Wisdom from Diversity: Bias Mitigation Through Hybrid Human-LLM Crowds

Axel Abels[1,2,3] , **Tom Lenaerts**[1,2,3,4]
[1]Machine Learning Group, Université Libre de Bruxelles
[2]AI Lab, Vrije Universiteit Brussel
[3]FARI, AI for the Common-Good Institute, ULB-VUB
[4]Center for Human-Compatible AI, UC Berkeley
{axel.abels, tom.lenaerts}@ulb.be

Abstract

Despite their performance, large language models (LLMs) can inadvertently perpetuate biases found in the data they are trained on. By analyzing LLM responses to bias-eliciting headlines, we find that these models often mirror human biases. To address this, we explore crowd-based strategies for mitigating bias through response aggregation. We first demonstrate that simply averaging responses from multiple LLMs, intended to leverage the "wisdom of the crowd", can exacerbate existing biases due to the limited diversity within LLM crowds. In contrast, we show that locally weighted aggregation methods more effectively leverage the wisdom of the LLM crowd, achieving both bias mitigation and improved accuracy. Finally, recognizing the complementary strengths of LLMs (accuracy) and humans (diversity), we demonstrate that hybrid crowds containing both significantly enhance performance and further reduce biases across ethnic and gender-related contexts.

1 Introduction

The increasing adoption of LLM assistants raises concerns about their potential to perpetuate or amplify societal biases [Bolukbasi *et al.*, 2016; Caliskan *et al.*, 2017; Zhao *et al.*, 2017; Zhang *et al.*, 2020]. These often subtle yet pervasive stereotypes pose a significant challenge to their responsible use. While prior work has identified biases stemming from training data and model architectures [Blodgett *et al.*, 2020], understanding how these biases manifest—especially in comparison to human biases—remains crucial. This understanding is key not only for evaluating LLM fairness and reliability but also for designing systems that complement human decision-making without perpetuating harmful biases.

In this work, we analyze LLM responses to a set of bias-eliciting headlines, comparing them to previously collected human responses [Abels *et al.*, 2024]. This allows us to assess how LLM biases align or differ from human biases, as well as to evaluate bias patterns across various LLMs.

Building on these insights, we explore strategies to mitigate biases in LLM outputs. Unlike previous works which tackled biases of individuals LLMs [Zhang *et al.*, 2020; Zhao *et al.*, 2018; Tamkin *et al.*, 2023], we draw on the principles of collective intelligence [Surowiecki, 2005]. Specifically, we investigate the effectiveness of aggregating responses from multiple LLMs—creating "LLM crowds". We hypothesize that, similar to human crowds [Abels *et al.*, 2024], this aggregation can effectively diminish bias by leveraging the diversity of responses. Using methods like simple averaging and locally weighted averages—where weights are tailored to the specializations of the crowd—we evaluate how effectively these approaches address biases while enhancing performance.

Finally, given the complementary strengths of humans (high diversity, lower individual accuracy) and LLMs (high individual accuracy, lower diversity), and recognizing that both are crucial for effective collective intelligence [Hong and Page, 2004; Surowiecki, 2005], we examine the potential of hybrid crowds. Our findings demonstrate that hybrid crowds significantly improve accuracy while reducing biases to negligible levels, outperforming both LLM-only and human-only groups. These results highlight the potential of hybrid approaches to enhance fairness and accuracy in applications such as content moderation and hiring systems.

To summarize, our contributions include: (1) a comparative analysis of LLM and human biases on bias-eliciting headlines, (2) an evaluation of the potential and limitations of LLM crowds, and (3) a demonstration of the superior performance and fairness of hybrid crowds.

2 Background

Social biases are deeply ingrained cognitive patterns that influence human perception, judgment, and behavior. These biases, often unconscious, arise from stereotypes, cultural norms, and societal structures. While they are often the product of heuristics—mental shortcuts which enable quick decision-making—they frequently result in discrimination, inequality, and harm to marginalized groups [Greenwald *et al.*, 1998; Devine, 1989]. For instance, racial biases perpetuate systemic inequalities in education, employment, and healthcare [Williams, 1996; Reskin, 2000], while gender biases limit opportunities and reinforce harmful stereotypes about abilities and roles [Eagly and Johnson, 1990].

Despite growing awareness and efforts to combat them, social biases persist, influencing hiring practices, educational access, healthcare delivery, and legal outcomes [Bertrand and

Mullainathan, 2004; Pager, 2008]. These biases are often perpetuated through societal institutions and media representations [Dovidio *et al.*, 2017]. The growing integration of artificial intelligence into daily life adds a new dimension to this challenge. Widely adopted LLM assistants could expose users to these biases on a broader scale, reflecting and potentially reinforcing prejudices embedded in their training data.

2.1 Social Biases in LLMs

LLMs are a class of neural networks trained on extensive text corpora to perform a wide range of natural language processing (NLP) tasks, including text generation, summarization, translation, and conversational interactions [Vaswani, 2017; Radford *et al.*, 2019; Brown *et al.*, 2020]. Building on earlier NLP advancements, such as word embeddings [Mikolov, 2013], LLMs use transformer-based architectures to create dynamic, context-aware representations [Vaswani, 2017], enabling them to achieve high accuracy and fluency. However, their reliance on vast amounts of real-world data also makes them susceptible to inheriting and amplifying societal biases present in their training data [Bolukbasi *et al.*, 2016].

These biases manifest in various forms, including biased word associations [Caliskan *et al.*, 2017], stereotypical sentence generation [Zhao *et al.*, 2017], and unequal performance across demographic groups [Zhang *et al.*, 2020]. This phenomenon has been extensively documented across various social dimensions, including gender, race, and socioeconomic status [Bolukbasi *et al.*, 2016; Davidson *et al.*, 2019; Caliskan *et al.*, 2017]. For instance, word embedding association tests reveal that LLMs often replicate harmful stereotypes, such as associating certain professions predominantly with one gender [Zhao *et al.*, 2018]. Such biases extend to conversational systems, which risk exposing users to outputs that reinforce stereotypes and perpetuate societal prejudices.

Understanding how LLMs manifest social biases has therefore become a crucial component in the development of socially responsible AI systems [Gemini *et al.*, 2023; Anthropic, 2024; Achiam *et al.*, 2023]. However, mitigating these biases is a complex challenge. Early efforts focused on post-processing techniques, such as neutralizing gender-specific dimensions in word embeddings [Bolukbasi *et al.*, 2016]. Recent research includes interventions during the generation phase, such as reweighting or filtering outputs to suppress stereotypes [Sheng *et al.*, 2020; Liang *et al.*, 2021].

Beyond algorithmic solutions, socio-technical approaches have gained attention as complementary strategies. These include curating representative datasets, incorporating diverse evaluation metrics, and prioritizing interdisciplinary collaboration to address ethical concerns [Birhane *et al.*, 2021]. Such approaches aim to tackle biases at their source, emphasizing the importance of combining technical innovations with ethical and societal considerations.

Despite these efforts, several challenges persist. Bias mitigation techniques can struggle to generalize across different bias types, and often degrade overall model performance. Additionally, the high cost of training state-of-the-art models has incentivized the development of closed-source LLMs, complicating efforts to ensure transparency and accountability.

2.2 Collective Intelligence

Beyond model-centric solutions, we argue in this work that leveraging collective intelligence offers a promising complementary approach to bias mitigation. The principle that groups can achieve more accurate and reliable outcomes than individuals is well-established [Condorcet, 1785; Hong and Page, 2004; Surowiecki, 2005]. This “wisdom of the crowd” effect leverages diverse perspectives and independent errors to improve decision-making.

