma-2-27b-it	30.4	38.9	56.1	47.7	0.302	0.220	-0.560	-0.574	0.881	0.690	0.354	0.496	-0.019	-0.020	-0.506	-0.361
falcon-7b-instruct	22.9	48.8	59.5	32.9	0.314	0.091	-0.489	-0.470	0.378	0.371	0.565	0.527	0.019	-0.374	0.372	0.521
Llama-3.1-8B-Instructruct	52.0	61.3	36.5	26.7	0.195	0.110	-0.626	-0.564	0.755	0.667	0.582	0.620	-0.180	-0.113	-0.388	-0.185
phi-4	53.0	60.6	35.6	27.6	0.189	0.118	-0.628	-0.577	0.816	0.714	0.570	0.620	-0.154	-0.135	-0.434	-0.282
gemma-2-9b-it	29.1	37.8	56.5	46.9	0.311	0.224	-0.550	-0.565	0.889	0.693	0.341	0.489	0.012	-0.003	-0.477	-0.336
zephyr-7b-beta	45.3	52.0	37.5	30.7	0.216	0.152	-0.578	-0.553	0.888	0.749	0.489	0.576	-0.076	-0.148	-0.447	-0.320
Llama-2-7b-hf	29.1	55.8	55.3	27.7	0.295	0.063	-0.546	-0.462	0.371	0.369	0.568	0.515	-0.066	-0.348	0.411	0.493
Qwen2.5-3B	53.5	61.8	34.5	25.8	0.181	0.102	-0.619	-0.546	0.714	0.643	0.580	0.608	-0.157	-0.095	-0.368	-0.153
Llama-3.2-1B	33.1	50.6	44.6	27.0	0.273	0.113	-0.552	-0.502	0.515	0.468	0.524	0.528	-0.125	-0.071	-0.043	0.220
Llama-2-13b-hf	38.6	54.6	45.7	28.4	0.246	0.097	-0.575	-0.491	0.516	0.472	0.535	0.540	-0.114	-0.108	-0.030	0.242
gemma-3-4b-it	24.9	36.3	60.0	47.7	0.332	0.219	-0.526	-0.557	0.793	0.615	0.344	0.476	0.062	0.144	-0.484	-0.253
phi-1_5	36.5	51.0	44.9	30.1	0.250	0.118	-0.555	-0.502	0.562	0.501	0.510	0.532	-0.022	-0.006	-0.180	0.116
Qwen2.5-0.5B	39.0	51.5	38.4	25.6	0.230	0.113	-0.550	-0.492	0.610	0.541	0.506	0.534	-0.069	-0.008	-0.215	0.031
gemma-3-1b-it	25.8	40.0	57.1	41.8	0.320	0.184	-0.519	-0.536	0.663	0.546	0.413	0.493	0.085	0.158	-0.410	-0.097

Table 27: Comprehensive debiasing results for **QASC: S4 Direct Contest**. An improvement in an 'After' cell is colored green; a regression is colored red.

Model	Accuracy (%)		HPSR (%)		PopGap		Corr		Confidence		Alignment		Spearman Conf Pop		Spearman Align Pop	
	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After
Llama-3.2-3B	45.2	47.0	50.5	45.5	0.117	0.075	-0.352	-0.284	0.633	0.573	0.534	0.548	-0.016	-0.103	-0.139	-0.066
Mistral-7B-Instructruct-v0.3	44.8	45.4	49.4	47.0	0.112	0.093	-0.328	-0.295	0.833	0.747	0.508	0.535	-0.104	-0.371	-0.179	-0.175
gemma-3-12b-it	23.8	25.9	49.7	45.6	0.145	0.110	-0.244	-0.214	0.760	0.661	0.368	0.433	0.060	-0.152	-0.147	-0.011
phi-2	46.5	47.8	46.9	43.5	0.096	0.067	-0.309	-0.253	0.656	0.593	0.538	0.551	-0.001	-0.145	-0.177	-0.139
Meta-Llama-3-8B-Instructruct	52.1	52.6	48.5	47.5	0.101	0.092	-0.336	-0.321	0.866	0.775	0.565	0.579	-0.030	-0.436	-0.200	-0.276
Qwen2.5-7B	55.9	57.2	48.2	45.6	0.096	0.076	-0.384	-0.330	0.782	0.712	0.604	0.604	-0.058	-0.260	-0.196	-0.239
Qwen2.5-1.5B	48.2	49.9	49.2	45.4	0.105	0.073	-0.337	-0.279	0.690	0.626	0.555	0.566	-0.010	-0.131	-0.178	-0.136
Mistral-7B-Instructruct-v0.2	45.6	46.2	46.8	46.0	0.100	0.093	-0.303	-0.292	0.939	0.835	0.481	0.515	-0.045	-0.510	-0.201	-0.217
gemma-2-2b-it	30.4	32.2	47.1	41.7	0.114	0.073	-0.240	-0.184	0.635	0.565	0.450	0.488	0.035	-0.090	-0.151	-0.044
gemma-2-27b-it	28.8	29.1	45.8	44.6	0.113	0.104	-0.233	-0.220	0.880	0.773	0.340	0.406	-0.031	-0.369	-0.143	-0.048
falcon-7b-instruct	25.0	30.1	48.8	32.0	0.136	0.006	-0.228	-0.064	0.377	0.353	0.561	0.561	0.012	-0.144	0.147	0.285
Llama-3.1-8B-Instructruct	52.0	53.0	47.8	45.0	0.100	0.075	-0.331	-0.285	0.756	0.691	0.573	0.583	-0.014	-0.165	-0.171	-0.176
phi-4	52.5	52.8	47.8	46.2	0.097	0.084	-0.345	-0.317	0.815	0.736	0.564	0.575	-0.048	-0.338	-0.182	-0.231
gemma-2-9b-it	28.8	28.9	47.5	46.2	0.123	0.112	-0.246	-0.233	0.882	0.770	0.340	0.408	-0.006	-0.392	-0.147	-0.020
zephyr-7b-beta	42.0	42.5	49.0	47.5	0.110	0.098	-0.309	-0.292	0.882	0.781	0.470	0.509	-0.051	-0.438	-0.186	-0.165
Llama-2-7b-hf	30.2	35.3	48.9	31.0	0.128	-0.007	-0.264	-0.054	0.368	0.348	0.563	0.556	-0.038	-0.059	0.152	0.258
Qwen2.5-3B	52.8	54.4	49.5	46.3	0.101	0.073	-0.380	-0.324	0.699	0.638	0.573	0.579	-0.026	-0.157	-0.177	-0.153
Llama-3.2-1B	32.9	35.9	49.1	40.9	0.127	0.053	-0.285	-0.195	0.519	0.468	0.516	0.534	0.034	-0.018	-0.052	0.051
Llama-2-13b-hf	37.5	39.9	47.9	41.0	0.114	0.058	-0.276	-0.185	0.521	0.470	0.529	0.543	0.005	-0.074	-0.027	0.078
gemma-3-4b-it	24.2	24.9	46.4	44.4	0.122	0.106	-0.230	-0.214	0.801	0.700	0.341	0.408	-0.019	-0.301	-0.145	-0.006
phi-1_5	36.9	38.6	45.4	40.2	0.107	0.061	-0.262	-0.200	0.559	0.507	0.509	0.528	0.077	-0.012	-0.100	-0.024
Qwen2.5-0.5B	38.3	39.8	48.3	42.7	0.109	0.062	-0.282	-0.212	0.602	0.543	0.508	0.529	0.007	-0.087	-0.125	-0.031
gemma-3-1b-it	24.9	26.1	45.4	41.6	0.121	0.088	-0.220	-0.188	0.669	0.589	0.411	0.463	0.036	-0.152	-0.151	-0.026

confidence decreases slightly (more calibrated behavior). Spearman correlations tend toward zero for the majority of models.

S3 Popularity Gradient (Table 26). Results mirror S2 with graded difficulty: accuracy up, HPSR down, PopGap/Corr nearer zero across sizes. Improvements are especially pronounced for compact and mid-sized models, narrowing the gap to larger systems.

Table 28: Comprehensive debiasing results for **QASC: S5 Reverse Control**. An improvement in an 'After' cell is colored green; a regression is colored red.

Model	Accuracy (%)		HPSR (%)		PopGap		Corr		Confidence		Alignment		Spearman Conf Pop		Spearman Align Pop	
	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After
Llama-3.2-3B	47.6	47.2	54.2	53.1	-0.088	-0.093	0.188	0.204	0.619	0.605	0.552	0.549	0.069	0.079	0.167	0.152
Mistral-7B-Instructruct-v0.3	48.3	48.4	53.2	53.1	-0.089	-0.089	0.195	0.196	0.836	0.816	0.537	0.533	0.074	0.069	0.236	0.230
gemma-3-12b-it	25.6	25.4	48.5	48.0	-0.133	-0.136	0.197	0.200	0.751	0.713	0.374	0.374	0.026	-0.078	0.111	0.079
phi-2	47.4	47.4	51.3	51.0	-0.099	-0.101	0.225	0.233	0.640	0.629	0.561	0.559	0.096	0.111	0.166	0.167
Meta-Llama-3-8B-Instructruct	55.5	55.5	55.1	55.0	-0.070	-0.071	0.192	0.194	0.865	0.850	0.595	0.592	0.089	0.129	0.228	0.246
Qwen2.5-7B	60.5	60.5	54.6	54.4	-0.066	-0.068	0.223	0.232	0.782	0.768	0.636	0.630	0.119	0.126	0.199	0.202
Qwen2.5-1.5B	52.0	52.1	53.9	53.6	-0.082	-0.083	0.216	0.221	0.683	0.669	0.575	0.570	0.105	0.105	0.167	0.160
Mistral-7B-Instructruct-v0.2	48.1	48.1	52.8	52.8	-0.096	-0.096	0.208	0.208	0.937	0.918	0.504	0.503	0.051	0.025	0.224	0.213
gemma-2-2b-it	30.4	29.9	50.0	49.0	-0.121	-0.128	0.188	0.198	0.637	0.616	0.448	0.449	0.063	0.044	0.101	0.062
gemma-2-27b-it	27.1	27.0	48.7	48.4	-0.140	-0.142	0.206	0.210	0.881	0.854	0.329	0.333	0.014	-0.064	0.121	0.095
falcon-7b-instruct	25.8	23.4	46.4	43.2	-0.137	-0.161	0.206	0.231	0.376	0.366	0.561	0.569	0.006	-0.056	-0.107	-0.065
Llama-3.1-8B-Instructruct	54.4	54.5	54.4	54.2	-0.074	-0.075	0.206	0.212	0.745	0.730	0.602	0.596	0.111	0.115	0.220	0.216
phi-4	55.8	55.6	55.0	55.0	-0.071	-0.072	0.186	0.188	0.814	0.798	0.601	0.597	0.116	0.145	0.208	0.221
gemma-2-9b-it	29.4	29.4	48.5	48.4	-0.132	-0.133	0.200	0.202	0.879	0.850	0.349	0.350	0.069	-0.019	0.142	0.122
zephyr-7b-beta	46.0	46.0	53.8	53.5	-0.089	-0.091	0.181	0.185	0.883	0.861	0.506	0.504	0.022	0.000	0.233	0.214
Llama-2-7b-hf	31.8	31.4	51.6	49.0	-0.123	-0.136	0.188	0.222	0.365	0.358	0.564	0.564	-0.006	0.005	-0.135	-0.110
Qwen2.5-3B	57.0	57.0	54.0	53.9	-0.072	-0.073	0.230	0.233	0.698	0.685	0.609	0.602	0.111	0.111	0.178	0.174
Llama-3.2-1B	33.8	33.0	50.0	48.5	-0.126	-0.135	0.205	0.221	0.511	0.497	0.520	0.521	0.091	0.092	0.019	-0.020
Llama-2-13b-hf	41.1	40.7	51.4	50.6	-0.111	-0.115	0.233	0.245	0.514	0.502	0.535	0.534	0.072	0.075	0.043	0.020
gemma-3-4b-it	23.4	23.2	47.5	47.2	-0.142	-0.144	0.185	0.188	0.791	0.762	0.339	0.344	-0.030	-0.092	0.146	0.129
phi-1_5	37.0	36.9	50.0	49.1	-0.121	-0.125	0.215	0.226	0.544	0.530	0.525	0.523	0.138	0.132	0.059	0.035
Qwen2.5-0.5B	36.8	36.2	49.7	48.8	-0.121	-0.127	0.224	0.240	0.596	0.576	0.515	0.511	0.057	0.027	0.095	0.059
gemma-3-1b-it	23.2	23.1	47.3	47.0	-0.146	-0.149	0.190	0.195	0.673	0.651	0.402	0.406	0.037	-0.001	0.079	0.057

Table 29: Comprehensive debiasing results for **QASC: S6 None of the Above**. An improvement in an 'After' cell is colored green; a regression is colored red.

