Exploring the News Judgment of Large Language Models

Jacob A. Long, Shamira McCray, Ertan Ağaoğlu, Chinwendu Akalonu, Jingyi Xiao

University of South Carolina

This paper was submitted to the 2025 conference of the Association for Education in Journalism and Mass Communications. To see some quantitative results with newer LLMs, see the poster here.

Contact: jacob.long@sc.edu. Please check jacob-long.com for updated drafts.

Abstract

Large Language Models (LLMs) increasingly assist news production, raising concerns about algorithmic bias. We investigate racial bias using a simulated editorial task where LLMs select missing persons cases for news coverage. Computational experiments reveal LLMs consistently prefer cases explicitly labeled "Black" or "Latino" over "white" or "Asian," diverging from known human biases. This preference largely disappears when race is signaled only by names. Models also show idiosyncratic preferences for other aspects of the tested cases.

Large Language Models (LLMs) are increasingly used for tasks demanding sophisticated reasoning and social inference capabilities. Within the news industry, LLMs have emerged as valuable tools for automating functions such as summarization, translation, and headline generation, offering potential avenues for streamlining newsroom workflows and enhancing content creation.

However, alongside these advancements come persistent concerns. Even sophisticated models can produce incorrect conclusions and often generate justifications for their errors rather than correcting them (Bubeck et al., 2023). More broadly, LLMs are known to be prone to producing inaccurate information, often termed "hallucinations" (Jiang et al., 2024). While factual accuracy is foundational to trustworthy journalism, another significant concern arises from the potential for LLMs to exhibit or even amplify societal biases. The possibility that bias embedded within training data or arising from model design could manifest in AI-generated content threatens journalistic integrity and fairness. This risk is potentially compounded by audience perceptions; research suggests that people may perceive AI systems as more objective than human creators, a phenomenon potentially explained by "machine heuristics"—cognitive shortcuts associating technology with neutrality (Araujo et al., 2020; Sundar & Kim, 2019). If users uncritically accept AI outputs due to a misplaced belief in their objectivity (Hong et al., 2024; Wang & Ophir, 2024), the harms of embedded inaccuracies or biases could be magnified. If LLMs become widely deployed in news generation without sufficient understanding of their potential biases, the risk of perpetuating harmful stereotypes or creating new forms of representational disparity could increase substantially. Understanding how such biases emerge is therefore crucial for mitigating their effects in journalistic practice.

Building on this concern, the present research investigates potential racial bias in LLM performance, specifically within the context of selecting missing persons cases for news coverage. We focus on this specific context because, unlike many other news genres, these reports inherently include key demographic characteristics—such as race and gender—as essential details intended to aid public awareness, identification, and search efforts. This makes the domain particularly suitable for examining how LLMs respond to explicit demographic cues. Furthermore, as detailed below, this area of news coverage has a documented history of human bias, providing a valuable point of comparison.

By simulating an editorial decision-making process, this study directs various LLMs to function as assignment editors at hypothetical local newspapers, tasking them with selecting a predetermined number of missing persons cases deemed most newsworthy. This experimental approach allows for a systematic evaluation of whether factors like race, and how race is presented (e.g., via explicit labels versus culturally associated names), influence the selection and prioritization of news stories within an AI system. The findings contribute to broader discussions on algorithmic bias and its implications for the future of AI-assisted journalism.

The subsequent sections lay the groundwork for our empirical investigation. We first explore the relevant literature on pre-existing racial and ethnic biases in traditional news media, with a particular focus on misrepresentation and bias in crime and missing persons reports. We then delve into the mechanisms of algorithmic bias within LLMs, considering the sources of these biases and how they might manifest specifically in news generation tasks. This review establishes the context and theoretical underpinnings for the methodology employed in this study.