Wisdom of LLM Crowds. The concept of crowd wisdom, widely studied in human groups and exploited by traditional machine learning ensembles [Surowiecki, 2005; Wolpert, 1992; Breiman, 1996], is now being explored in the context of LLMs. Recent research demonstrates the existence of a “wisdom of LLM crowds”. For instance, Schoenegger *et al.* [2024] show that LLM ensembles achieve forecasting accuracy comparable to human crowds. Additionally, they show that exposing LLMs to the median prediction of human crowds enhances their accuracy, consistent with findings from human studies [Becker *et al.*, 2017].

Exploring bias dynamics, Chuang *et al.* [2024] show that LLM crowds, when simulating partisan personas, can mimic human-like partisan biases. However, deliberation among these personas leads to more accurate beliefs, mirroring the benefits of discussion and information sharing which are sometimes observed in human groups [Becker *et al.*, 2019].

While these studies reproduce key results in LLMs, they leave open critical questions. Specifically, how should LLMs be aggregated to systematically mitigate biases while enhancing performance? Furthermore, the potential synergy between LLMs and humans remains largely unexplored. To enable us to address these gaps, we now review prior methods to mitigate biases and restore crowd wisdom in human groups.

2.3 Bias Mitigation in Human Crowds

Understanding how biases influence crowd wisdom in human groups provides valuable insights for developing aggregation strategies in LLM and hybrid crowds. The following experiment illustrates the prevalence of social biases in individuals and their impact on collective decision-making.

The Headline Experiment

In [Abels *et al.*, 2024], participants evaluated a balanced set of genuine and altered news headlines, where demographic groups were swapped to create counterfactual pairs. For example, the headline *“Men more likely than women to say they are financially better off since last year”* was altered to *“Women more likely than men to say they are financially better off since last year”*. Headlines described positive or negative outcomes for various demographic groups (gender, ethnicity, age), and participants rated their authenticity on a scale from “very unlikely” to “very likely.”

This design allowed the authors to measure bias by comparing error rates across demographic groups and outcomes. For example, discrepancies in error rates for positive vs. negative outcomes for white individuals revealed underlying biases. Further analyses of the responses revealed that factors like responders’ demographics, headline categories, and question framing significantly influenced individual judgments, often leading to systematic errors.

Restoring Crowd Wisdom in Human Groups
While the "wisdom of the crowd" suggests that diverse perspectives could mitigate individual biases, Abels *et al.* [2024] demonstrated that dominant societal beliefs can still lead to collective errors. To address this, they proposed using locally-weighted aggregation to counter individual biases and variations in expertise. Specifically, they used ExpertiseTrees [Abels *et al.*, 2023], which leverage the diversity within a group by partitioning the context space (e.g., headline categories) and fitting specialized models to each region. Similar to decision trees, ExpertiseTrees dynamically introduce splits only when they improve group aggregation.

This hierarchical approach offers several advantages compared to other aggregation methods. While techniques like stacking [Breiman, 1996] or simple averaging provide static aggregation, ExpertiseTrees allow for more nuanced aggregation by adjusting individuals' weights to the problem (here, the headline) at hand. This can be particularly beneficial for bias mitigation, as different individuals may exhibit different biases in different contexts. In particular, Abels *et al.* [2024] show that when identifying fake news, human performance varies significantly with the demographic group contained in the headline. ExpertiseTrees partition the context space to route inputs to individuals most likely to provide unbiased responses, improving accuracy and fairness.

While this research highlights the promise of advanced aggregation methods in human crowds, their applicability to LLM crowds remains uncertain. LLMs may lack the diversity found in humans and exhibit distinct biases, potentially limiting the effectiveness of aggregation strategies. Additionally, we hypothesize that the complementary strengths of humans and LLMs—humans contributing greater diversity and LLMs offering higher individual accuracy—present a unique opportunity for hybrid approaches. With these considerations in mind, we now present how we addressed these challenges.

3 Methods

Our aim in this work is two-fold. First, we aim to compare LLM biases with those of humans. Second, we aim to study the benefits offered by crowds, first crowds of LLMs, and then hybrid crowds, containing both LLMs and humans. As a first step, we replicate the headline experiment conducted by Abels *et al.* [2024] on LLMs, enabling a direct comparison of LLM biases and performance with those of humans.

To ensure broad coverage, we include both closed-source and open-source LLMs from OpenAI [2023], Anthropic [2024], Google [2023], Meta [2024], Mistral [2023], Alibaba [2023], and DeepSeek AI [2024]. Supplementary Table S.2 provides a detailed comparison of the selected models. These LLMs were selected to capture a diverse range of architectures and training paradigms, ensuring a comprehensive analysis of LLM performance and biases.

3.1 Experimental Procedure

We closely replicate the methodology of [Abels *et al.*, 2024] to enable a direct comparison between humans and LLMs.

We therefore prompted LLMs to estimate the likelihood that the given headlines were real. Following the human study, LLMs were instructed to respond using a 5-point Likert scale ranging from "very unlikely" to "very likely." The response for every headline h was then mapped onto a numerical likelihood: $p_h \in \{0, 0.25, 0.5, 0.75, 1\}$. All LLMs were prompted in a 4-shot setting with instructions designed to replicate the guidance provided to human participants. Prompts included a brief explanation of the task and example responses. Details on the prompting procedure and parameters are given in the supplementary materials.

3.2 Metrics

We evaluate LLMs on their accuracy, diversity, and susceptibility to biases, comparing them with previously collected human data.

Accuracy. Accuracy is calculated as the proportion of correctly identified headlines, whether genuine or altered. Since the dataset is balanced across class (genuine vs. altered), sentiment (positive vs. negative), and demographic groups (man $\leftrightarrow$ woman, young $\leftrightarrow$ old, White $\leftrightarrow$ African American), this metric provides an overall indication of performance. To identify potential specializations, accuracy is also reported for each combination of class and demographic group.

To further identify whether there are systematic differences in accuracy which can be attributed to prejudices against certain groups, we explore the following types of bias.

Counterfactual Bias. In line with conditional statistical parity [Corbett-Davies *et al.*, 2017], we define counterfactual bias as a systematic tendency to favor certain outcomes for specific demographic groups. For a demographic group g and its counterpart g', this bias is quantified as:

$$\Delta_{g,g'}(s,\sigma) = \mathbb{E}_{h\sim H_{s,\sigma,g}}[p_h] - \mathbb{E}_{h'\sim H_{s,\sigma,g'}}[p_{h'}], \quad (1)$$

where p_h is the likelihood assigned to headline h, and $H_{s,\sigma,g}$ is the subset of headlines of status s (Genuine or Altered), sentiment σ (positive or negative), and demographic group g (e.g., man, woman, young , old, White, African American).

Positive values indicate a bias towards group g, whereas negative values indicate a bias towards g'. For example, if a model is more likely to believe headlines reporting positive outcomes for men over women, $\Delta_{man,woman}(genuine,+)$ will be positive, indicating a bias in favor of men. In practice, $\mathbb{E}_{h\sim H_{s,\sigma,g}}[p_h]$ is estimated as the mean likelihood assigned to all headlines in $H_{s,\sigma,g}$. To determine the significance of counterfactual biases, we use Mann-Whitney U tests [Mann and Whitney, 1947] to compare the distributions of likelihoods assigned to headlines from different demographic groups.

Framing Effects. Framing effects occur when responses differ based on how identical information is presented. In the headline dataset, a responder's likelihood for a headline h may differ from $1-p_{h'}$, where h' is the counterfactual variant of h obtained by swapping demographic groups. Systematic differences indicate susceptibility to framing effects.