Model	Accuracy (%)		HPSR (%)		PopGap		Corr		Confidence		Alignment		Spearman Conf Pop		Spearman Align Pop	
	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After
Llama-3.2-3B	1.5	8.2	65.5	43.0	0.386	0.053	-0.423	-0.660	0.581	0.496	0.417	0.480	0.024	-0.004	0.004	0.249
Mistral-7B-Instructruct-v0.3	12.1	22.6	55.8	34.4	0.342	0.056	-0.808	-0.846	0.826	0.650	0.227	0.369	0.131	-0.000	-0.292	-0.072
gemma-3-12b-it	37.4	49.6	41.2	24.2	0.235	0.033	-0.925	-0.912	0.773	0.689	0.467	0.530	-0.169	-0.169	-0.666	-0.515
phi-2	8.0	21.1	58.6	37.4	0.319	0.043	-0.734	-0.824	0.595	0.511	0.393	0.442	0.163	0.219	-0.051	0.188
Meta-Llama-3-8B-Instructruct	16.8	32.6	54.1	31.1	0.328	0.045	-0.830	-0.855	0.780	0.633	0.285	0.391	0.154	0.123	-0.385	-0.201
Qwen2.5-7B	14.3	28.4	57.3	32.8	0.362	0.047	-0.837	-0.853	0.678	0.557	0.342	0.428	0.150	0.110	-0.261	0.010
Qwen2.5-1.5B	4.8	10.1	63.3	42.2	0.373	0.054	-0.666	-0.704	0.626	0.519	0.372	0.462	0.137	0.017	-0.108	0.174
Mistral-7B-Instructruct-v0.2	4.8	12.1	60.4	40.4	0.359	0.075	-0.642	-0.743	0.949	0.709	0.085	0.290	0.091	-0.196	-0.149	0.180
gemma-2-2b-it	38.1	48.5	40.0	24.8	0.224	0.037	-0.924	-0.911	0.566	0.511	0.442	0.477	0.201	0.046	-0.236	0.040
gemma-2-27b-it	16.4	26.8	53.6	35.8	0.304	0.058	-0.839	-0.856	0.834	0.676	0.244	0.372	0.178	0.019	-0.380	-0.205
falcon-7b-instruct	2.1	5.2	62.8	44.2	0.344	0.046	-0.456	-0.560	0.362	0.361	0.631	0.623	-0.017	-0.093	0.097	0.260
Llama-3.1-8B-Instructruct	6.4	15.7	61.3	38.4	0.363	0.052	-0.695	-0.762	0.679	0.550	0.316	0.420	0.164	0.098	-0.126	0.125
phi-4	12.2	22.4	58.7	36.7	0.363	0.056	-0.809	-0.823	0.784	0.624	0.269	0.398	0.112	0.013	-0.324	-0.086
gemma-2-9b-it	2.5	13.8	63.0	41.3	0.368	0.067	-0.524	-0.763	0.850	0.659	0.153	0.305	0.062	0.054	-0.070	0.102
zephyr-7b-beta	8.8	18.9	60.0	38.2	0.354	0.060	-0.771	-0.805	0.876	0.675	0.161	0.330	0.147	-0.005	-0.233	-0.017
Llama-2-7b-hf	13.2	26.1	56.4	34.1	0.311	0.030	-0.802	-0.837	0.340	0.340	0.611	0.572	0.151	0.061	0.303	0.552
Qwen2.5-3B	15.1	32.3	56.4	31.2	0.343	0.038	-0.840	-0.867	0.583	0.494	0.409	0.456	0.203	0.165	-0.127	0.224
Llama-3.2-1B	0.3	1.6	64.0	45.1	0.360	0.054	-0.229	-0.353	0.549	0.484	0.450	0.511	0.032	-0.015	-0.021	0.076
Llama-2-13b-hf	4.8	14.7	63.3	39.6	0.376	0.042	-0.664	-0.762	0.466	0.417	0.521	0.542	0.083	0.051	0.068	0.348
gemma-3-4b-it	6.8	18.6	60.2	39.0	0.349	0.055	-0.690	-0.796	0.711	0.573	0.298	0.399	0.130	0.115	-0.173	0.058
phi-1_5	22.4	41.7	50.6	27.4	0.290	0.030	-0.861	-0.879	0.507	0.460	0.459	0.467	0.244	0.222	0.076	0.330
Qwen2.5-0.5B	5.1	13.9	62.0	40.2	0.349	0.044	-0.640	-0.745	0.579	0.497	0.413	0.469	0.054	0.088	0.027	0.215
gemma-3-1b-it	16.2	31.1	54.1	33.4	0.301	0.042	-0.828	-0.859	0.634	0.546	0.375	0.432	0.211	0.196	-0.263	-0.043

S4 Direct Contest (Table 27). Moderate accuracy improvements and consistent HPSR reductions are observed. PopGap and Corr again compress toward zero. Alignment is stable or slightly higher; Spearman indicators typically drift toward zero with a few architecture-specific exceptions.

S5 Reverse Control (Table 28). As a sanity check, accuracy and alignment remain essentially unchanged. Minor shifts in PopGap/Corr appear but are small in magnitude, suggesting the method does not induce unintended bias reversals.

S6 None of the Above (Table 29). Large gains in accuracy and sharp HPSR declines show improved rejection of plausible but incorrect options. PopGap tightens considerably. Corr and Spearman terms sometimes move non-monotonically (calibration effects), while alignment remains stable or improves slightly.

F.3 Natural Questions (NQ)

S1 Baseline Control (Table 30). Accuracy improves slightly for most models; HPSR decreases reliably. PopGap and Corr move closer to zero, indicating less reliance on popularity priors. Alignment is stable with small gains; Spearman terms generally trend toward zero.

S2 Popular Trap (Table 31). Debiasing yields clear gains: accuracy increases substantially (particularly for smaller models), HPSR drops, and PopGap/Corr compress toward zero. Alignment typically improves or remains steady. Spearman correlations mostly move closer to zero, with occasional variability tied to model size.

Table 30: Comprehensive debiasing results for **NQ: S1 Baseline Control**. An improvement in an 'After' cell is colored green; a regression is colored red.

Model	Accuracy (%)		HPSR (%)		PopGap		Corr		Confidence		Alignment		Spearman Conf Pop		Spearman Align Pop	
	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After
Llama-3.2-3B	48.5	49.8	48.3	43.1	0.057	0.008	-0.211	-0.081	0.527	0.493	0.572	0.569	0.162	0.145	-0.028	-0.025
Mistral-7B-Instructruct-v0.3	55.5	55.8	46.5	44.4	0.048	0.029	-0.184	-0.127	0.835	0.777	0.603	0.604	0.047	-0.187	-0.049	-0.089
gemma-3-12b-it	27.9	28.3	45.5	41.9	0.067	0.033	-0.139	-0.087	0.734	0.660	0.399	0.439	0.018	-0.194	-0.074	0.042
phi-2	42.9	44.0	46.6	41.0	0.053	0.004	-0.152	-0.050	0.530	0.494	0.551	0.553	0.052	-0.000	-0.002	0.018
Meta-Llama-3-8B-Instructruct	56.0	56.2	45.9	45.1	0.034	0.027	-0.153	-0.132	0.812	0.761	0.620	0.615	0.093	-0.149	-0.068	-0.133
Qwen2.5-7B	55.9	57.4	49.7	46.0	0.064	0.031	-0.250	-0.162	0.682	0.630	0.620	0.615	0.100	0.012	-0.066	-0.050
Qwen2.5-1.5B	46.2	47.1	48.1	42.6	0.063	0.015	-0.180	-0.069	0.574	0.533	0.581	0.582	0.095	0.042	0.008	0.059
Mistral-7B-Instructruct-v0.2	53.5	53.6	46.9	46.6	0.043	0.040	-0.171	-0.165	0.935	0.862	0.558	0.563	0.020	-0.468	-0.055	-0.099
gemma-2-2b-it	33.1	34.5	46.5	41.0	0.072	0.020	-0.154	-0.070	0.570	0.518	0.496	0.514	0.057	-0.035	-0.061	0.010
gemma-2-27b-it	40.0	40.2	43.5	42.8	0.040	0.034	-0.097	-0.081	0.850	0.786	0.451	0.470	0.068	-0.315	-0.071	0.006
falcon-7b-instruct	24.1	26.8	43.0	29.2	0.056	-0.071	-0.083	0.107	0.387	0.369	0.561	0.562	0.038	-0.163	0.026	0.110
gemma-2-2b-it_mcq_results_new_popularity.json	33.1	34.5	46.5	41.0	0.072	0.020	-0.154	-0.070	0.570	0.518	0.496	0.514	0.057	-0.035	-0.061	0.010
gemma-2-2b-it_mcq_results_frequency.json	33.1	34.5	46.5	41.0	0.072	0.020	-0.154	-0.070	0.570	0.518	0.496	0.514	0.057	-0.035	-0.061	0.010
Llama-3.1-8B-Instructruct	58.5	59.3	47.4	44.4	0.037	0.010	-0.178	-0.089	0.702	0.660	0.628	0.622	0.067	-0.008	-0.056	-0.068
phi-4	55.4	56.5	48.0	45.4	0.054	0.030	-0.181	-0.116	0.784	0.726	0.614	0.612	0.075	-0.113	-0.090	-0.121
gemma-2-9b-it	34.8	35.5	45.8	44.3	0.064	0.049	-0.135	-0.114	0.832	0.746	0.420	0.452	0.130	-0.238	-0.108	0.024
zephyr-7b-beta	53.5	53.6	48.4	46.9	0.054	0.041	-0.206	-0.175	0.858	0.788	0.582	0.587	0.043	-0.288	-0.071	-0.113
Llama-2-7b-hf	31.5	31.9	45.2	37.4	0.053	-0.016	-0.128	-0.009	0.433	0.402	0.536	0.548	0.070	-0.037	0.018	0.048
Qwen2.5-3B	51.2	52.9	49.5	44.7	0.066	0.021	-0.213	-0.092	0.593	0.551	0.601	0.594	0.075	0.033	-0.030	-0.011
Llama-3.2-1B	36.5	37.0	52.2	40.2	0.097	-0.002	-0.237	-0.027	0.416	0.385	0.558	0.563	0.160	0.167	0.007	-0.019
Llama-2-13b-hf	49.0	50.9	47.0	40.5	0.050	-0.009	-0.181	-0.019	0.471	0.443	0.563	0.554	0.073	0.042	0.063	0.041
gemma-3-4b-it	28.1	28.3	46.2	44.5	0.066	0.050	-0.143	-0.120	0.776	0.699	0.381	0.424	0.001	-0.294	-0.086	0.073
phi-1_5	31.4	33.9	48.1	36.7	0.074	-0.032	-0.165	0.018	0.408	0.377	0.550	0.551	0.066	0.006	0.044	0.082
Qwen2.5-0.5B	38.3	40.2	49.1	40.2	0.080	0.002	-0.182	-0.014	0.461	0.425	0.547	0.548	0.093	0.060	0.058	0.086
gemma-3-1b-it	24.1	24.8	44.9	41.1	0.060	0.028	-0.111	-0.062	0.665	0.603	0.409	0.450	0.129	-0.085	-0.148	0.013