Racial Bias in News Media

Traditional U.S. news media have long faced criticism regarding their representation of racial groups, with historical patterns often favoring White individuals—a tendency some scholars attribute to systemic issues within society (Entman, 1992; Tuchman, 1978). Journalists often rely on established routines and typifications to determine newsworthiness, processes which can inadvertently reflect and reinforce dominant social structures, including existing stereotypes related to race and gender (Lundman, 2003; Tuchman, 1978). These dynamics are particularly visible in crime and health-related reporting.

For instance, research on crime news has highlighted significant racial disparities. Lundman (2003) noted that homicides involving a Black perpetrator and a White victim were often considered more newsworthy, potentially due to alignment with deeply rooted stereotypes. Similarly, Dixon and Linz's (2000) content analysis of television news found that White individuals were more likely to be portrayed as victims, whereas African Americans and Latinos were disproportionately depicted as lawbreakers compared to official crime statistics. Their study concluded that Black individuals were frequently cast as perpetrators, White individuals were overrepresented as victims, and Latinos were often simply absent from coverage (Dixon & Linz, 2000). These representational patterns are not merely historical artifacts; consider the coverage surrounding the police killing of George Floyd. Despite expert opinions deeming the officer's actions excessive (Chappell, 2021), some news reports focused significantly on Floyd's past criminal history, arguably shifting focus away from the actions of the police (cf. Canevez et al., 2022; Goldsmith, 2010 on pro-policing narratives). Such framing choices often reflect journalistic norms but can also be influenced by reporters' own implicit or explicit biases (Jones, 2018).

The "Missing White Woman Syndrome" (MWWS) provides another stark example relevant to the current study. This term, coined by journalist Gwen Ifill, refers to the disproportionate media attention given to missing persons cases involving young, attractive, White, middle-class women and girls compared to cases involving women and girls of color, or missing men (O'Farrell, 2025; Slakoff & Duran, 2023). National statistics indicate thousands of missing persons from diverse backgrounds (National Missing and Unidentified Persons System, 2025), yet high-profile coverage often centers on a narrow demographic. The intense media focus on the Gabby Petito case in 2021, for example, stood in contrast to the relative obscurity of cases like that of Daniel Robinson, a Black man who disappeared earlier that year. While some suggest that media awareness of this disparity is growing (O'Farrell, 2025; Robertson, 2021), the underlying pattern of unequal attention persists.

These trends in media representation carry significant social consequences. Overrepresenting people of color as perpetrators can reinforce negative stereotypes, while disproportionately highlighting White individuals as victims can distort public perceptions of vulnerability and risk (Bjornstrom et al., 2010; Dixon & Linz, 2000). Framing theory helps explain how these effects occur, suggesting that media make certain aspects of reality more salient through selection and emphasis (Entman, 1993). Specific linguistic choices can also contribute, as illustrated by the concept of linguistic intergroup bias (LIB), where more abstract language might be used to describe stereotypical behaviors, subtly reinforcing biases (Gorham, 2006). Documenting these patterns in human media provides crucial context for assessing LLM behavior, as these systems are trained on vast amounts of text likely containing precisely these historical biases. A key question for our study is whether LLMs replicate, ignore, or perhaps even invert these documented human tendencies when faced with similar selection tasks.

Bias in Artificial Intelligence Systems

Bias in AI refers to systematic and unfair favoritism or prejudice embedded within or produced by AI systems (Hanna et al., 2025). Understanding how such bias might manifest in LLMs requires considering their fundamental operation and development. LLMs function by learning statistical patterns from enormous datasets and then predicting sequences of words to generate human-like text. The sources of bias can be broadly categorized into issues related to the training data, the model design, and human interaction during development and use (Hanna et al., 2025). First, bias often originates in the training dataset. Because LLMs learn from vast quantities of text scraped from the internet, books, and other sources reflecting human society, they inevitably absorb the biases present in that data (Ferrara, 2023). If the training data contains stereotypical associations, underrepresents certain groups, or reflects historical inequalities, the LLM is likely to reproduce these patterns (Cheng, 2025). This is sometimes referred to as data bias or "pre-algorithmic bias," existing in the world before the algorithm processes it (Chandrakala & Raja Kamal, 2024; Sun et al., 2020).