For a group g and sentiment σ, average framing effects are:

$$\Delta_F(\sigma, g) = \frac{1}{|H_{\sigma,g}|} \sum_{h\in H_{\sigma,g}} p_h - (1 - p_{h'})$$

where $H_{\sigma,g}$ is the set of headlines with that sentiment and group. For example, for humans $\Delta_F(positive, young)$ and $\Delta_F(negative, young)$ are positive, indicating they tend to assign the same belief to a headline reporting the opposite outcome for age groups. We use Wilcoxon signed-rank tests [Wilcoxon, 1992] to assess significance.

Diversity. Diversity is crucial for crowd wisdom, as it leverages individuals' independent errors [Wood *et al.*, 2023]. When group members consistently make the same mistakes, their combination offers little benefit, as they merely amplify their shared errors and biases. Conversely, when individuals exhibit diverse behavior—where one member's errors are offset by other members' correct predictions—the group can achieve higher performance. This is particularly relevant for bias mitigation, as one group member might exhibit gender bias in a specific context, while another model, trained on a different dataset, or using a different architecture, might not.

To quantify diversity, we use the Q-statistic [Yule, 1900], which measures the extent to which two classifiers make the same predictions (correct or incorrect). It is defined as

$$Q = \frac{N_{11}\,N_{00} - N_{10}\,N_{01}}{N_{11}\,N_{00} + N_{10}\,N_{01}}, \tag{2}$$

where N_{11} is the number of instances both classifiers predict correctly, N_{00} is the number of instances both classifiers predict incorrectly, N_{10} is the number of instances in which C_1 is correct while C_2 is incorrect, and N_{01} is the number of instances in which C_1 is incorrect while C_2 is correct.

Combined with high individual accuracies, a lower Q-statistic often signals useful diversity, yielding better ensembles as the collective can compensate for individual mistakes.

3.3 Ensembling Strategies

Our second major contribution is the exploration of ensemble-based strategies to mitigate LLM biases. We compare LLM crowds against individual models to highlight the benefits offered by different aggregation strategies, including potential trade-offs between accuracy and bias mitigation.

In addition to using simple averages, we investigate two types of weighted averages. First, we explore traditional stacking [Breiman, 1996], wherein group members are assigned a constant weight. While such weights effectively prioritize consistently well-performing members, they cannot adapt to variations in context, such as biases or specializations specific to certain headline categories. To address this limitation, we also study weights tailored to headline categories.

Specifically, let p_h^m be the likelihood group member m assigns to headline h. The aggregated likelihood is $\bar{p}_h = \sum_m w_{\phi(h)}^m\, p_h^m$, where $\phi(h)$ is h's category (age, gender, ethnicity), and $w_{\phi(h)}^m$ is individual m's weight for that category.

We use ExpertiseTrees (see [Abels *et al.*, 2023], as well as our detailed description in the Supplementary Information) to learn these localized weights[1]. Unlike traditional stacking, which learns static weights for combining model predictions, ExpertiseTrees use a tree structure to assign weights based on input-specific characteristics, such as headline categories. This allows ExpertiseTrees to amplify the contributions of less biased members while down-weighting biased predictions, thereby improving fairness and accuracy. In the headline setting, localized weights allow us to adapt the aggregation to headline categories. For instance, Abels *et al.* [2024] demonstrate that human performance varies significantly across headline categories. ExpertiseTrees exploit this specialization by learning a distinct set of weights for each category, allowing for context-sensitive aggregation.

Benchmark-based Sampling. When forming LLM groups, we employ two sampling approaches. First, we use random sampling, where LLMs are selected uniformly at random from the pool of available models. This serves as a baseline for evaluating the benefits of ensemble methods. Second, we select LLMs based on their scores on the widely used MMLU benchmark [Hendrycks *et al.*, 2020]. This method prioritizes high-performing models, which we hypothesize will better correlate with improved performance on the headline task. However, selecting high-performing models may reduce diversity due to increased similarity among the selected models, presenting a trade-off between individual strength and collective diversity. Groups formed using this performance-based approach are denoted with a "+" sign, e.g., LLM+.

To evaluate the potential of hybrid ensembles, we combined human responders from the headline experiment dataset with LLMs. We hypothesize that hybrid groups leverage the complementary strengths of humans and LLMs: humans provide greater diversity, while LLMs offer higher individual accuracy. The LLMs in hybrid groups were selected using the same two sampling approaches, resulting in two types of hybrid groups: hybrid, with randomly sampled LLMs, and hybrid+, with performance-selected LLMs.

To assess the impact of group size on collective performance, we evaluated groups ranging in size from 2 to 16 responders. This range reflects practical limits, as the available pool of LLMs consisted of 18 models. Although larger groups were not tested, observed trends provide a basis for inferring potential outcomes for larger groups.

4 Results

We now present the results of our evaluation of the headline experiment on LLMs[2]. Specifically, we first compare the performance and bias of individual LLMs to human participants. Next, we demonstrate the limitations of LLM crowds and the advantages of hybrid crowds, emphasizing the potential of locally weighted aggregations in achieving improved outcomes.

4.1 LLM Performance

To evaluate the potential for collective intelligence, we begin by analyzing individual LLM performances. Table 1 reports the accuracy of various LLMs across headline categories, along with their counterfactual biases. For comparison, the table also includes the average human participant.

[1] For both weighted average methods, we use cross-validation to ensure weights were not trained on headlines they are evaluated on.

[2] Code available at [Abels and Lenaerts, 2025].