Table 31: Comprehensive debiasing results for **NQ: S2 Popular Trap**. An improvement in an 'After' cell is colored green; a regression is colored red.

Model	Accuracy (%)		HPSR (%)		PopGap		Corr		Confidence		Alignment		Spearman Conf Pop		Spearman Align Pop	
	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After
Llama-3.2-3B	48.4	60.0	33.7	21.6	0.168	0.062	-0.609	-0.526	0.532	0.506	0.575	0.565	-0.007	0.022	-0.039	0.158
Mistral-7B-Instructruct-v0.3	56.8	61.7	29.5	24.3	0.144	0.096	-0.645	-0.607	0.832	0.754	0.617	0.651	-0.118	-0.126	-0.360	-0.228
gemma-3-12b-it	27.5	36.9	45.9	35.6	0.245	0.152	-0.517	-0.521	0.742	0.630	0.398	0.476	-0.024	0.003	-0.294	-0.091
phi-2	45.5	58.3	37.0	23.7	0.187	0.072	-0.624	-0.556	0.548	0.517	0.559	0.553	-0.119	-0.082	-0.098	0.092
Meta-Llama-3-8B-Instructruct	58.5	62.0	27.7	24.0	0.118	0.085	-0.626	-0.592	0.820	0.753	0.633	0.657	-0.137	-0.164	-0.343	-0.251
Qwen2.5-7B	57.0	65.2	29.9	21.8	0.146	0.072	-0.658	-0.600	0.690	0.643	0.627	0.632	-0.109	-0.037	-0.255	-0.054
Qwen2.5-1.5B	47.0	59.5	36.2	23.3	0.189	0.073	-0.632	-0.556	0.587	0.547	0.575	0.573	-0.065	0.005	-0.155	0.077
Mistral-7B-Instructruct-v0.2	56.4	59.0	28.7	25.9	0.135	0.110	-0.626	-0.607	0.934	0.832	0.587	0.638	-0.168	-0.284	-0.379	-0.324
gemma-2-2b-it	33.5	45.1	42.1	30.1	0.230	0.118	-0.550	-0.534	0.569	0.509	0.500	0.524	0.035	0.040	-0.175	0.072
gemma-2-27b-it	39.7	45.3	38.7	32.8	0.199	0.144	-0.577	-0.564	0.856	0.738	0.448	0.526	0.010	-0.093	-0.383	-0.246
falcon-7b-instruct	26.6	44.9	46.2	24.7	0.243	0.070	-0.492	-0.435	0.387	0.378	0.554	0.531	0.002	-0.285	0.300	0.415
Llama-3.1-8B-Instructruct	61.2	67.3	26.5	20.1	0.115	0.056	-0.641	-0.574	0.708	0.669	0.646	0.649	-0.100	-0.034	-0.242	-0.084
phi-4	59.0	63.6	28.6	23.6	0.128	0.084	-0.646	-0.602	0.790	0.727	0.632	0.654	-0.105	-0.104	-0.315	-0.190
gemma-2-9b-it	33.6	39.6	42.8	36.0	0.226	0.164	-0.549	-0.550	0.831	0.705	0.406	0.493	0.132	0.011	-0.423	-0.231
zephyr-7b-beta	55.5	59.5	30.9	26.8	0.147	0.109	-0.642	-0.621	0.859	0.770	0.590	0.632	-0.099	-0.161	-0.362	-0.255
Llama-2-7b-hf	29.8	46.6	44.8	26.2	0.233	0.075	-0.528	-0.478	0.436	0.416	0.538	0.526	-0.005	-0.118	0.213	0.358
Qwen2.5-3B	53.4	63.5	30.9	20.4	0.157	0.063	-0.625	-0.533	0.600	0.563	0.603	0.600	-0.093	-0.043	-0.167	0.039
Llama-3.2-1B	34.4	52.0	43.5	24.4	0.237	0.073	-0.578	-0.504	0.419	0.402	0.552	0.525	0.112	-0.014	0.124	0.285
Llama-2-13b-hf	48.5	61.3	33.2	20.1	0.160	0.048	-0.601	-0.484	0.468	0.452	0.562	0.544	-0.088	-0.103	0.116	0.240
gemma-3-4b-it	24.5	33.6	47.4	38.0	0.257	0.169	-0.485	-0.504	0.782	0.648	0.356	0.455	-0.009	0.005	-0.300	-0.126
phi-1_5	31.9	48.0	43.8	26.6	0.230	0.080	-0.546	-0.505	0.408	0.391	0.551	0.532	0.049	-0.157	0.205	0.341
Qwen2.5-0.5B	34.4	50.4	42.7	25.5	0.232	0.086	-0.566	-0.517	0.465	0.438	0.556	0.542	-0.001	-0.078	0.046	0.244
gemma-3-1b-it	22.9	34.1	48.9	36.6	0.264	0.155	-0.478	-0.490	0.666	0.566	0.409	0.478	0.158	0.149	-0.330	-0.076

Table 32: Comprehensive debiasing results for **NQ: S3 Popularity Gradient**. An improvement in an 'After' cell is colored green; a regression is colored red.

Model	Accuracy (%)		HPSR (%)		PopGap		Corr		Confidence		Alignment		Spearman Conf Pop		Spearman Align Pop	
	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After
Llama-3.2-3B	49.9	61.9	34.8	22.4	0.166	0.062	-0.637	-0.558	0.530	0.504	0.568	0.559	0.010	0.010	-0.033	0.136
Mistral-7B-Instructruct-v0.3	57.7	62.2	29.8	25.0	0.141	0.098	-0.648	-0.615	0.845	0.762	0.610	0.647	-0.087	-0.130	-0.348	-0.237
gemma-3-12b-it	27.9	37.6	46.9	36.1	0.245	0.147	-0.517	-0.525	0.737	0.627	0.396	0.474	-0.076	-0.038	-0.221	-0.063
phi-2	44.2	57.6	38.0	24.6	0.186	0.069	-0.604	-0.533	0.549	0.516	0.559	0.553	-0.151	-0.091	-0.096	0.100
Meta-Llama-3-8B-Instructruct	60.2	64.5	27.5	23.2	0.125	0.084	-0.642	-0.609	0.823	0.755	0.640	0.664	-0.142	-0.159	-0.336	-0.241
Qwen2.5-7B	56.7	65.7	31.9	22.1	0.157	0.071	-0.684	-0.612	0.689	0.641	0.627	0.632	-0.144	-0.040	-0.267	-0.056
Qwen2.5-1.5B	46.6	58.7	35.9	23.6	0.188	0.076	-0.627	-0.562	0.588	0.548	0.574	0.577	-0.050	-0.008	-0.136	0.089
Mistral-7B-Instructruct-v0.2	56.7	59.2	30.2	27.6	0.140	0.115	-0.635	-0.619	0.938	0.834	0.590	0.641	-0.116	-0.252	-0.359	-0.303
gemma-2-2b-it	33.1	46.7	48.0	33.6	0.248	0.116	-0.578	-0.565	0.567	0.505	0.503	0.525	0.058	0.068	-0.191	0.084
gemma-2-27b-it	42.0	47.5	44.9	39.2	0.218	0.164	-0.642	-0.639	0.851	0.719	0.465	0.547	0.067	-0.059	-0.437	-0.272
falcon-7b-instruct	23.8	46.5	59.1	35.4	0.284	0.085	-0.520	-0.519	0.387	0.378	0.561	0.530	0.043	-0.302	0.308	0.448
Llama-3.1-8B-Instructruct	61.2	68.2	28.9	21.6	0.125	0.061	-0.654	-0.585	0.710	0.666	0.647	0.649	-0.114	-0.053	-0.274	-0.106
phi-4	59.1	63.6	30.3	25.6	0.131	0.089	-0.666	-0.637	0.790	0.725	0.629	0.652	-0.099	-0.112	-0.334	-0.204
gemma-2-9b-it	33.9	41.5	50.1	42.1	0.245	0.172	-0.597	-0.604	0.836	0.695	0.405	0.501	0.195	0.065	-0.441	-0.229
zephyr-7b-beta	55.5	59.5	30.4	26.4	0.145	0.107	-0.621	-0.595	0.867	0.779	0.589	0.633	-0.076	-0.156	-0.346	-0.253
Llama-2-7b-hf	29.9	49.1	56.1	35.6	0.268	0.089	-0.578	-0.564	0.437	0.413	0.539	0.525	0.023	-0.099	0.238	0.384
Qwen2.5-3B	53.2	64.6	31.7	20.1	0.161	0.059	-0.649	-0.560	0.602	0.567	0.605	0.596	-0.102	-0.019	-0.140	0.065
Llama-3.2-1B	37.0	54.7	44.1	24.9	0.234	0.070	-0.600	-0.532	0.418	0.403	0.548	0.519	0.086	-0.033	0.161	0.290
Llama-2-13b-hf	47.5	61.1	38.8	24.4	0.181	0.056	-0.653	-0.553	0.469	0.453	0.563	0.544	-0.112	-0.132	0.070	0.204
gemma-3-4b-it	26.1	36.5	57.4	46.3	0.290	0.189	-0.547	-0.573	0.785	0.629	0.359	0.473	0.087	0.077	-0.390	-0.171
phi-1_5	33.1	50.1	49.0	30.6	0.243	0.083	-0.575	-0.522	0.408	0.390	0.544	0.525	0.096	-0.122	0.195	0.343
Qwen2.5-0.5B	36.4	52.0	37.4	20.8	0.220	0.074	-0.570	-0.508	0.468	0.443	0.550	0.536	-0.017	-0.063	0.142	0.276
gemma-3-1b-it	25.6	37.5	55.1	42.2	0.289	0.169	-0.529	-0.549	0.670	0.553	0.415	0.498	0.219	0.171	-0.392	-0.086

S3 Popularity Gradient (Table 32). Under graded pressure, we again see accuracy up, HPSR down, and PopGap/Corr nearer zero across the board. Smaller models benefit most in relative terms; larger models consolidate modest but consistent improvements.