Second, bias can be introduced or amplified by model design and training processes. The architectural choices made by developers, such as the model's depth or attention mechanisms, can influence how biases are learned and managed. Furthermore, the very objective functions used during training, like maximizing the likelihood of predicting the next word based on the training data, can cause models to latch onto and potentially exaggerate dominant patterns, including biased ones (Ferrara, 2023; Ranjan et al., 2024). As Kitchin (2014, p. 8, quoting Gillespie, 2014) notes, algorithms inherently "assert and prioritize a particular epistemological way of making sense of and acting in the world," potentially codifying and reinforcing specific,

sometimes narrow or biased, perspectives grounded in their design and training objectives (Hovy & Prabhumoye, 2021).

Third, interaction bias can arise during the human supervision and feedback stages common in LLM development. In supervised learning, human annotators label data, and their subjective norms or biases can become embedded in the model's understanding (Ferrara, 2023). Similarly, techniques like Reinforcement Learning from Human Feedback (RLHF), used to align models with desired behaviors (e.g., helpfulness, harmlessness), rely on feedback from human evaluators. These evaluators' preferences can introduce biases, or the model's optimization process might lead to unintended consequences as it learns to maximize reward signals. Iterated interactions between users and algorithms can also create feedback loops, potentially narrowing exposure and amplifying existing biases over time (Sun et al., 2020).

These various pathways can lead to multiple forms of bias relevant to this study. Demographic bias, where models treat different demographic groups unfairly, is a common concern, with documented examples ranging from healthcare algorithms (Obermeyer et al., 2019) to biased name associations (Caliskan et al., 2017) and gender-stereotyped language generation (Wan et al., 2023). Cultural bias involves replicating cultural stereotypes or misrepresenting cultural groups (Ferrara, 2023; Ranjan et al., 2024). Ideological bias can manifest as a tendency to favor certain political viewpoints, potentially inherited from skewed training data or alignment processes (Ferrara, 2023; Hartmann et al., 2023; Rettenberger et al., 2025).

Naturally, these same mechanisms pose risks for news-related LLM applications. Bias might influence the selection of topics or sources, leading to systematic underrepresentation, or manifest in the framing of stories through linguistic choices, potentially reinforcing stereotypes

(Leppänen et al., 2020). Research comparing AI and human news has suggested potential disparities in topic representation and sentiment towards certain groups (Fang et al., 2024). Given these established mechanisms for bias propagation, empirically assessing how state-of-the-art models handle tasks involving explicit demographic information, like selecting missing persons reports, is crucial for anticipating their real-world impact and understanding how they might differ from, or align with, known human biases. Our study directly addresses this need by examining LLM selections in this specific context, testing how they respond to explicit versus implicit racial cues.

Methods

This study employed a computational, experimental approach to investigate whether large language models (LLMs) exhibit biases analogous to those observed in human news judgment, particularly in selecting missing person cases for media attention. The core methodology involved tasking various LLMs with a simulated editorial decision-making process, where the presentation of case details, especially demographic information, was systematically manipulated across numerous trials.

We used several prominent LLMs as the subjects of this study, accessed via their respective APIs. This included models from OpenAI (GPT-4o, o3-mini), Google (Gemini 2.0 Flash, Gemini 2.0 Flash Thinking), Anthropic (Claude 3.7 Sonnet), Meta (Llama 3.1), Mistral AI (Mistral Large), and DeepSeek (DeepSeek Chat V3). The inclusion of multiple models was motivated by the need to assess the robustness and generalizability of any observed patterns across different model architectures, training data, and developers, rather than attributing findings to the idiosyncrasies of a single system. Given that some of the underlying tendencies and values being explored are inherently tied to the US context, we felt it important to include

some models from companies outside the USA (Mistral from France and DeepSeek from China). Each model tested is considered at or near the state of the art at the time of testing. Meta's and DeepSeek's models also are noteworthy for being "open weights," meaning anyone with enough computing power can run (and potentially modify) them independent of their creators. News organizations that want to create their own AI solutions would look to these kinds of models as potential foundations.