	AGE		ETHNICITY		GENDER		AVERAGE
MODEL	altered	genuine	altered	genuine	altered	genuine	
human	0.44 (Δ=+0.01)	0.65 (Δ=-0.01)	0.65 (Δ=+0.09)	0.48 (Δ=+0.17)	0.57 (Δ=-0.07)	0.53 (Δ=-0.04)	0.550
Qwen2.5-72B-Instruct	0.57 (Δ=+0.04)	0.82 (Δ=-0.05)	0.65 (Δ=+0.12)	0.62 (Δ=+0.14)	0.65 (Δ=+0.04)	0.50 (Δ=+0.04)	0.637
claude-3-5-haiku	0.78 (Δ=+0.03)	0.35 (Δ=+0.09)	0.52 (Δ=+0.16)	0.47 (Δ=+0.33)	0.80 (Δ=+0.00)	0.52 (Δ=-0.12)	0.575
claude-3-5-sonnet	0.55 (Δ=+0.06)	0.88 (Δ=+0.11)	0.55 (Δ=+0.35)	0.75 (Δ=+0.32)	0.68 (Δ=-0.11)	0.82 (Δ=-0.03)	0.704
claude-3-opus	0.50 (Δ=+0.10)	0.93 (Δ=+0.08)	0.48 (Δ=+0.18)	0.73 (Δ=+0.06)	0.65 (Δ=-0.15)	0.62 (Δ=+0.01)	0.650
deepseek	0.40 (Δ=-0.03)	0.85 (Δ=+0.08)	0.42 (Δ=+0.23)	0.65 (Δ=+0.27)	0.45 (Δ=-0.17)	0.75 (Δ=-0.01)	0.588
gemini-1.5-flash	0.48 (Δ=+0.03)	0.60 (Δ=+0.00)	0.68 (Δ=+0.22)	0.50 (Δ=+0.28)	0.78 (Δ=-0.02)	0.42 (Δ=-0.02)	0.575
gemini-1.5-pro	0.62 (Δ=+0.06)	0.53 (Δ=-0.01)	0.62 (Δ=+0.24)	0.68 (Δ=+0.24)	0.75 (Δ=-0.14)	0.45 (Δ=-0.05)	0.608
gemini-2.0-flash	0.45 (Δ=+0.04)	0.77 (Δ=+0.08)	0.45 (Δ=+0.19)	0.70 (Δ=+0.32)	0.60 (Δ=-0.16)	0.70 (Δ=-0.15)	0.612
gemma2-9b	0.33 (Δ=+0.01)	0.78 (Δ=-0.01)	0.47 (Δ=+0.16)	0.62 (Δ=+0.27)	0.57 (Δ=-0.10)	0.65 (Δ=+0.06)	0.571
gemma-2-27b	0.53 (Δ=+0.02)	0.62 (Δ=-0.01)	0.80 (Δ=+0.01)	0.38 (Δ=+0.27)	0.82 (Δ=-0.14)	0.40 (Δ=-0.11)	0.592
gpt-4	0.65 (Δ=-0.01)	0.57 (Δ=-0.05)	0.70 (Δ=+0.06)	0.45 (Δ=+0.31)	0.70 (Δ=-0.14)	0.50 (Δ=-0.02)	0.596
gpt-4-turbo	0.55 (Δ=-0.01)	0.68 (Δ=-0.01)	0.65 (Δ=+0.03)	0.55 (Δ=+0.22)	0.75 (Δ=-0.10)	0.60 (Δ=+0.04)	0.629
gpt-4o	0.57 (Δ=-0.08)	0.80 (Δ=-0.04)	0.68 (Δ=+0.18)	0.75 (Δ=+0.24)	0.78 (Δ=-0.08)	0.75 (Δ=-0.05)	**0.721**
gpt-4o-mini	0.28 (Δ=-0.06)	0.82 (Δ=+0.01)	0.62 (Δ=+0.08)	0.50 (Δ=+0.29)	0.42 (Δ=-0.16)	0.72 (Δ=-0.11)	0.562
Llama-3.3-70B	0.43 (Δ=+0.13)	0.80 (Δ=+0.10)	0.33 (Δ=+0.20)	0.80 (Δ=+0.17)	0.53 (Δ=-0.13)	0.65 (Δ=-0.10)	0.588
mistral-large	0.50 (Δ=+0.04)	0.80 (Δ=+0.05)	0.62 (Δ=+0.20)	0.57 (Δ=+0.33)	0.78 (Δ=-0.10)	0.68 (Δ=-0.07)	0.658
mixtral-8x7b	0.28 (Δ=+0.06)	0.90 (Δ=+0.03)	0.28 (Δ=+0.04)	0.80 (Δ=+0.19)	0.35 (Δ=-0.15)	0.88 (Δ=-0.01)	0.579
open-mistral	0.73 (Δ=-0.07)	0.38 (Δ=+0.05)	0.93 (Δ=-0.06)	0.25 (Δ=+0.15)	0.80 (Δ=-0.14)	0.25 (Δ=+0.02)	0.554
average(LLM)	0.53 (Δ=+0.02)	0.83 (Δ=+0.03)	0.57 (Δ=+0.14)	0.62 (Δ=+0.24)	0.72 (Δ=-0.11)	0.65 (Δ=-0.04)	0.652
average(human)	0.43 (Δ=+0.01)	0.75 (Δ=-0.02)	0.79 (Δ=+0.11)	0.48 (Δ=+0.17)	0.67 (Δ=-0.09)	0.55 (Δ=-0.04)	0.611
average(hybrid)	0.50 (Δ=+0.01)	0.80 (Δ=+0.00)	0.69 (Δ=+0.13)	0.56 (Δ=+0.21)	0.69 (Δ=-0.10)	0.60 (Δ=-0.04)	0.641
average(LLM+)	0.60 (Δ=+0.01)	0.83 (Δ=+0.03)	0.52 (Δ=+0.17)	0.62 (Δ=+0.24)	0.70 (Δ=-0.11)	0.65 (Δ=-0.02)	0.654
average(hybrid+)	0.55 (Δ=+0.01)	0.78 (Δ=+0.02)	0.73 (Δ=+0.18)	0.60 (Δ=+0.23)	0.77 (Δ=-0.11)	0.63 (Δ=-0.05)	**0.678**
WeightedAverage(LLM)	0.67 (Δ=-0.01)	0.68 (Δ=-0.02)	0.67 (Δ=+0.11)	0.70 (Δ=+0.14)	0.74 (Δ=-0.04)	0.75 (Δ=+0.01)	0.703
WeightedAverage(human)	0.62 (Δ=-0.02)	0.66 (Δ=-0.04)	0.69 (Δ=+0.08)	0.70 (Δ=+0.12)	0.70 (Δ=-0.09)	0.65 (Δ=-0.04)	0.671
WeightedAverage(hybrid)	0.66 (Δ=-0.01)	0.68 (Δ=-0.02)	0.69 (Δ=+0.07)	0.71 (Δ=+0.11)	0.75 (Δ=-0.07)	0.74 (Δ=-0.03)	0.704
WeightedAverage(LLM+)	0.76 (Δ=-0.02)	0.75 (Δ=-0.03)	0.69 (Δ=+0.12)	0.72 (Δ=+0.15)	0.72 (Δ=-0.02)	0.79 (Δ=+0.01)	0.738
WeightedAverage(hybrid+)	0.71 (Δ=-0.02)	0.74 (Δ=+0.00)	0.75 (Δ=+0.11)	0.77 (Δ=+0.13)	0.83 (Δ=-0.06)	0.83 (Δ=-0.03)	**0.772**
ExpertiseTree(LLM)	0.70 (Δ=-0.02)	0.70 (Δ=-0.03)	0.64 (Δ=+0.07)	0.65 (Δ=+0.14)	0.76 (Δ=-0.03)	0.75 (Δ=-0.04)	0.701
ExpertiseTree(human)	0.72 (Δ=-0.02)	0.76 (Δ=-0.02)	0.75 (Δ=+0.05)	0.77 (Δ=+0.08)	0.75 (Δ=-0.04)	0.75 (Δ=-0.02)	0.749
ExpertiseTree(hybrid)	0.73 (Δ=+0.01)	0.80 (Δ=-0.01)	0.73 (Δ=+0.05)	0.76 (Δ=+0.05)	0.82 (Δ=-0.04)	0.83 (Δ=-0.03)	0.777
ExpertiseTree(LLM+)	0.74 (Δ=-0.01)	0.75 (Δ=-0.03)	0.68 (Δ=+0.09)	0.71 (Δ=+0.11)	0.82 (Δ=-0.01)	0.80 (Δ=-0.01)	0.752
ExpertiseTree(hybrid+)	0.81 (Δ=-0.02)	0.82 (Δ=-0.03)	0.75 (Δ=+0.04)	0.75 (Δ=+0.04)	0.89 (Δ=-0.00)	0.86 (Δ=-0.03)	**0.813**

Table 1: Accuracy and counterfactual bias (Δ, see Equation 1) across headline categories. High counterfactual biases indicate a higher belief in positive headlines for historically privileged groups (older, white, male). Cell shading represents statistical significance of the counterfactual bias: darkest red for $p < 0.01$, medium red for $p < 0.05$, and light red for $p < 0.1$. Rows labeled *average(·)*,*WeightedAverage(·)*, and *ExpertiseTree(·)* give the performance of groups of 8 aggregated through respectively simple averages, weighted averages, or locally weighted averages. These groups can consist of humans, LLMs, or a mix of both (hybrid). Instead of randomly sampling available LLMs, LLM+ and Hybrid+ select the models with the highest scores on the MMLU benchmark.

LLMs show above-human-level accuracy. Our findings indicate that GPT-4o and Claude-3.5-Sonnet achieve the highest average accuracy, consistent with prior benchmarks like MMLU [Hendrycks *et al.*, 2020]. While smaller models, such as mistral-8x7b, generally perform worse, all tested LLMs outperform the average human on this task.