S4 Direct Contest (Table 33). Improvements are more modest but directionally consistent: small accuracy gains, lower HPSR, and reduced PopGap/Corr. Alignment remains steady. Spearman measures generally drift toward zero.

Table 33: Comprehensive debiasing results for **NQ: S4 Direct Contest**. An improvement in an 'After' cell is colored green; a regression is colored red.

Model	Accuracy (%)		HPSR (%)		PopGap		Corr		Confidence		Alignment		Spearman Conf Pop		Spearman Align Pop	
	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After
Llama-3.2-3B	47.5	51.5	39.8	32.9	0.092	0.033	-0.371	-0.255	0.533	0.500	0.565	0.562	0.112	0.091	-0.031	0.042
Mistral-7B-Instructruct-v0.3	57.0	57.8	39.4	37.4	0.077	0.059	-0.420	-0.380	0.836	0.773	0.604	0.615	-0.003	-0.201	-0.175	-0.205
gemma-3-12b-it	27.3	28.9	45.0	40.7	0.135	0.096	-0.321	-0.286	0.743	0.655	0.390	0.446	0.012	-0.158	-0.143	-0.002
phi-2	42.3	46.1	40.2	33.1	0.099	0.037	-0.318	-0.216	0.538	0.503	0.544	0.541	-0.022	-0.007	-0.021	0.038
Meta-Llama-3-8B-Instructruct	56.1	57.2	38.5	36.9	0.072	0.056	-0.381	-0.351	0.820	0.760	0.611	0.617	0.005	-0.177	-0.156	-0.183
Qwen2.5-7B	56.1	58.8	41.6	37.6	0.096	0.058	-0.457	-0.381	0.682	0.637	0.621	0.620	0.014	-0.011	-0.150	-0.091
Qwen2.5-1.5B	46.1	49.0	40.5	34.9	0.105	0.054	-0.356	-0.272	0.578	0.538	0.568	0.569	0.099	0.057	-0.032	0.053
Mistral-7B-Instructruct-v0.2	53.9	54.5	38.2	37.4	0.076	0.069	-0.371	-0.356	0.931	0.863	0.562	0.580	-0.060	-0.389	-0.175	-0.218
gemma-2-2b-it	31.6	35.4	45.6	37.5	0.132	0.062	-0.345	-0.263	0.573	0.522	0.491	0.511	0.031	-0.001	-0.126	-0.011
gemma-2-27b-it	39.8	40.6	41.5	39.6	0.096	0.080	-0.335	-0.313	0.850	0.774	0.448	0.485	0.023	-0.244	-0.221	-0.167
falcon-7b-instruct	25.1	32.9	46.2	30.0	0.143	0.003	-0.303	-0.145	0.387	0.371	0.556	0.548	0.039	-0.162	0.160	0.260
Llama-3.1-8B-Instructruct	57.6	59.5	39.3	35.9	0.076	0.044	-0.426	-0.348	0.704	0.658	0.628	0.627	0.035	-0.031	-0.126	-0.101
phi-4	56.1	57.5	39.5	36.5	0.087	0.059	-0.422	-0.366	0.780	0.722	0.617	0.625	-0.012	-0.128	-0.160	-0.154
gemma-2-9b-it	32.6	33.0	42.1	40.6	0.112	0.099	-0.305	-0.287	0.831	0.747	0.403	0.455	0.121	-0.185	-0.217	-0.099
zephyr-7b-beta	52.5	53.6	40.7	39.0	0.089	0.071	-0.409	-0.382	0.859	0.794	0.571	0.585	0.001	-0.193	-0.180	-0.186
Llama-2-7b-hf	29.6	35.0	44.1	32.1	0.122	0.021	-0.305	-0.184	0.433	0.406	0.544	0.542	0.037	0.029	0.070	0.174
Qwen2.5-3B	50.7	54.5	42.0	36.5	0.100	0.048	-0.418	-0.326	0.591	0.553	0.602	0.596	0.032	0.014	-0.088	-0.003
Llama-3.2-1B	35.0	41.6	46.4	33.2	0.144	0.034	-0.374	-0.222	0.419	0.392	0.555	0.543	0.119	0.125	0.048	0.128
Llama-2-13b-hf	46.9	51.3	41.1	31.5	0.092	0.013	-0.377	-0.213	0.466	0.442	0.561	0.553	-0.010	-0.021	0.036	0.092
gemma-3-4b-it	27.0	28.3	44.0	41.5	0.127	0.103	-0.284	-0.262	0.778	0.688	0.376	0.438	-0.005	-0.251	-0.168	-0.014
phi-1_5	31.1	37.5	45.4	33.4	0.132	0.026	-0.332	-0.207	0.409	0.382	0.548	0.541	0.071	0.016	0.141	0.211
Qwen2.5-0.5B	35.6	41.3	43.6	33.1	0.130	0.037	-0.354	-0.225	0.463	0.430	0.557	0.549	0.050	0.037	0.061	0.137
gemma-3-1b-it	24.6	26.4	44.6	40.8	0.128	0.092	-0.284	-0.254	0.664	0.583	0.412	0.470	0.191	-0.013	-0.244	-0.051

Table 34: Comprehensive debiasing results for **NQ: S5 Reverse Control**. An improvement in an 'After' cell is colored green; a regression is colored red.

Model	Accuracy (%)		HPSR (%)		PopGap		Corr		Confidence		Alignment		Spearman Conf Pop		Spearman Align Pop	
	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After
Llama-3.2-3B	51.5	51.6	49.1	48.6	-0.029	-0.034	0.104	0.119	0.525	0.515	0.592	0.586	0.231	0.214	0.072	0.056
Mistral-7B-Instructruct-v0.3	57.9	57.8	48.6	48.5	-0.034	-0.037	0.134	0.142	0.827	0.809	0.624	0.616	0.123	0.015	0.168	0.116
gemma-3-12b-it	28.6	28.4	46.4	45.5	-0.057	-0.065	0.124	0.136	0.741	0.711	0.401	0.405	0.037	-0.062	0.067	0.082
phi-2	43.2	43.0	48.0	46.8	-0.034	-0.046	0.132	0.157	0.520	0.507	0.560	0.557	0.104	0.079	0.083	0.076
Meta-Llama-3-8B-Instructruct	61.2	61.3	48.4	48.4	-0.029	-0.029	0.156	0.158	0.808	0.795	0.658	0.650	0.165	0.099	0.179	0.129
Qwen2.5-7B	60.5	60.7	51.1	50.8	-0.008	-0.011	0.053	0.062	0.684	0.663	0.655	0.642	0.182	0.120	0.145	0.105
Qwen2.5-1.5B	47.1	46.6	49.5	47.9	-0.020	-0.035	0.077	0.114	0.571	0.554	0.596	0.591	0.167	0.145	0.128	0.124
Mistral-7B-Instructruct-v0.2	55.0	54.9	48.0	47.9	-0.036	-0.037	0.142	0.146	0.931	0.907	0.575	0.568	0.064	-0.122	0.162	0.111
gemma-2-2b-it	33.1	32.6	46.7	45.0	-0.050	-0.064	0.129	0.151	0.567	0.545	0.495	0.495	0.068	0.010	0.048	0.038
gemma-2-27b-it	39.5	39.5	44.5	44.4	-0.059	-0.060	0.176	0.179	0.844	0.819	0.449	0.448	0.129	-0.010	0.033	0.046
falcon-7b-instruct	25.6	24.9	42.4	38.4	-0.069	-0.104	0.157	0.202	0.385	0.377	0.558	0.562	0.072	-0.040	-0.090	0.006
Llama-3.1-8B-Instructruct	62.2	62.0	48.2	48.0	-0.028	-0.030	0.121	0.127	0.695	0.679	0.661	0.650	0.160	0.108	0.142	0.100
phi-4	60.5	60.5	49.8	49.0	-0.024	-0.027	0.107	0.126	0.784	0.767	0.650	0.644	0.186	0.119	0.204	0.167
gemma-2-9b-it	35.6	35.5	45.3	45.0	-0.050	-0.053	0.144	0.148	0.834	0.801	0.425	0.425	0.192	-0.005	0.030	0.067
zephyr-7b-beta	56.3	56.1	49.0	48.6	-0.031	-0.034	0.124	0.131	0.850	0.830	0.605	0.597	0.109	-0.017	0.173	0.121
Llama-2-7b-hf	29.9	29.1	44.1	42.0	-0.066	-0.087	0.156	0.184	0.430	0.420	0.541	0.543	0.027	-0.021	-0.001	0.021
Qwen2.5-3B	55.0	54.9	49.8	48.9	-0.019	-0.027	0.079	0.101	0.590	0.574	0.626	0.617	0.178	0.143	0.134	0.120
Llama-3.2-1B	39.0	38.1	51.6	49.1	-0.025	-0.045	0.063	0.103	0.414	0.403	0.567	0.566	0.172	0.148	-0.018	-0.022
Llama-2-13b-hf	50.2	49.8	46.7	44.9	-0.046	-0.063	0.179	0.218	0.460	0.447	0.567	0.561	0.090	0.068	0.010	-0.018
gemma-3-4b-it	26.6	26.6	44.9	44.4	-0.067	-0.071	0.146	0.153	0.774	0.743	0.381	0.386	-0.020	-0.154	0.095	0.134
phi-1_5	31.2	31.1	45.8	42.9	-0.052	-0.080	0.128	0.172	0.405	0.395	0.557	0.556	0.047	0.017	-0.002	0.019
Qwen2.5-0.5B	36.0	36.0	50.0	47.1	-0.028	-0.054	0.083	0.129	0.458	0.444	0.572	0.567	0.100	0.079	0.058	0.063
gemma-3-1b-it	24.9	24.4	43.5	42.9	-0.070	-0.077	0.149	0.156	0.665	0.638	0.424	0.433	0.116	0.003	-0.032	0.004

Table 35: Comprehensive debiasing results for **NQ: S6 None of the Above**. An improvement in an 'After' cell is colored green; a regression is colored red.