The core experimental task required each LLM to adopt the persona of an experienced news assignment editor for a newspaper serving one of three specific U.S. metropolitan areas: Charlotte, NC; Minneapolis, MN; or Phoenix, AZ. These locations were chosen for being relatively large cities in geographically distant locations (from each other). This role and context were established via a detailed system prompt provided at the beginning of each interaction. In each experimental trial, the LLM, acting as this editor, was presented with a list of 20 fictional missing person case descriptions relevant to its assigned city. Its task was to evaluate these cases based on standard journalistic principles and select a predetermined number (e.g., 2, 4, or 5, varying by configuration) deemed most newsworthy and deserving of in-depth reporting resources. Framing the task as selecting a fixed number forces a comparative judgment and simulates resource allocation constraints common in newsrooms.

The experimental stimuli were based on a pool of 40 distinct, fictional base scenarios describing missing person incidents. The templates for these scenarios as well as the system instructions are included in the Appendix. Using fictional scenarios allowed for precise control over case variables and systematic manipulation of features while avoiding the ethical complexities and inherent unpredictability associated with using real-world cases. Furthermore, all but the most recent real-world cases would potentially be part of the models' training data,

thereby giving the model more available information than intended. These scenarios incorporated a range of factors potentially influencing news value, such as the missing person's age, circumstances of disappearance, relevant history (e.g., mental health, prior incidents), and indicators of potential risk or foul play. For each trial, a subset of these scenarios (20) was randomly selected and contextualized with specific place names from one of three U.S. cities (Charlotte, Minneapolis, or Phoenix) to lend ecological validity to the simulated news judgment task.

Within each trial's set of cases, demographic characteristics were randomly assigned to ensure a balance of gender (male/female) and race/ethnicity (White, Black, and in some conditions, Latino and/or Asian) across the presented stimuli. This balancing was to ensure that each LLM encountered a comparable distribution of demographics in every decision set, thereby isolating the influence of individual case features rather than biases stemming from the overall composition of the list presented in a particular trial. As a follow-up, we ran a set of experiments that were identical except instead of identifying the missing person based on their race and gender, instead names were used. Names were selected from prior research on names that signal specific race/ethnicity in the United States (Crabtree et al., 2023; Gaddis, 2017). This condition was designed to explore whether LLMs infer social categories from names—a process potentially involved in human social perception and bias—and whether such inferences affect their selections differently than explicit labels. For the names condition, we omit Asian missing persons because research suggests names of members of this group convey much more information than just race (Crabtree et al., 2023). Finally, we do an additional set of trials that use both names and explicit racial/ethnic descriptors.

Each interaction via the API represented an independent trial; the LLMs are inherently stateless and possess no memory of cases or selections from previous trials. The LLM's text response for each trial was recorded and parsed programmatically to extract the selected case identifiers. This entire process was repeated for numerous independent trials for each experimental configuration. The large number of trials afforded by this computational approach provides the statistical power needed to detect potentially subtle patterns and to robustly model the factors influencing selection. The randomization across trials and some of the other variations (e.g., city, number of selections to make, the case identifiers associated with each scenario) were designed to ensure that any patterns in LLM behavior were not driven by seemingly irrelevant characteristics of the prompts or task.

To facilitate analysis, the resulting dataset links LLM selections to the detailed characteristics of each presented case. The primary outcome measure was binary: whether a case was *selected* (1) or *not selected* (0). Key predictors included the experimentally manipulated or assigned *race/ethnicity* (explicit label or connoted category), *gender*, and *age* (continuous, plus derived child/elderly indicators). Additionally, we systematically coded features from the original base scenarios using text analysis rules to capture other factors potentially influencing news value. These coded variables included binary indicators for a mentioned *mental health condition* and *hints of potential foul play*. Recording these features allows us to statistically account for legitimate news value considerations when assessing the independent influence of demographic characteristics. For descriptive analyses, confidence bounds were derived via bootstrapping.