LLMs mirror human counterfactual biases. Similar to human responders, all tested LLMs show some degree of counterfactual bias, especially for headlines involving ethnic groups. LLMs tend to assign higher likelihoods to headlines reporting positive outcomes for White individuals than for African-American ones, especially when only considering genuine headlines (Table 1, ETHNICITY-Genuine column). All but one LLM show a significant effect for this bias. For age headlines, counterfactual bias is less pronounced in both LLMs and humans. Gender headlines elicit moderate bias, with fewer than half of the tested LLMs showing some form of counterfactual bias.

These findings suggest that while LLMs outperform humans in accuracy, they are likely to reinforce existing biases when assisting in decision-making.

LLMs are less susceptible to framing effects. One key result from [Abels *et al.*, 2024] was that humans showed significant framing effects. In particular, humans often gave similar likelihood ratings to headlines reporting opposite outcomes for age and ethnicity categories. This was attributed to varying levels of skepticism: humans showed low skepticism for age-related headlines, assigning high likelihoods regardless of content, but were more skeptical of ethnicity-related headlines, often assigning lower likelihoods. Humans were more discerning for gender-related headlines but had lower skepticism for headlines reporting negative outcomes for men.

In contrast, we found that LLM groups tend to be less susceptible to framing effects than human groups. Supplementary Figure S4 shows LLMs are not biased in their responses to headlines reporting outcomes for ethnic groups. Similarly, while humans show significant framing effects for age headlines, we found that LLMs display very little framing effects, suggesting they have stronger opinions on these headlines than human responders. In terms of framing effects for gender headlines, LLMs aligned closely with humans.

Note that, while the average response from LLMs shows framing effects, LLMs are not uniformly susceptible to them. Supplementary Figure S5 shows that 6 models (claude-3-5-haiku, claude-3-5-sonnet, gemini-1.5-pro, gemini-2.0-flash, gpt4, and gpt-4o) display no framing effects.

Lastly, we investigate how different prompts affects LLM performance (Figure 1). Results show that the original prompt ("how likely is it that this headline is true?") achieves the highest accuracy, suggesting it best captures what is being tested in the headline dataset. Variants such as “real” and “genuine” slightly reduce performance, while antonyms like “fake”, “false”, and “altered” cause significant declines. This shows that while LLMs are robust against headline-induced framing effects, they remain sensitive to prompt framing.

4.2 Wisdom and bias of LLM crowds

Building on our understanding of individual LLM behavior, we now examine how groups of LLMs, humans, and hybrid ensembles perform collectively.

Diversity of LLM crowds. Figure 1.C displays the Q-statistic (Equation 2) between the 18 LLMs and 40 of the human participants, while Figure 1.B clusters responders based on their Q-statistics.

Both the close proximity of LLM models in Figure 1.B and the high values in the LLM sub-matrix of Figure 1.C show that LLMs exhibit much higher correlation among themselves than humans. Specifically, the average Q-statistic within human ensembles is 0.387 ± 0.33, while that of LLM ensembles is 0.855 ± 0.08. This implies that when one LLM makes a mistake, others are likely to make the same mistake. Consequently, simply averaging LLM outputs is unlikely to improve performance and may even lead to consensus on incorrect answers, amplifying shared biases.

Notably, the average Q-statistic for hybrid ensembles is 0.548 ± 0.31, significantly lower than that of LLM-only ensembles. This suggests that hybrid groups could complement LLM accuracy with the diversity of human responders.

Static aggregates reinforce biases. The wisdom of the crowd relies on diversity within the group to cancel out individual mistakes. However, simple averages carry individual LLM biases into the aggregate (Table 1, average(LLM) row), as they lack the diversity—e.g., being biased in opposite directions—to allow averaging to mitigate biases. Notably, while only a minority of models show bias for the GENDER-altered categories, their aggregate is significantly biased.

The lack of diversity among LLMs also results in smaller gains from aggregation compared to human groups. For example, aggregation increases human performance from an individual average of 0.55 to 0.611. In contrast, LLM groups improve only slightly, from 0.61 to 0.652. Restricting LLM groups to high-performing LLMs (i.e., LLM+, see Section 3.3) slightly raises the aggregated performance to 0.654, but biases remain unmitigated.

Hybrid ensembles (average(hybrid)), achieve performance levels between purely human and LLM groups. In contrast, selective hybrid ensembles (average(hybrid+)) significantly boost performance, even outperforming groups of exclusively strong LLMs (average(LLM+)). While LLM groups offer strong individual accuracy, their high correlation limits the benefits of additional group members. Partially replacing LLMs with randomly sampled humans introduces greater diversity, allowing the collective to correct more errors and achieve stronger overall performance.

Group Size and Collective Intelligence. Condorcet’s Jury Theorem [Condorcet, 1785] and related principles suggests that larger, diverse groups should achieve higher accuracy, assuming members are reasonably independent and perform better than chance [Surowiecki, 2005].

Figure 2.A shows that, for the simple average, the limited diversity of LLM crowds results in diminishing returns as group size increases. In contrast, human groups exhibit steady performance gains due to greater diversity. Hybrid groups initially perform between LLM and human groups but catch up to LLMs as group size increases, suggesting they benefit from both LLM accuracy and human diversity.

When LLMs are selected for inclusion based on their MMLU benchmark scores (Figure 2.B), the simple average improves significantly. However, as additional LLMs tend to be similar to, but weaker than, already included LLMs, increasing group size leads to decreased performance, as the aggregate is unable to benefit from weaker or redundant contributions. This also explains the stagnation of the simple average for hybrid ensembles beyond group size 4; adding LLMs introduces redundancy, while adding humans provides beneficial diversity. Beyond size 4, these effects cancel each other out, resulting in plateauing performance.

ExpertiseTrees promote the wisdom of the crowd. Figure 2.A demonstrates that locally weighted averages derived from ExpertiseTrees consistently outperform simple averages. While simple LLM averages stagnate beyond group size 4, locally weighted averages continue to improve by leveraging additional LLMs. Interestingly, although small LLM groups outperform human groups, human diversity leads to higher performance in large groups.

The most significant improvements occur in hybrid groups, which outperform both human and LLM groups. To balance the high accuracy of LLMs with the diversity of humans, ExpertiseTrees assign higher weights to well-performing individuals (typically LLMs) and to complementary subgroups of humans. In addition, by potentially maintaining a distinct set of weights for each category, the ExpertiseTree can also capitalize on the specialized strengths of certain human responders. A comparison to regular weighted averages highlights the benefits of this specialization, as ExpertiseTrees consistently outperform them when the group contains humans.

When LLMs and hybrid groups are selected based on MMLU benchmark scores (Figure 2.B), performance further improves. In particular, by having a more informed selection of LLMs, their aggregation outperforms human groups for more group sizes. Note that despite the use of ExpertiseTrees, including more LLMs still decreases performance, as without any useful diversity or improved individual accuracy, the additional LLMs simply introduce more noise. Conversely, hybrid groups continue to benefit from the inclusion of humans, as their diversity complements LLM accuracy.