Model	Accuracy (%)		HPSR (%)		PopGap		Corr		Confidence		Alignment		Spearman Conf Pop		Spearman Align Pop	
	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After
Llama-3.2-3B	0.1	8.7	66.3	38.0	0.433	0.087	-0.104	-0.567	0.500	0.424	0.499	0.550	0.148	-0.056	-0.144	0.333
Mistral-7B-Instructruct-v0.3	18.6	33.9	54.8	31.1	0.347	0.078	-0.833	-0.828	0.801	0.627	0.294	0.424	0.166	0.032	-0.463	-0.198
gemma-3-12b-it	26.1	38.2	49.0	28.9	0.311	0.070	-0.864	-0.851	0.722	0.609	0.406	0.492	-0.015	-0.130	-0.524	-0.255
phi-2	6.0	22.2	59.2	32.8	0.360	0.066	-0.645	-0.768	0.490	0.434	0.497	0.517	0.140	0.063	0.007	0.385
Meta-Llama-3-8B-Instructruct	26.6	44.6	48.6	26.6	0.292	0.053	-0.855	-0.853	0.722	0.602	0.381	0.457	0.115	0.053	-0.506	-0.227
Qwen2.5-7B	8.8	25.9	64.0	34.7	0.410	0.076	-0.728	-0.787	0.611	0.496	0.400	0.476	0.125	0.036	-0.192	0.227
Qwen2.5-1.5B	9.3	22.9	61.0	34.8	0.384	0.073	-0.731	-0.762	0.521	0.443	0.472	0.518	0.157	-0.034	-0.073	0.382
Mistral-7B-Instructruct-v0.2	8.0	19.6	62.1	39.6	0.369	0.103	-0.689	-0.749	0.928	0.689	0.129	0.318	0.126	-0.068	-0.256	0.027
gemma-2-2b-it	56.8	69.2	28.6	13.8	0.184	0.029	-0.929	-0.901	0.548	0.513	0.495	0.510	0.019	-0.120	-0.229	0.038
gemma-2-27b-it	25.0	39.1	47.9	27.8	0.294	0.065	-0.839	-0.836	0.778	0.640	0.335	0.432	0.164	0.026	-0.505	-0.276
falcon-7b-instruct	0.1	6.6	64.8	39.6	0.388	0.079	-0.122	-0.524	0.386	0.383	0.614	0.604	0.037	-0.275	-0.030	0.375
Llama-3.1-8B-Instructruct	17.8	35.9	54.5	29.1	0.337	0.058	-0.812	-0.824	0.601	0.501	0.404	0.464	0.195	0.068	-0.197	0.194
phi-4	26.8	45.3	49.6	25.6	0.316	0.058	-0.871	-0.861	0.727	0.602	0.410	0.492	0.008	-0.038	-0.548	-0.231
gemma-2-9b-it	5.0	17.4	61.2	36.8	0.403	0.096	-0.583	-0.707	0.807	0.599	0.200	0.361	0.150	0.089	-0.172	0.136
zephyr-7b-beta	12.9	30.6	58.5	33.7	0.370	0.083	-0.777	-0.820	0.822	0.622	0.234	0.377	0.240	0.139	-0.407	-0.117
Llama-2-7b-hf	4.9	18.2	62.4	35.0	0.386	0.064	-0.578	-0.703	0.382	0.370	0.600	0.585	0.149	-0.100	0.062	0.448
Qwen2.5-3B	9.9	29.9	62.6	31.6	0.416	0.064	-0.729	-0.794	0.512	0.430	0.477	0.512	0.200	0.001	-0.086	0.449
Llama-3.2-1B	0.0	5.9	69.1	37.8	0.465	0.088	0.000	-0.493	0.450	0.395	0.550	0.589	0.144	-0.258	-0.144	0.399
Llama-2-13b-hf	12.8	30.5	59.4	31.1	0.374	0.055	-0.789	-0.805	0.404	0.378	0.562	0.551	0.132	-0.166	0.199	0.608
gemma-3-4b-it	12.3	28.8	56.7	32.5	0.363	0.083	-0.768	-0.814	0.715	0.567	0.307	0.410	0.151	0.134	-0.250	0.021
phi-1_5	22.5	46.5	52.6	23.1	0.347	0.036	-0.853	-0.853	0.392	0.377	0.544	0.506	0.126	-0.106	0.410	0.691
Qwen2.5-0.5B	1.9	13.7	62.3	34.7	0.412	0.073	-0.420	-0.655	0.464	0.411	0.531	0.558	0.045	-0.140	0.010	0.417
gemma-3-1b-it	23.4	42.2	50.0	26.6	0.309	0.059	-0.849	-0.856	0.603	0.517	0.414	0.463	0.189	0.086	-0.278	0.084

S5 Reverse Control (Table 34). No substantive accuracy change is expected or observed. Popularity metrics show only minor, symmetric movement, supporting the claim that the procedure does not inject reverse popularity bias.

S6 None of the Above (Table 35). Accuracy gains are large and widespread, with strong HPSR reductions. PopGap compresses markedly; Corr and Spearman terms sometimes fluctuate as models become more cautious, while alignment typically holds or improves.

F.4 MuSiQue — Explanations for Tables 36–41

Table 36: Comprehensive debiasing results for **MuSiQue: S1 Baseline Control**. An improvement in an 'After' cell is colored green; a regression is colored red.

Model	Accuracy (%)		HPSR (%)		PopGap		Corr		Confidence		Alignment		Spearman Conf Pop		Spearman Align Pop	
	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After
Llama-3.2-3B	42.0	44.1	50.4	39.2	0.069	-0.023	-0.212	0.009	0.452	0.418	0.554	0.552	0.169	0.086	0.087	0.094
Mistral-7B-Instructruct-v0.3	47.2	47.8	46.3	43.5	0.037	0.012	-0.117	-0.052	0.707	0.649	0.554	0.558	0.093	-0.116	-0.027	-0.005
gemma-3-12b-it	27.3	27.9	48.0	44.0	0.069	0.031	-0.134	-0.079	0.715	0.641	0.407	0.448	-0.018	-0.270	-0.061	0.084
phi-2	42.1	43.4	43.9	37.5	0.022	-0.037	-0.081	0.058	0.505	0.474	0.539	0.539	0.050	-0.033	0.019	0.002
Meta-Llama-3-8B-Instructruct	49.6	49.9	48.9	46.6	0.053	0.036	-0.186	-0.126	0.750	0.691	0.558	0.561	0.129	-0.121	-0.040	-0.047
Qwen2.5-7B	49.5	50.1	51.2	46.0	0.070	0.027	-0.230	-0.103	0.554	0.505	0.573	0.572	0.135	-0.007	0.030	0.043
Qwen2.5-1.5B	42.3	42.9	48.5	39.8	0.064	-0.017	-0.167	0.000	0.471	0.432	0.557	0.558	0.148	0.042	0.056	0.050
Mistral-7B-Instructruct-v0.2	42.8	42.8	45.8	45.0	0.043	0.036	-0.117	-0.103	0.880	0.801	0.477	0.494	0.064	-0.442	-0.056	-0.013
gemma-2-2b-it	33.6	35.0	48.1	42.9	0.060	0.014	-0.174	-0.090	0.569	0.517	0.494	0.513	0.051	-0.137	-0.078	-0.001
gemma-2-27b-it	39.8	40.2	46.0	43.5	0.040	0.014	-0.132	-0.084	0.777	0.705	0.463	0.482	0.098	-0.229	-0.066	-0.003
falcon-7b-instruct	25.4	28.3	46.5	31.9	0.053	-0.078	-0.114	0.093	0.395	0.374	0.554	0.557	0.033	-0.279	0.084	0.233
Llama-3.1-8B-Instructruct	48.8	50.2	45.1	40.2	0.034	-0.008	-0.115	0.003	0.588	0.545	0.580	0.575	0.123	0.020	0.019	0.011
phi-4	53.0	53.4	46.5	44.0	0.035	0.013	-0.111	-0.054	0.700	0.647	0.591	0.587	0.116	-0.102	0.029	0.002
gemma-2-9b-it	32.1	32.2	45.5	44.1	0.049	0.034	-0.118	-0.097	0.799	0.710	0.403	0.443	0.097	-0.336	-0.128	0.023
zephyr-7b-beta	44.0	44.5	48.0	45.2	0.057	0.032	-0.167	-0.118	0.761	0.692	0.525	0.536	0.036	-0.246	-0.055	-0.019
Llama-2-7b-hf	31.8	33.4	47.1	30.6	0.052	-0.093	-0.131	0.129	0.355	0.338	0.560	0.560	0.008	-0.108	0.120	0.133
Qwen2.5-3B	45.0	45.7	48.8	41.3	0.066	0.002	-0.190	-0.026	0.484	0.444	0.567	0.567	0.127	0.012	0.033	0.034
Llama-3.2-1B	35.1	36.1	51.5	39.0	0.078	-0.025	-0.219	-0.014	0.400	0.370	0.550	0.555	0.143	0.016	0.061	0.078
Llama-2-13b-hf	41.2	43.1	48.0	38.7	0.059	-0.028	-0.145	0.029	0.436	0.404	0.550	0.546	0.080	-0.038	0.040	0.032
gemma-3-4b-it	28.6	29.3	48.6	46.4	0.063	0.045	-0.155	-0.122	0.792	0.703	0.385	0.431	0.036	-0.381	-0.099	0.092
phi-1_5	30.6	33.4	49.0	34.1	0.081	-0.055	-0.171	0.050	0.400	0.367	0.555	0.558	0.114	-0.062	0.029	0.082
Qwen2.5-0.5B	35.6	37.3	46.0	37.0	0.055	-0.022	-0.138	0.016	0.423	0.393	0.549	0.549	0.155	0.050	0.023	0.035
gemma-3-1b-it	26.2	27.0	48.5	43.8	0.063	0.018	-0.136	-0.076	0.643	0.567	0.433	0.476	0.124	-0.146	-0.139	0.025

Table 37: Comprehensive debiasing results for **MuSiQue: S2 Popular Trap**. An improvement in an 'After' cell is colored green; a regression is colored red.