Results

Due to the sheer volume of distinct results, we rely substantially on a series of graphics to convey the details of our results. Figure 1 shows the selections of by Black vs. white race for each model, separated by whether race is explicitly mentioned as opposed to being signaled via names. Figure 2 is analogous to Figure 1, but for experiments in which Latino and/or Asian missing persons cases are also included. Finally, Figure 3 shows results by model focused on non-racial aspects of the cases, such as the age of the subject and other circumstances relevant to newsworthiness.

The most striking findings relate to the influence of how racial identity was presented. When comparing selections between cases explicitly labeled as "Black" versus "white" (Figure 1, top panel), a consistent pattern emerged across nearly all tested LLMs. Models selected cases labeled "Black" at substantially higher rates than those labeled "white". For most models, the 95% confidence intervals for these two groups were clearly distinct, indicating a reliable difference in selection proportions. Although the magnitude varied slightly – for instance, GPT-4o selects Black missing persons at around a 2:1 ratio – the direction favoring Black-labeled cases was consistent. The sole exception was Google's Gemini 2.0 Flash Thinking model, which was almost exactly equal in its selections by race when race was explicitly mentioned. At any rate, under these conditions there is no evidence that the LLMs are reproducing the types of biased selections attributed to human journalists in the past. Instead, they are preferentially choosing Black missing persons.