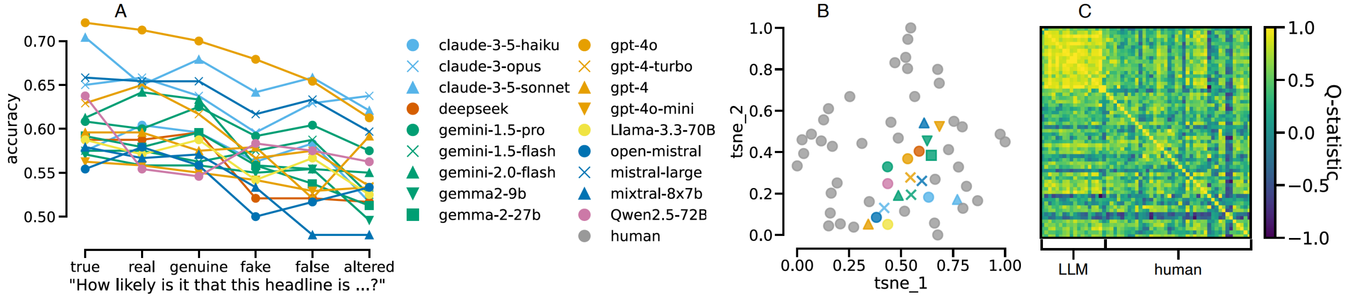

Figure 1: **A.** Model accuracy across prompt variations, with each dot representing the accuracy of a model for a specific prompt ending. **B.** t-SNE visualization of responder diversity. **C.** Q-statistics (see Equation 2) matrix. Each cell gives the Q-statistic between responder pairs.

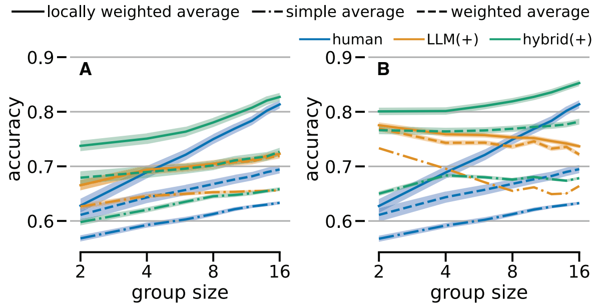

Figure 2: Accuracy for different group sizes and aggregators. Shaded areas show 95% confidence intervals. LLMs are either sampled randomly (**A**) or based on their MMLU scores (**B**).

ExpertiseTrees mitigate biases. Beyond simply improving accuracy, Table 1 shows that ExpertiseTrees mitigate biases. In particular, the ExpertiseTree(LLM) and ExpertiseTree(LLM+) results in fewer significant biases compared to simple and weighted LLM(+) averages. However, the most prevalent bias in individual LLMs persists in the ExpertiseTree aggregates, likely because the lack of diversity within LLM-only groups limits the capacity for bias mitigation.

In contrast, the diversity within human and hybrid crowds allows ExpertiseTrees to mitigate biases displayed by individuals and (weighted) averages of human or hybrid groups.

5 Conclusion

In this work we investigated whether LLMs exhibit similar biases to humans on the headline dataset [Abels *et al.*, 2024]. Human responses to this dataset revealed susceptibility to counterfactual biases and framing effects. Our findings confirm that LLMs reflect those counterfactual biases. All tested LLMs exhibited significant biases, with many favoring headlines reporting positive outcomes for White individuals over Black individuals. Approximately half of the LLMs showed a similar bias favoring women over men. However, LLMs were less affected by framing effects from headline variations, showing greater consistency. Nevertheless, LLMs were susceptible to prompt-induced framing effects, with subtle changes in wording significantly impacting accuracy.

To mitigate individual errors and biases, we explored the “wisdom of the LLM crowd”. Our initial experiments revealed that simple averaging of LLM outputs marginally improved accuracy but reinforced existing biases due to the lack of diversity among LLMs. In contrast, locally weighted averages partially restored the benefits of the crowd, mitigating some biases while improving performance.

Recognizing the complementary strengths of humans (greater diversity) and LLMs (higher individual accuracy), we investigated hybrid crowds that combine both. We found that hybrid crowds outperformed purely human and purely LLM groups in both simple and weighted aggregation approaches. Notably, while locally weighted averages of LLM groups still exhibited counterfactual biases, hybrid crowds achieved improved accuracy without significant biases.

To conclude, our findings highlight the potential of integrating humans and AI within collective intelligence systems. Even modestly sized hybrid ensembles demonstrated advantages, combining the accuracy of LLMs with the diversity of human perspectives to achieve more robust and fair outcomes.

Limitations. While our findings are promising, there are a few important considerations. First, the analysis was conducted using a single dataset. While this provided a controlled environment for systematically comparing biases and performance across LLMs and humans, results may differ for datasets that have different cultural or demographic contexts.

Second, the observed biases and performance reflect the specific versions of LLMs tested. As these models are frequently updated and retrained, future versions may exhibit different behaviors, potentially affecting our findings.

Third, our exploration of diversity focused on hybrid ensembles but did not incorporate techniques to engineer diversity within LLMs, such as fine-tuning or prompting models to adopt varied personas or perspectives. Such approaches could further enhance ensemble diversity and offer additional opportunities for bias mitigation.

Acknowledgments

A.A. is supported by a post-doctoral grant (Chargé de Recherche) by the National Fund for Scientific Research (F.N.R.S.) of Belgium. T.L. is supported by the F.N.R.S. [grant numbers 31257234 and 40007793], the Fonds Wetenschappelijk Onderzoek (F.W.O.) [grant number G.0391.13N], the Service Public de Wallonie Recherche [grant n°2010235–ARIAC by DigitalWallonia4.ai], the Flemish Government through the AI Research Program, and TAILOR, a project funded by EU Horizon 2020 research and innovation program [grant number 952215]. The resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation - Flanders (FWO) and the Flemish Government.

References

[Abels and Lenaerts, 2025] Axel Abels and Tom Lenaerts. Wisdom from diversity: Bias mitigation through hybrid human-llm crowds. https://github.com/axelabels/HybridCrowds, 2025. Accessed: 2025-05-22.

[Abels *et al.*, 2023] Axel Abels, Tom Lenaerts, Vito Trianni, and Ann Nowe. Expertise trees resolve knowledge limitations in collective decision-making. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 79–90. PMLR, 23–29 Jul 2023.

[Abels *et al.*, 2024] Axel Abels, Elias Fernandez Domingos, Ann Nowé, and Tom Lenaerts. Mitigating biases in collective decision-making: Enhancing performance in the face of fake news. *arXiv preprint arXiv:2403.08829*, 2024.

[Achiam *et al.*, 2023] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[Anthropic, 2024] Anthropic. The claude 3 model family: Opus, sonnet, haiku. 2024.

[Bai *et al.*, 2023] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[Becker *et al.*, 2017] Joshua Becker, Devon Brackbill, and Damon Centola. Network dynamics of social influence in the wisdom of crowds. *Proceedings of the national academy of sciences*, 114(26):E5070–E5076, 2017.

[Becker *et al.*, 2019] Joshua Becker, Ethan Porter, and Damon Centola. The wisdom of partisan crowds. *Proceedings of the National Academy of Sciences*, 116(22):10717–10722, 2019.

[Bertrand and Mullainathan, 2004] Marianne Bertrand and Sendhil Mullainathan. Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American economic review*, 94(4):991–1013, 2004.

[Birhane *et al.*, 2021] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021.

[Blodgett *et al.*, 2020] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in nlp. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, 2020.

[Bolukbasi *et al.*, 2016] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

[Breiman, 1996] Leo Breiman. Stacked regressions. *Machine learning*, 24:49–64, 1996.

[Brown *et al.*, 2020] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[Caliskan *et al.*, 2017] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

[Chuang *et al.*, 2024] Yun-Shiuan Chuang, Nikunj Harlalka, Siddharth Suresh, Agam Goyal, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy T Rogers. The wisdom of partisan crowds: Comparing collective intelligence in humans and llm-based agents. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46, 2024.

[Condorcet, 1785] Marquis de Condorcet. Essay on the application of analysis to the probability of majority decisions. *Paris: Imprimerie Royale*, page 1785, 1785.

[Corbett-Davies *et al.*, 2017] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pages 797–806, 2017.

[Davidson *et al.*, 2019] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, 2019.