Model	Accuracy (%)		HPSR (%)		PopGap		Corr		Confidence		Alignment		Spearman Conf Pop		Spearman Align Pop	
	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After
Llama-3.2-3B	43.1	57.6	41.0	22.7	0.207	0.050	-0.613	-0.473	0.459	0.436	0.549	0.529	0.021	-0.034	0.097	0.240
Mistral-7B-Instructruct-v0.3	48.8	55.0	35.1	28.0	0.172	0.105	-0.588	-0.535	0.712	0.639	0.561	0.593	-0.044	-0.131	-0.313	-0.138
gemma-3-12b-it	26.9	37.4	49.8	37.4	0.260	0.146	-0.512	-0.508	0.713	0.598	0.402	0.479	-0.038	-0.048	-0.242	-0.018
phi-2	46.0	57.2	37.3	24.2	0.165	0.050	-0.575	-0.465	0.524	0.491	0.556	0.550	0.002	-0.012	-0.039	0.122
Meta-Llama-3-8B-Instructruct	51.0	55.7	35.5	30.0	0.158	0.108	-0.625	-0.583	0.762	0.684	0.570	0.604	0.022	-0.116	-0.309	-0.172
Qwen2.5-7B	51.0	61.5	36.2	23.7	0.174	0.065	-0.639	-0.536	0.567	0.526	0.573	0.569	-0.009	-0.040	-0.057	0.110
Qwen2.5-1.5B	44.3	58.1	39.8	23.5	0.198	0.053	-0.611	-0.486	0.480	0.454	0.553	0.538	0.034	-0.035	0.103	0.231
Mistral-7B-Instructruct-v0.2	47.7	50.3	34.8	31.6	0.171	0.142	-0.569	-0.549	0.883	0.766	0.515	0.591	-0.053	-0.349	-0.382	-0.286
gemma-2-2b-it	33.0	45.0	45.0	31.4	0.233	0.106	-0.542	-0.499	0.566	0.502	0.491	0.513	-0.029	-0.032	-0.106	0.140
gemma-2-27b-it	41.1	47.0	38.5	31.6	0.193	0.129	-0.561	-0.534	0.775	0.671	0.476	0.538	0.079	-0.071	-0.389	-0.195
falcon-7b-instruct	26.1	45.5	47.4	24.2	0.253	0.054	-0.492	-0.418	0.395	0.385	0.551	0.527	0.006	-0.306	0.285	0.411
Llama-3.1-8B-Instructruct	50.6	59.4	35.5	24.9	0.162	0.070	-0.615	-0.527	0.601	0.559	0.583	0.585	-0.014	-0.014	-0.130	0.044
phi-4	55.8	60.6	32.8	27.3	0.139	0.090	-0.616	-0.565	0.708	0.653	0.601	0.617	-0.045	-0.121	-0.239	-0.110
gemma-2-9b-it	31.0	37.2	44.4	36.8	0.233	0.164	-0.502	-0.486	0.793	0.667	0.396	0.486	0.099	-0.087	-0.378	-0.142
zephyr-7b-beta	47.7	54.4	37.9	29.6	0.182	0.108	-0.615	-0.563	0.768	0.675	0.549	0.593	-0.082	-0.153	-0.379	-0.201
Llama-2-7b-hf	29.1	50.4	47.0	23.1	0.246	0.039	-0.516	-0.402	0.354	0.357	0.569	0.521	-0.051	-0.287	0.338	0.376
Qwen2.5-3B	47.4	61.1	38.6	22.3	0.189	0.049	-0.629	-0.499	0.489	0.463	0.564	0.544	0.020	-0.017	0.076	0.218
Llama-3.2-1B	35.9	54.0	44.5	22.9	0.234	0.048	-0.573	-0.440	0.396	0.382	0.549	0.518	0.102	-0.132	0.197	0.311
Llama-2-13b-hf	41.5	57.8	41.1	23.1	0.206	0.045	-0.597	-0.471	0.434	0.419	0.550	0.521	-0.036	-0.122	0.189	0.272
gemma-3-4b-it	26.0	34.6	46.1	36.5	0.251	0.160	-0.481	-0.486	0.789	0.649	0.363	0.464	0.063	-0.063	-0.329	-0.075
phi-1_5	32.4	50.1	46.0	24.6	0.245	0.055	-0.555	-0.464	0.406	0.389	0.548	0.524	0.099	-0.120	0.171	0.317
Qwen2.5-0.5B	36.0	53.2	44.6	23.9	0.226	0.047	-0.580	-0.460	0.430	0.410	0.548	0.528	0.121	-0.056	0.107	0.266
gemma-3-1b-it	25.5	38.0	48.7	33.6	0.257	0.121	-0.490	-0.483	0.637	0.539	0.427	0.489	0.120	0.066	-0.254	0.023

Table 38: Comprehensive debiasing results for **MuSiQue: S3 Popularity Gradient**. An improvement in an 'After' cell is colored green; a regression is colored red.

Model	Accuracy (%)		HPSR (%)		PopGap		Corr		Confidence		Alignment		Spearman Conf Pop		Spearman Align Pop	
	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After
Llama-3.2-3B	43.7	57.5	40.3	23.4	0.202	0.056	-0.608	-0.482	0.460	0.436	0.548	0.531	0.042	-0.039	0.121	0.263
Mistral-7B-Instructruct-v0.3	49.9	57.1	37.1	28.2	0.169	0.093	-0.612	-0.549	0.709	0.640	0.570	0.597	-0.080	-0.133	-0.299	-0.129
gemma-3-12b-it	25.6	35.2	49.1	37.7	0.249	0.145	-0.484	-0.480	0.719	0.610	0.391	0.468	-0.005	-0.066	-0.236	-0.010
phi-2	47.5	59.8	40.8	25.7	0.176	0.044	-0.612	-0.481	0.525	0.493	0.559	0.554	-0.028	-0.058	-0.062	0.100
Meta-Llama-3-8B-Instructruct	51.2	56.6	35.4	29.4	0.163	0.108	-0.637	-0.598	0.762	0.686	0.571	0.601	-0.003	-0.125	-0.303	-0.179
Qwen2.5-7B	52.3	63.1	36.5	24.1	0.175	0.067	-0.654	-0.555	0.568	0.526	0.572	0.568	-0.003	-0.039	-0.063	0.112
Qwen2.5-1.5B	43.0	58.4	40.5	23.2	0.209	0.055	-0.598	-0.473	0.482	0.455	0.555	0.537	0.053	-0.012	0.098	0.250
Mistral-7B-Instructruct-v0.2	45.7	50.2	38.9	33.7	0.177	0.129	-0.593	-0.571	0.882	0.765	0.507	0.578	-0.082	-0.296	-0.373	-0.249
gemma-2-2b-it	31.1	45.2	50.2	34.7	0.257	0.111	-0.562	-0.536	0.556	0.490	0.495	0.518	0.025	-0.017	-0.159	0.102
gemma-2-27b-it	42.1	48.5	48.0	40.5	0.231	0.161	-0.628	-0.606	0.779	0.653	0.482	0.559	0.089	-0.095	-0.444	-0.199
falcon-7b-instruct	25.1	51.4	64.3	35.8	0.312	0.066	-0.536	-0.479	0.395	0.382	0.554	0.517	0.030	-0.340	0.330	0.456
Llama-3.1-8B-Instructruct	51.0	61.5	39.8	27.7	0.173	0.072	-0.639	-0.541	0.597	0.552	0.585	0.586	-0.033	-0.055	-0.146	0.058
phi-4	55.8	61.5	35.5	28.7	0.145	0.083	-0.641	-0.577	0.717	0.658	0.606	0.621	-0.065	-0.106	-0.248	-0.109
gemma-2-9b-it	33.9	43.2	55.1	44.9	0.272	0.173	-0.596	-0.591	0.801	0.652	0.412	0.514	0.151	-0.017	-0.444	-0.160
zephyr-7b-beta	47.7	53.8	34.5	27.8	0.170	0.107	-0.602	-0.558	0.767	0.682	0.549	0.588	-0.061	-0.146	-0.341	-0.169
Llama-2-7b-hf	30.6	55.9	58.3	29.2	0.279	0.036	-0.573	-0.441	0.354	0.354	0.561	0.505	0.014	-0.304	0.337	0.361
Qwen2.5-3B	48.4	61.6	37.9	22.4	0.186	0.048	-0.637	-0.518	0.492	0.468	0.565	0.547	0.009	-0.047	0.069	0.194
Llama-3.2-1B	36.1	54.5	45.9	24.1	0.240	0.054	-0.593	-0.469	0.397	0.382	0.546	0.513	0.112	-0.110	0.220	0.328
Llama-2-13b-hf	40.6	57.3	46.2	27.1	0.219	0.055	-0.599	-0.478	0.432	0.412	0.549	0.521	-0.013	-0.133	0.193	0.297
gemma-3-4b-it	27.8	37.8	62.9	51.6	0.317	0.212	-0.575	-0.597	0.793	0.619	0.373	0.497	0.049	-0.057	-0.390	-0.083
phi-1_5	29.8	52.3	56.8	30.6	0.285	0.058	-0.566	-0.474	0.404	0.383	0.552	0.520	0.130	-0.137	0.173	0.367
Qwen2.5-0.5B	37.6	51.2	32.5	15.7	0.196	0.043	-0.551	-0.421	0.427	0.409	0.545	0.527	0.086	-0.022	0.159	0.279
gemma-3-1b-it	25.2	39.8	59.2	42.2	0.303	0.149	-0.536	-0.549	0.644	0.525	0.430	0.504	0.174	0.114	-0.355	0.024

The six MuSiQue tables mirror the structure used for TriviaQA/QASC/NQ. Each row shows a model's metric **Before** and **After** debiasing. Green indicates improvement, red indicates regression. Columns: *Accuracy*, *HPSR* (High–Popularity Success Rate; lower is better in trap/contest settings), *PopGap* (gap between popular vs. less popular choices; lower is better), *Corr* (correlation between popularity and choosing it; closer to 0 is better), *Confidence* and *Alignment* (desired: cali-

Table 39: Comprehensive debiasing results for **MuSiQue: S4 Direct Contest**. An improvement in an 'After' cell is colored green; a regression is colored red.

Model	Accuracy (%)		HPSR (%)		PopGap		Corr		Confidence		Alignment		Spearman Conf Pop		Spearman Align Pop	
	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After
Llama-3.2-3B	42.1	44.7	45.8	34.4	0.108	0.011	-0.364	-0.183	0.457	0.420	0.551	0.551	0.122	0.018	0.068	0.143
Mistral-7B-Instructruct-v0.3	47.0	48.4	43.9	39.5	0.086	0.050	-0.363	-0.289	0.708	0.645	0.553	0.562	0.002	-0.166	-0.121	-0.099
gemma-3-12b-it	26.4	28.1	48.1	41.2	0.131	0.069	-0.313	-0.251	0.715	0.620	0.395	0.453	-0.016	-0.238	-0.112	0.041
phi-2	43.0	44.1	39.0	32.8	0.061	0.008	-0.265	-0.168	0.507	0.477	0.547	0.549	0.065	-0.003	0.044	0.056
Meta-Llama-3-8B-Instructruct	48.5	49.7	42.8	40.5	0.080	0.061	-0.377	-0.340	0.747	0.684	0.566	0.571	0.088	-0.111	-0.122	-0.133
Qwen2.5-7B	49.9	51.4	45.8	40.0	0.100	0.053	-0.423	-0.323	0.556	0.513	0.572	0.571	0.104	-0.000	0.009	0.035
Qwen2.5-1.5B	42.1	45.0	45.2	36.0	0.098	0.022	-0.354	-0.221	0.472	0.437	0.553	0.550	0.138	0.036	0.084	0.129
Mistral-7B-Instructruct-v0.2	43.0	43.0	42.8	41.8	0.076	0.068	-0.322	-0.303	0.875	0.791	0.479	0.509	0.041	-0.422	-0.185	-0.137
gemma-2-2b-it	31.9	33.4	46.7	40.5	0.117	0.062	-0.350	-0.277	0.566	0.510	0.494	0.517	0.013	-0.155	-0.087	0.026
gemma-2-27b-it	39.5	40.1	43.4	41.6	0.092	0.076	-0.316	-0.297	0.777	0.701	0.468	0.501	0.102	-0.240	-0.198	-0.109
falcon-7b-instruct	25.6	31.8	47.1	30.2	0.133	-0.021	-0.282	-0.117	0.395	0.374	0.553	0.549	0.039	-0.199	0.149	0.258
Llama-3.1-8B-Instructruct	48.9	50.1	41.4	35.2	0.069	0.020	-0.332	-0.223	0.590	0.548	0.578	0.577	0.054	-0.026	-0.004	0.022
phi-4	52.8	54.4	41.0	37.7	0.072	0.042	-0.343	-0.282	0.706	0.652	0.597	0.594	0.025	-0.149	-0.078	-0.085
gemma-2-9b-it	30.9	31.4	43.3	40.4	0.103	0.077	-0.277	-0.244	0.796	0.714	0.388	0.438	0.078	-0.277	-0.184	-0.027
zephyr-7b-beta	45.4	46.0	44.1	41.6	0.087	0.066	-0.352	-0.312	0.766	0.700	0.525	0.545	0.012	-0.241	-0.172	-0.147
Llama-2-7b-hf	30.1	36.0	46.6	31.0	0.108	-0.020	-0.298	-0.126	0.355	0.342	0.565	0.550	-0.017	-0.097	0.172	0.215
Qwen2.5-3B	46.4	49.0	44.8	35.2	0.098	0.018	-0.374	-0.198	0.486	0.449	0.562	0.557	0.090	0.039	0.066	0.071
Llama-3.2-1B	35.9	37.6	48.2	35.4	0.124	0.013	-0.360	-0.185	0.397	0.368	0.549	0.552	0.126	0.006	0.115	0.167
Llama-2-13b-hf	41.2	45.0	47.2	34.8	0.104	-0.000	-0.392	-0.190	0.432	0.403	0.543	0.536	0.022	-0.052	0.074	0.106
gemma-3-4b-it	28.1	28.6	46.0	43.5	0.113	0.089	-0.298	-0.272	0.788	0.701	0.383	0.436	0.041	-0.315	-0.181	-0.010
phi-1_5	30.6	36.5	47.9	31.9	0.125	-0.013	-0.327	-0.147	0.403	0.372	0.549	0.547	0.147	-0.023	0.061	0.198
Qwen2.5-0.5B	34.9	39.5	46.0	32.6	0.120	0.004	-0.341	-0.171	0.424	0.390	0.550	0.544	0.170	0.091	0.056	0.112
gemma-3-1b-it	24.7	26.7	47.1	41.1	0.129	0.074	-0.292	-0.242	0.641	0.565	0.429	0.475	0.107	-0.092	-0.161	0.006