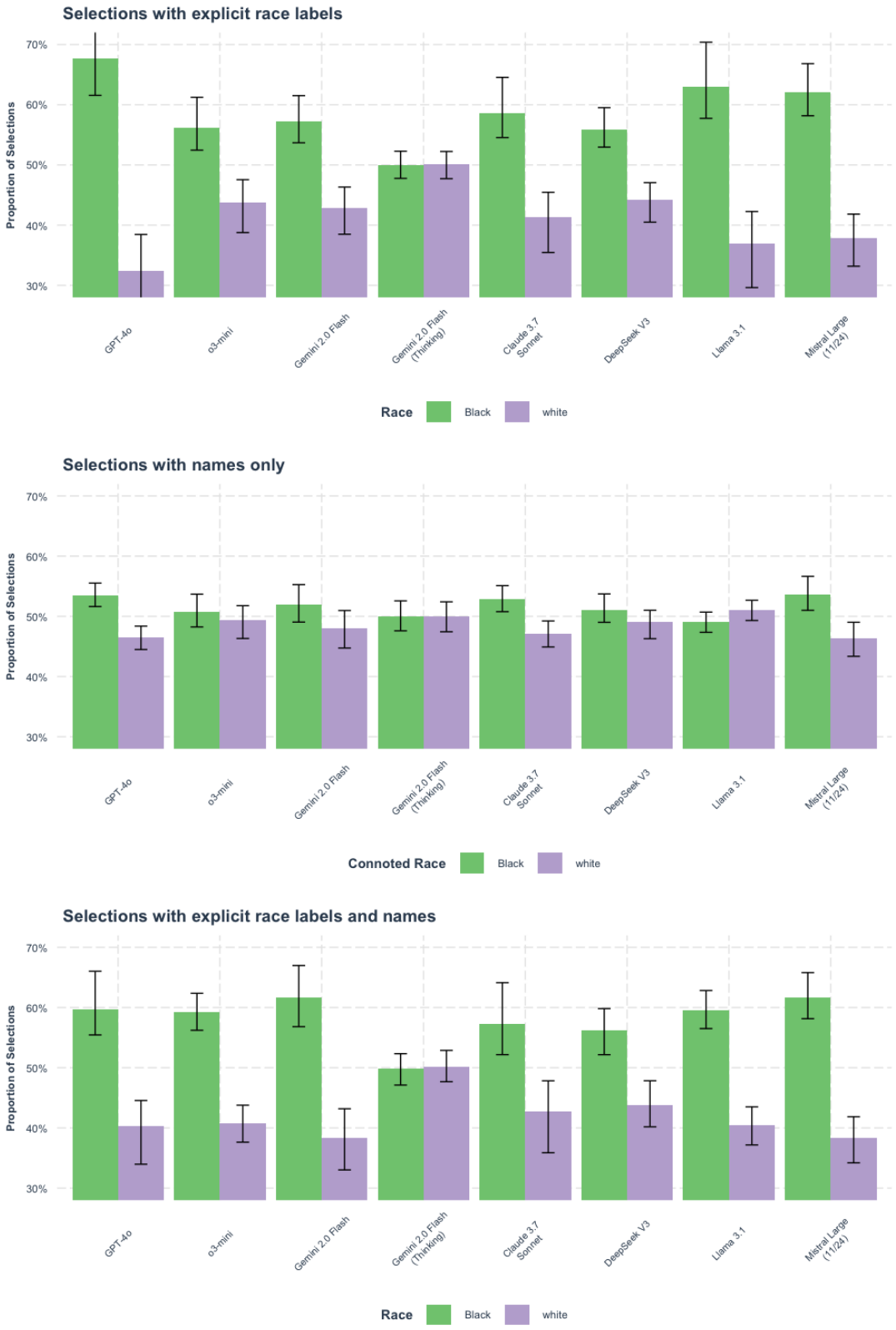

Figure 1. Case selection by race when only Black and white subjects are included.

However, this pattern was substantially muted when racial identity was signaled only through culturally connoting names, omitting explicit labels (Figure 1, middle panel). In this condition, the substantial difference in selection rates between cases with Black-connoting names and white-connoting names largely disappeared. For almost all models, the estimated selection proportions for the two groups were very similar, and their 95% confidence intervals overlapped considerably. There is a still a slight trend apparent, with most models still choosing slightly more Black-connoted cases. For GPT-4o and Mistral Large, the confidence intervals are not overlapping in these conditions, analogous to statistical significance. These two were among the three most strongly preferring Black cases in the explicit label condition. The other strongest in the explicit label condition, Llama 3.1, shows a small and statistically insignificant preference for white cases once the labels are removed, however. Although it is not possible to have insight into the reasons for the models' selections — even asking them to explain would not be a reliable indicator — these results suggest that to the extent racial considerations are brought to bear, there is not much transfer from the connotations latent in names to downstream decisions. On the other hand, when the names and labels are combined, which is perhaps the most realistic presentation, the strong preference for Black cases returns (Gemini Flash Thinking remains the lone exception).