[Devine, 1989] Patricia G Devine. Stereotypes and prejudice: Their automatic and controlled components. *Journal of personality and social psychology*, 56(1):5, 1989.

[Dovidio *et al.*, 2017] John F Dovidio, Angelika Love, Fabian MH Schellhaas, and Miles Hewstone. Reducing intergroup bias through intergroup contact: Twenty years of progress and future directions. *Group Processes & Intergroup Relations*, 20(5):606–620, 2017.

[Dubey *et al.*, 2024] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[Eagly and Johnson, 1990] Alice H Eagly and Blair T Johnson. Gender and leadership style: A meta-analysis. *Psychological bulletin*, 108(2):233, 1990.

[Gemini *et al.*, 2023] Team Gemini, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[Greenwald *et al.*, 1998] Anthony G Greenwald, Debbie E McGhee, and Jordan LK Schwartz. Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology*, 74(6):1464, 1998.

[Hendrycks *et al.*, 2020] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2020.

[Hong and Page, 2004] Lu Hong and Scott E Page. Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proceedings of the National Academy of Sciences*, 101(46):16385–16389, 2004.

[Liang *et al.*, 2021] Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. Towards understanding and mitigating social biases in language models. In *International Conference on Machine Learning*, pages 6565–6576. PMLR, 2021.

[Liu *et al.*, 2024] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

[Mann and Whitney, 1947] Henry B Mann and Donald R Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60, 1947.

[Mikolov, 2013] Tomas Mikolov. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 3781, 2013.

[MistralAI, 2023] MistralAI. Mistral chat: Advanced conversational ai, 2023. Available at https://mistral.ai.

[Pager, 2008] Devah Pager. *Marked: Race, crime, and finding work in an era of mass incarceration*. University of Chicago Press, 2008.

[Radford *et al.*, 2019] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[Reskin, 2000] Barbara F Reskin. The proximate causes of employment discrimination. *Contemporary sociology*, 29(2):319–328, 2000.

[Schoenegger *et al.*, 2024] Philipp Schoenegger, Indre Tuminauskaite, Peter S Park, and Philip E Tetlock. Wisdom of the silicon crowd: Llm ensemble prediction capabilities match human crowd accuracy. *arXiv preprint arXiv:2402.19379*, 2024.

[Sheng *et al.*, 2020] Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. Towards controllable biases in language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3239–3254, 2020.

[Surowiecki, 2005] James Surowiecki. The wisdom of crowds. *Surowiecki, J*, 2005.

[Tamkin *et al.*, 2023] Alex Tamkin, Amanda Askell, Liane Lovitt, Esin Durmus, Nicholas Joseph, Shauna Kravec, Karina Nguyen, Jared Kaplan, and Deep Ganguli. Evaluating and mitigating discrimination in language model decisions. *arXiv preprint arXiv:2312.03689*, 2023.

[Vaswani, 2017] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[Wilcoxon, 1992] Frank Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in statistics: Methodology and distribution*, pages 196–202. Springer, 1992.

[Williams, 1996] David R Williams. Racism and health: a research agenda. *Ethnicity & disease*, 6(1/2):1–6, 1996.

[Wolpert, 1992] David H Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.

[Wood *et al.*, 2023] Danny Wood, Tingting Mu, Andrew M Webb, Henry WJ Reeve, Mikel Lujan, and Gavin Brown. A unified theory of diversity in ensemble learning. *Journal of Machine Learning Research*, 24(359):1–49, 2023.

[Yule, 1900] George Udny Yule. Vii. on the association of attributes in statistics: with illustrations from the material of the childhood society, &c. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 194(252-261):257–319, 1900.

[Zhang *et al.*, 2020] Guanhua Zhang, Bing Bai, Junqi Zhang, Kun Bai, Conghui Zhu, and Tiejun Zhao. Demographics should not be the reason of toxicity: Mitigating discrimination in text classifications with instance weighting. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4134–4145, 2020.

[Zhao *et al.*, 2017] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2017.

[Zhao *et al.*, 2018] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 2, 2018.

Supplementary Information for
Wisdom from Diversity: Bias Mitigation Through Hybrid Human-LLM Crowds

S1 Supplementary Information

Table S2 presents an overview of the LLMs we considered in this work, including the version, estimates of the number of parameters, their MMLU [Hendrycks *et al.*, 2020] score, the company which initially developed them, as well as the provider we used. Note that several models are proprietary, and their sizes are therefore estimated based on their cost, their response times, as well as their performance.

S1.1 Prompt

Figure S3 shows the prompt we presented to the LLMs:

Where, like in the prompt presented to human participants in [Abels *et al.*, 2024], `target_str` is "true" unless specified otherwise. For Figure 1 we explore different values for `target_str`.

As multi-shot learning [Brown *et al.*, 2020] has been shown to improve accuracy and compliance with instructions (here, the expected answer format), we include for each headline 4 example headline-response sequences. Example headlines are sampled randomly from the same category (i.e., age headlines will be presented with examples on age headlines) and are balanced across the different statuses (genuine or altered) and sentiments (positive or negative). Since the dataset contains both genuine and altered variants of the same headline, we ensured that the examples never contained the alternative version of the queried headline. To reduce variability across models, we fixed the temperature to 0, such that the most likely token was returned. The responses are then parsed to extract the label ([1-5]), which, as in [Abels *et al.*, 2024], are then mapped to numeric values $\{0, 0.25, 0.5, 0.75, 1\}$.

Responses which failed to answer the query, such as those containing system-level disclaimers ("As an AI model, I...") were re-queried for consistency.

S1.2 ExpertiseTrees

ExpertiseTrees [Abels *et al.*, 2023] are an advanced aggregation method designed to partition the context space (e.g., headline categories) and fit localized aggregations for more effective decision-making. They extend the principles of decision trees by incorporating model-based predictions at the leaves, enabling them to dynamically adapt to varying contexts while maintaining interpretability.

Structure and Function

Similar to decision trees, ExpertiseTrees recursively split the problem space, with each split chosen to maximize the performance improvement of the tree. Nodes in the tree correspond to subsets of the context space, such as gender-related or ethnicity-related headlines. This partitioning enables ExpertiseTrees to isolate specific contexts where distinct prediction patterns or biases may occur, tailoring the aggregation process to the nuances of each subset.

Unlike traditional decision trees, where the leaves contain constant values or simple averages, the leaves of an ExpertiseTree contain models. Specifically, these models are linear combinations of individual predictions, weighted to reflect the relevance and reliability of each member's contribution within that context. This allows ExpertiseTrees to dynamically adapt aggregation weights based on context-specific patterns, enhancing both accuracy and fairness.

Advantages Over Traditional Approaches

ExpertiseTrees offer several advantages compared to static aggregation methods like simple averaging or traditional stacking [Breiman, 1996]:

- **Context Sensitivity**: By partitioning the context space, ExpertiseTrees can identify and exploit differences in performance or bias across contexts (e.g., headline categories). For example, an ExpertiseTree might assign higher weights to certain predictors for gender-related headlines and adjust those weights for ethnicity-related headlines.
- **Dynamic Weighting**: Unlike static methods that assign fixed weights to predictors, ExpertiseTrees adjust weights at the leaf level based on the specific context. This ensures that the most reliable and unbiased predictors contribute more heavily to the aggregate output within each subset of the context space.
- **Improved Bias Mitigation**: By isolating subsets of the data where certain individuals or models perform better or exhibit less bias, ExpertiseTrees can mitigate biases more effectively than methods that aggregate across the entire dataset. For example, if certain predictors are prone to counterfactual bias in ethnicity-related headlines but perform well in gender-related contexts, ExpertiseTrees can selectively downweight their contributions.
- **Interpretability**: The linear combination models at the leaves ensure that the aggregation remains interpretable, providing clear insight into how individual predictions are combined within each context.