Table 40: Comprehensive debiasing results for **MuSiQue: S5 Reverse Control**. An improvement in an 'After' cell is colored green; a regression is colored red.

Model	Accuracy (%)		HPSR (%)		PopGap		Corr		Confidence		Alignment		Spearman Conf Pop		Spearman Align Pop	
	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After
Llama-3.2-3B	44.7	44.4	49.7	48.6	-0.045	-0.054	0.070	0.092	0.449	0.438	0.562	0.558	0.183	0.142	0.058	0.036
Mistral-7B-Instructruct-v0.3	46.4	46.4	49.0	48.6	-0.059	-0.062	0.124	0.130	0.691	0.673	0.569	0.561	0.133	0.074	0.157	0.127
gemma-3-12b-it	25.1	24.9	45.8	44.8	-0.075	-0.081	0.104	0.115	0.717	0.688	0.395	0.398	-0.016	-0.147	0.109	0.117
phi-2	39.7	39.8	43.5	43.3	-0.075	-0.077	0.188	0.193	0.488	0.483	0.548	0.546	0.126	0.109	0.036	0.023
Meta-Llama-3-8B-Instructruct	50.9	50.9	48.6	48.5	-0.043	-0.044	0.128	0.129	0.735	0.716	0.584	0.575	0.187	0.105	0.192	0.146
Qwen2.5-7B	52.4	52.0	49.6	49.0	-0.030	-0.037	0.091	0.107	0.543	0.528	0.591	0.584	0.188	0.151	0.159	0.128
Qwen2.5-1.5B	42.2	42.1	48.4	47.5	-0.048	-0.057	0.105	0.123	0.463	0.450	0.572	0.565	0.210	0.166	0.085	0.039
Mistral-7B-Instructruct-v0.2	43.9	43.8	47.5	47.3	-0.067	-0.069	0.135	0.138	0.868	0.847	0.489	0.484	0.130	-0.039	0.146	0.105
gemma-2-2b-it	31.5	30.7	48.0	46.6	-0.068	-0.080	0.109	0.126	0.560	0.538	0.500	0.501	0.077	-0.002	0.027	0.004
gemma-2-27b-it	37.6	37.2	46.7	46.2	-0.065	-0.071	0.138	0.145	0.761	0.731	0.450	0.447	0.135	0.008	0.074	0.062
falcon-7b-instruct	26.6	24.1	43.0	39.6	-0.081	-0.120	0.137	0.173	0.395	0.386	0.552	0.561	0.028	-0.127	-0.053	0.069
Llama-3.1-8B-Instructruct	48.8	48.5	46.4	46.2	-0.051	-0.054	0.151	0.156	0.579	0.568	0.587	0.581	0.162	0.141	0.117	0.101
phi-4	53.3	53.4	48.2	48.1	-0.043	-0.044	0.138	0.141	0.692	0.671	0.607	0.595	0.183	0.080	0.205	0.145
gemma-2-9b-it	30.1	29.4	45.0	44.4	-0.073	-0.079	0.145	0.150	0.795	0.762	0.393	0.395	0.063	-0.106	0.043	0.058
zephyr-7b-beta	45.1	44.5	48.0	47.5	-0.059	-0.065	0.139	0.147	0.749	0.726	0.547	0.539	0.120	0.045	0.160	0.129
Llama-2-7b-hf	29.8	27.9	45.6	40.8	-0.072	-0.119	0.127	0.188	0.353	0.345	0.565	0.572	0.003	-0.080	-0.062	0.026
Qwen2.5-3B	48.3	48.0	50.0	49.4	-0.043	-0.050	0.120	0.136	0.475	0.461	0.581	0.573	0.222	0.200	0.077	0.049
Llama-3.2-1B	37.3	35.1	51.2	48.1	-0.043	-0.072	0.073	0.120	0.394	0.381	0.555	0.559	0.145	0.063	-0.070	-0.045
Llama-2-13b-hf	41.3	40.5	48.0	46.3	-0.057	-0.074	0.116	0.145	0.426	0.413	0.552	0.549	0.092	0.028	-0.020	-0.048
gemma-3-4b-it	27.9	27.8	44.5	44.1	-0.077	-0.081	0.127	0.131	0.783	0.750	0.380	0.381	0.044	-0.147	0.058	0.084
phi-1_5	33.4	31.8	46.7	44.9	-0.064	-0.085	0.115	0.140	0.399	0.388	0.548	0.551	0.097	0.004	-0.099	-0.085
Qwen2.5-0.5B	35.0	34.0	46.5	44.5	-0.059	-0.080	0.125	0.157	0.419	0.404	0.554	0.553	0.210	0.133	-0.039	-0.052
gemma-3-1b-it	24.8	24.1	45.4	43.8	-0.073	-0.087	0.115	0.131	0.640	0.612	0.430	0.436	0.092	-0.021	-0.013	-0.017

Table 41: Comprehensive debiasing results for **MuSiQue: S6 None of the Above**. An improvement in an 'After' cell is colored green; a regression is colored red.

Model	Accuracy (%)		HPSR (%)		PopGap		Corr		Confidence		Alignment		Spearman Conf Pop		Spearman Align Pop	
	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After	Before	After
Llama-3.2-3B	0.2	5.8	63.5	39.8	0.350	0.050	-0.145	-0.504	0.437	0.400	0.562	0.586	0.106	-0.077	-0.099	0.210
Mistral-7B-Instructruct-v0.3	22.8	33.1	47.4	30.4	0.241	0.047	-0.833	-0.820	0.669	0.581	0.384	0.450	0.142	0.044	-0.372	-0.147
gemma-3-12b-it	29.5	39.6	43.8	28.0	0.236	0.052	-0.857	-0.850	0.722	0.638	0.426	0.490	-0.076	-0.169	-0.525	-0.310
phi-2	8.9	19.2	52.9	35.8	0.238	0.043	-0.693	-0.750	0.460	0.430	0.520	0.525	0.082	0.081	0.144	0.331
Meta-Llama-3-8B-Instructruct	39.6	51.5	36.9	21.8	0.199	0.035	-0.901	-0.888	0.670	0.597	0.453	0.501	0.097	-0.037	-0.545	-0.276
Qwen2.5-7B	8.2	19.1	58.4	36.6	0.325	0.049	-0.703	-0.750	0.501	0.442	0.492	0.524	0.144	-0.002	-0.051	0.324
Qwen2.5-1.5B	2.1	8.0	61.5	40.5	0.331	0.053	-0.424	-0.570	0.448	0.411	0.546	0.570	0.108	-0.046	-0.034	0.251
Mistral-7B-Instructruct-v0.2	22.3	30.1	46.8	33.1	0.228	0.060	-0.825	-0.817	0.869	0.739	0.293	0.398	0.081	-0.169	-0.507	-0.311
gemma-2-2b-it	60.3	69.1	25.0	14.7	0.133	0.023	-0.934	-0.914	0.570	0.543	0.512	0.525	-0.006	-0.097	-0.355	-0.150
gemma-2-27b-it	56.1	65.0	26.9	16.2	0.149	0.030	-0.916	-0.894	0.754	0.693	0.543	0.585	-0.007	-0.144	-0.732	-0.547
falcon-7b-instruct	0.1	4.2	62.2	40.1	0.330	0.049	-0.126	-0.456	0.389	0.387	0.610	0.605	-0.002	-0.199	0.010	0.266
Llama-3.1-8B-Instructruct	17.9	27.6	51.0	32.8	0.262	0.040	-0.805	-0.792	0.501	0.455	0.491	0.515	0.065	-0.027	0.050	0.291
phi-4	22.0	32.9	47.1	30.2	0.241	0.042	-0.813	-0.818	0.617	0.555	0.435	0.478	0.016	-0.042	-0.282	-0.063
gemma-2-9b-it	10.1	21.6	53.6	35.9	0.290	0.071	-0.710	-0.779	0.760	0.623	0.265	0.364	0.174	0.091	-0.269	-0.008
zephyr-7b-beta	13.8	25.7	54.5	34.9	0.287	0.059	-0.760	-0.782	0.721	0.606	0.318	0.406	0.118	0.081	-0.284	-0.100
Llama-2-7b-hf	10.3	21.9	57.2	34.6	0.283	0.034	-0.718	-0.758	0.326	0.331	0.634	0.601	0.166	-0.032	0.187	0.468
Qwen2.5-3B	11.6	24.9	56.0	33.1	0.305	0.041	-0.739	-0.779	0.415	0.383	0.555	0.556	0.195	-0.032	0.089	0.458
Llama-3.2-1B	0.0	3.8	65.2	40.9	0.382	0.059	0.000	-0.418	0.444	0.404	0.556	0.585	0.071	-0.149	-0.071	0.267
Llama-2-13b-hf	11.4	23.8	56.1	33.0	0.292	0.038	-0.752	-0.769	0.397	0.377	0.575	0.565	0.033	-0.116	0.237	0.507
gemma-3-4b-it	9.8	24.9	55.8	34.8	0.315	0.075	-0.703	-0.794	0.739	0.594	0.278	0.377	0.194	0.144	-0.265	0.028
phi-1_5	20.7	39.5	51.7	26.8	0.286	0.025	-0.807	-0.819	0.388	0.374	0.552	0.520	0.157	-0.031	0.390	0.651
Qwen2.5-0.5B	1.9	7.6	58.1	38.1	0.300	0.045	-0.391	-0.538	0.411	0.387	0.583	0.597	0.074	-0.087	-0.005	0.237
gemma-3-1b-it	19.6	33.2	50.7	31.4	0.278	0.053	-0.809	-0.822	0.593	0.508	0.401	0.452	0.242	0.136	-0.179	0.156

brated confidence and higher alignment), plus Spearman correlations of confidence/alignment with popularity (closer to 0 is better unless noted).