By routing predictions through the tree structure, ExpertiseTrees effectively balance the strengths and weaknesses of individual contributors, resulting in improved accuracy and fairness across headline categories.

S1.3 Framing Effects

Figure S4 visualizes the framing effects observed across the three group types (LLM-only, human-only, and hybrid) under either human or ExpertiseTree aggregation. Each boxplot represents the distribution of framing effects (see Section 3.2), quantified as $\Delta_F(\text{sentiment}, \text{man})$ for gender, $\Delta_F(\text{sentiment}, \text{old})$ for age, and $\Delta_F(\text{sentiment}, \text{white})$ for ethnicity. Since every headline presents a contrasting statement for a complementary demographic group, the distributions for the complementary groups (not shown here) would simply be inverted versions of those shown in Figure S4.

Name	Version	Size	MMLU	Company	Provider
Qwen2.5-72B-Instruct		72B	86.1	Alibaba	HuggingFace
claude-3-5-haiku	20241022	20B[1]	75.2	Anthropic	Anthropic
claude-3-5-sonnet	20240229	137B[1]	88.7	Anthropic	Anthropic
claude-3-opus	20240229	137B[1]	86.8	Anthropic	Anthropic
DeepSeek	V3	18x37B	87.1	DeepSeek AI	DeepSeek AI
gemini-1.5-flash	September 2024	32B[1]	78.9	Google	Google
gemini-1.5-pro	September 2024	120B, 1.5T[1]	85.9	Google	Google
gemini-2.0-flash-exp	December 2024	32B[1]	87	Google	Google
gemma-2-9B		9B	71.3	Google	Groq
gemma-2-27B		27B	75.2	Google	HuggingFace
gpt-4	0613	175B[1]	86.4	OpenAI	OpenAI
gpt-4-turbo	2024-08-06	175B[1]	86.4	OpenAI	OpenAI
gpt-4o	2024-08-06	1T[1]	88.7	OpenAI	OpenAI
gpt-4o-mini	2024-07-18	8B[1]	82	OpenAI	OpenAI
Llama3.3-70B		70B	82	Meta	HuggingFace
mistral-large-latest	24.11	123B	84	Mistral	Mistral
mixtral-8x7b		8x7B	70.6	Mistral	Groq
open-mistral-nemo		12B	68	Mistral	Mistral

Table S2: Model summary. [1] denotes estimates from various sources of closed models based on their performance, response time, and cost. MMLU column provides the reported performance of the models on the MMLU benchmark.

```
How likely is it that this headline is {target_str}.

 Choose one of the following options and return only the number of that option:
1. very unlikely, 2. unlikely, 3. undecided, 4. likely, 5. very likely.

<examples>
"{example_headline1}"
Response: {expected_response1}
"{example_headline2}"
Response: {expected_response2}
"{example_headline3}"
Response: {expected_response3}
"{example_headline4}"
Response: {expected_response4}
</examples>

"{response_headline}"
Response:
```

Figure S3: The prompt presented to the LLMs.

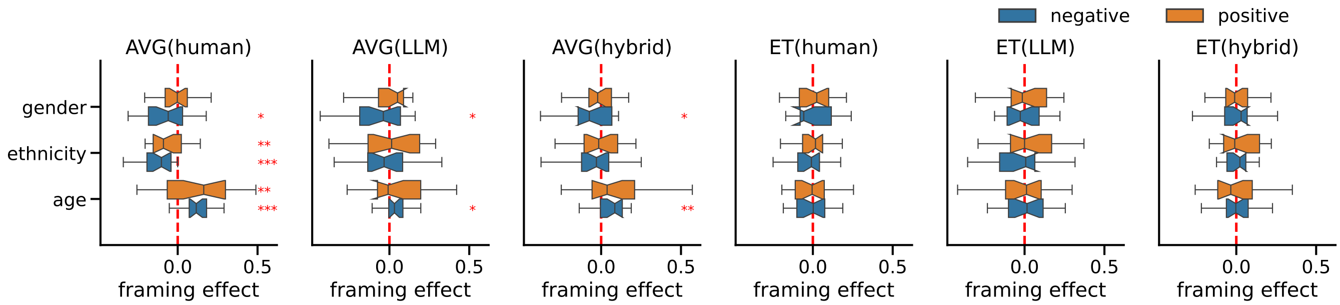

Figure S4: Distribution of framing effects for headlines reporting positive or negative outcomes across three demographic groups. Asterisks indicate statistical significance of the framing effects (Wilcoxon test, *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$).

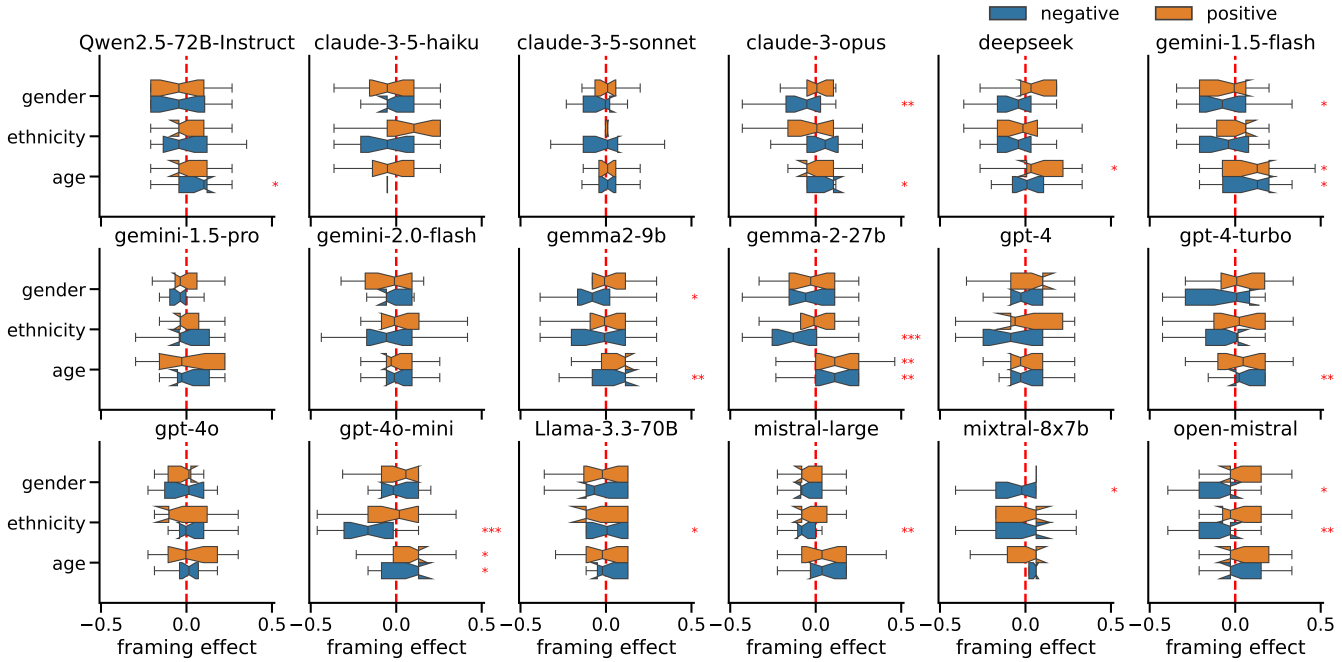

Figure S5: Distribution of framing effects for headlines reporting positive or negative outcomes across three demographic groups. Asterisks indicate statistical significance of the framing effects (Wilcoxon test, *: $p < 0.10$, **: $p < 0.05$, ***: $p < 0.01$).