S1 Baseline Control (Table 36). Plain multiple–choice without adversarial framing. **Goal:** lift Accuracy while reducing popularity coupling (*PopGap* $\downarrow$, *Corr* $\rightarrow$ 0) without harming *Alignment*. **Read:** look for green in Accuracy alongside green (lower) PopGap/Corr.

S2 Popular Trap (Table 37). Choices include a "popular but wrong" distractor. **Goal:** reduce *HPSR* and *PopGap* substantially, weaken *Corr*, and ideally increase Accuracy. **Read:** green (lower) HPSR/PopGap/Corr with stable or higher Accuracy; Spearman (Conf/Align vs. Pop) should move toward 0.

S3 Popularity Gradient (Table 38). Popularity varies smoothly across options. **Goal:** retain/increase Accuracy while flattening response sensitivity to popularity (PopGap ↓, Corr ↓). **Read:** green Accuracy with green (lower) PopGap/Corr; Spearman coefficients closer to 0 indicate weaker popularity tracking.

S4 Direct Contest (Table 39). Correct answer competes head–to–head with a popular incorrect one. **Goal:** improve Accuracy and reduce *HPSR*, PopGap, and Corr; maintain or raise *Alignment*. **Read:** green in Accuracy/*HPSR* plus green (lower) PopGap/Corr; Alignment should not regress.

S5 Reverse Control (Table 40). A sanity–check where the "popularity prior" is flipped. **Goal:** avoid overfitting to the debias signal: Accuracy and Alignment should remain stable; correlations should not increase in magnitude. **Read:** minimal color (stability) is good; large red/green swings in Corr/PopGap suggest over–correction.

S6 None of the Above (Table 41). Includes an explicit "None" option that can attract popularity–driven errors. **Goal:** raise Accuracy and sharply lower *HPSR*/PopGap while keeping confidence calibrated (no overconfident wrong "None"). **Read:** green (higher) Accuracy, green (lower) HPSR/PopGap, Corr closer to 0, and Spearman coefficients moving toward 0; Alignment should increase or hold.

Comparison guidance across models. When scanning rows, prioritize: (i) **Accuracy** (↑) and **HPSR** (↓) within each scenario's intent; (ii) **PopGap/Corr** (↓→ 0) as direct bias indicators; (iii) **Spearman Conf/Align vs. Pop** (→ 0) to confirm reduced popularity–driven calibration; (iv) **Alignment** (↑ or stable) to ensure the method doesn't harm normative responses.

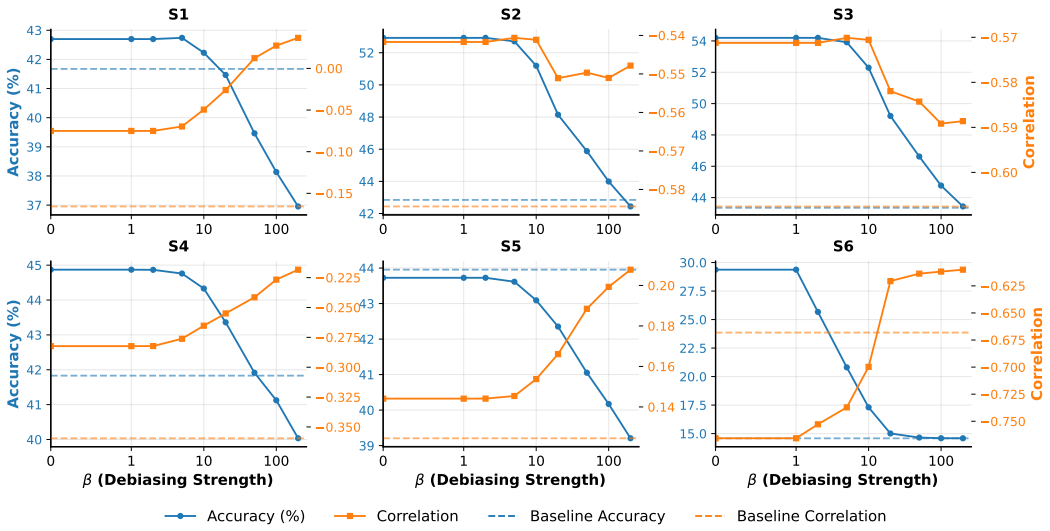

Figure 10: **NQ: β sweep (S1–S6).** $|\mathrm{Corr}|$ drops steadily as β grows, with clear gains by $\beta \in [10, 50]$ under popularity pressure (S2–S4, S6). Accuracy is stable for S1 and improves slightly on S2–S4 in this range. For S5, very large β begins to shave accuracy, reflecting mild over-tempering when popularity genuinely helps.

G TEMPERING COEFFICIENT β

This section interprets the full β sweeps for all datasets (NQ, TriviaQA, MuSiQue, QASC). Across settings S1–S6 we plot accuracy (higher is better) and $|\mathrm{Corr}|$ (lower is better) against β. The consistent pattern is: modest tempering ($\beta \approx 10$–20) yields most of the benefit under popularity pressure (S2–S4, S6), while extremely large β can slightly blunt performance in the reverse-control S5 where popularity is sometimes genuinely informative.

Natural Questions (Fig. 10). As β increases from 0 to $\sim$10–50, $|\mathrm{Corr}|$ steadily drops for S2–S4 and S6, indicating effective suppression of popularity chasing, while accuracy is stable for S1 and mildly improves for S2–S4. Very large values (> 50) begin to shave accuracy on S5, consistent with over-tempering when popularity aligns with truth. *Takeaway:* $\beta \in [10, 20]$ is a safe operating point that reduces $|\mathrm{Corr}|$ without hurting accuracy.

MuSiQue (Fig. 11). This dataset shows the *largest* benefit from tempering: increasing β strongly reduces $|\mathrm{Corr}|$ on S2–S4 and S6 and yields a tangible accuracy bump for S2–S4 around $\beta \in [10, 50]$. S1 stays stable; extremely large β (> 100) slightly dampens S5 accuracy, consistent with over-tempering. *Takeaway:* $\beta \approx 10$–20 offers an excellent trade-off and visibly boosts robustness under pressure.

QASC (Fig. 12). The pattern mirrors MuSiQue: sizable reductions in $|\mathrm{Corr}|$ on S2–S4/S6 alongside noticeable accuracy gains around $\beta \in [10, 50]$. S1 is flat; S5 shows the predictable minor cost as β grows, confirming the classic tempering trade-off. *Takeaway:* moderate tempering again hits the best balance—lower chasing with stable or improved accuracy.

Practical choice. Across all datasets, $\beta \approx 10$–20 consistently captures most of the accuracy gains (S2/S3/S6) and sharply lowers $|\mathrm{Corr}|$, with only mild S5 regressions at much larger values. We therefore set $\beta = 10$ by default and reserve larger values for stress tests requiring stronger suppression.

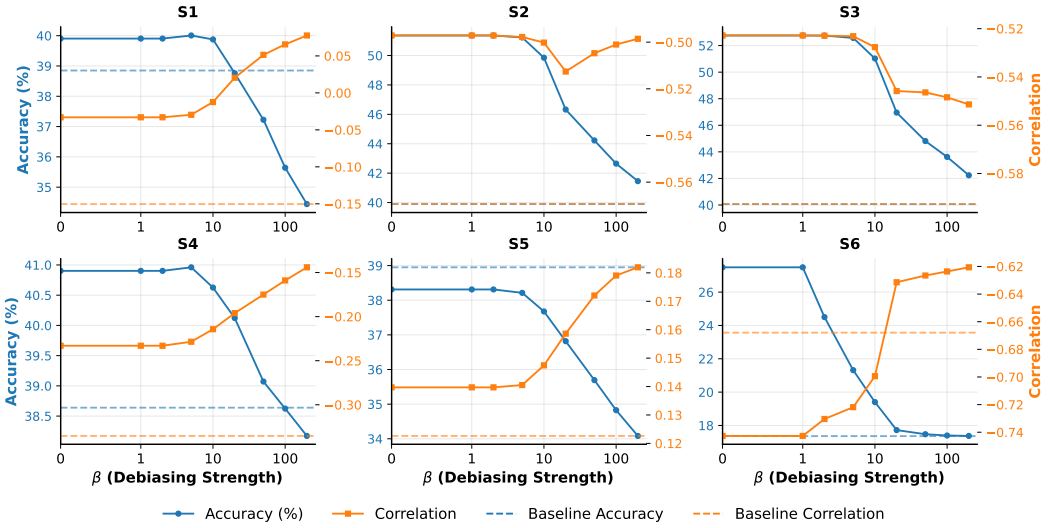

Figure 11: **MuSiQue: β sweep (S1–S6).** Largest benefit among the four: increasing β sharply reduces $|\mathrm{Corr}|$ on S2–S4 and improves S6, with a tangible accuracy bump for S2–S4 around $\beta \in [10, 50]$. S1 remains stable. Extremely large β (> 100) slightly dampens S5 accuracy due to over-tempering.

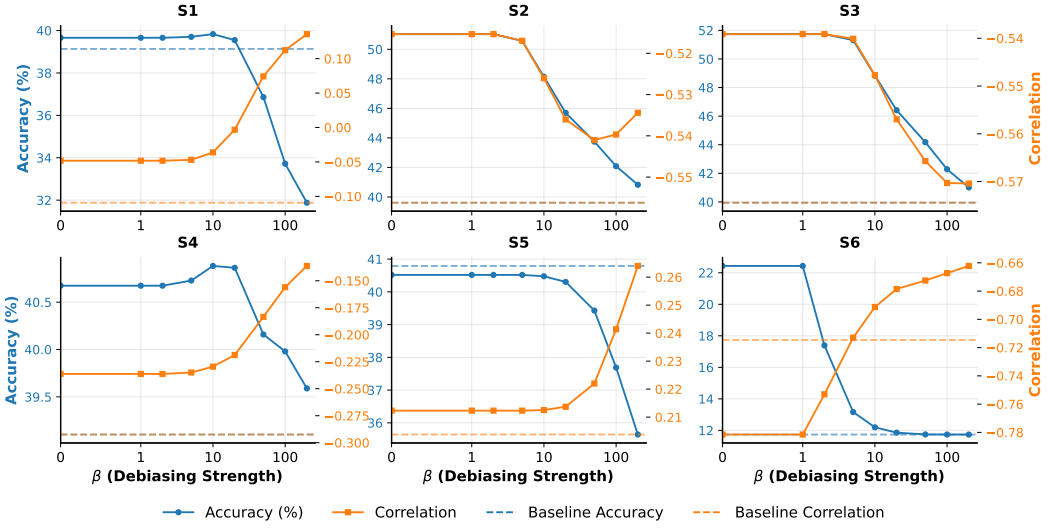

Figure 12: **QASC: β sweep (S1–S6).** Mirrors MuSiQue: sizable $|\mathrm{Corr}|$ reductions on S2–S4/S6 and noticeable accuracy gains at $\beta \in [10, 50]$. S1 is stable; S5 shows the predictable small cost as β grows, confirming that over-tempering can counteract alignment when popularity and correctness coincide.