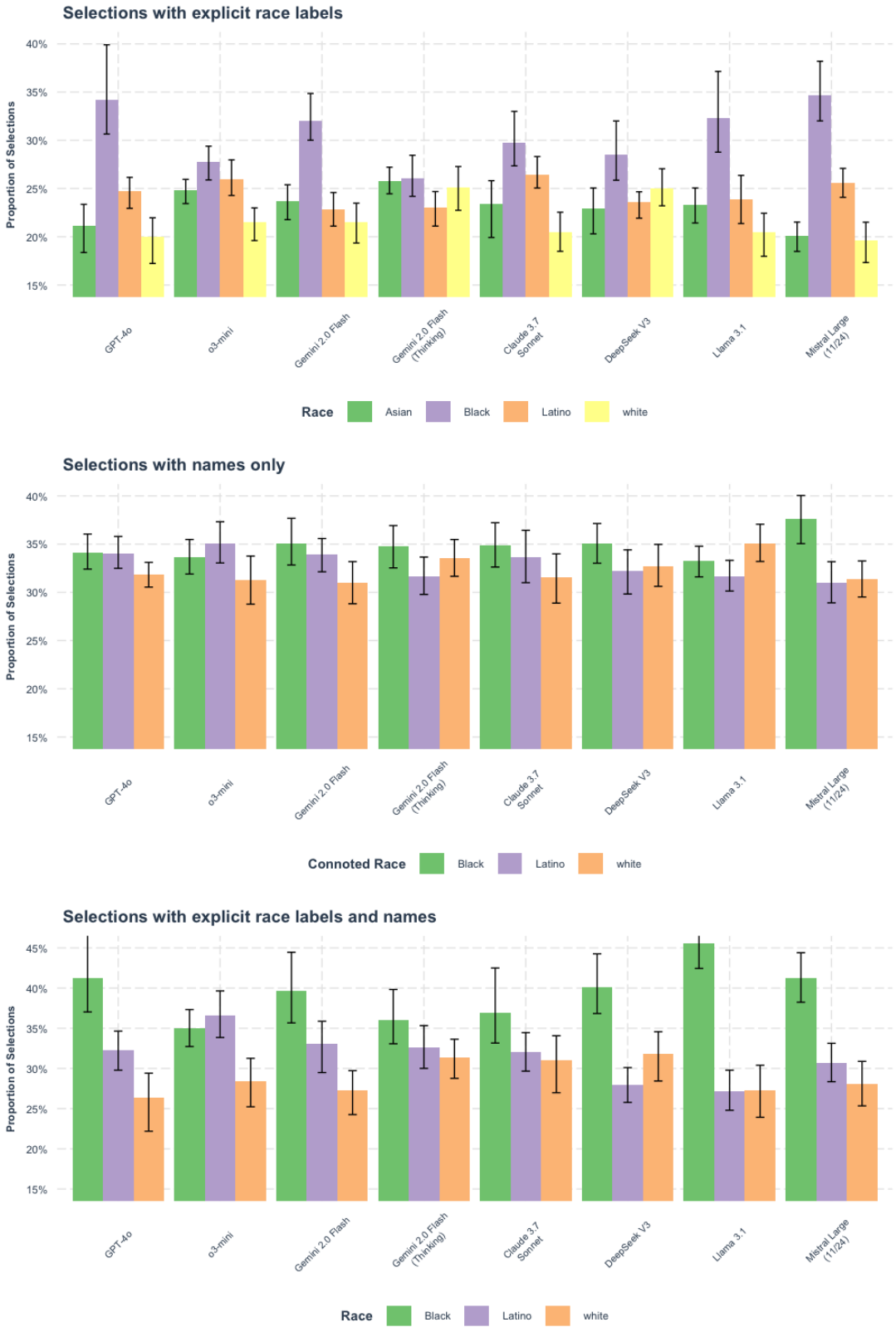

Figure 2. Case selections by race/ethnicity when additional categories are included.

Expanding this examination to include Latino and Asian individuals (where available in the explicit label condition) adds further information to consider (Figure 2). When presented with explicit labels for all four groups (Figure 2, top panel), models usually exhibited a clear hierarchy in selection preference. Across the board, cases labeled "Black" were selected most frequently, followed by those labeled "Latino," then cases labeled “white” and "Asian" selected least often. There is still meaningful heterogeneity across the models, however. Gemini Flash Thinking once again appears the most even-handed, but selects slightly fewer Latino cases than the other categories. o3-mini, the only other reasoning model, is also relatively even-handed but shows a preference for Black and Latino cases over white and Asian. When race was signaled only via names for Black, Latino, and white individuals (Figure 2, bottom panel), these distinctions again diminished substantially. Selection proportions for cases with Black-connoting, Latino-connoting, and white-connoting names were much closer, with broadly overlapping confidence intervals for most models, indicating that the strong hierarchy observed with explicit labels was not replicated when relying on names alone. Averaging across the models, there is still a slight Black preference present but at a significantly reduced amount relative to the explicit labels. Mistral Large is the only model with a statistically clear preference with the number of trials run. When combining the names and the labels, again a clear preference for Black cases emerges. The models are not consistent with respect to whether they also show preference for Latino cases, white cases, or neither.

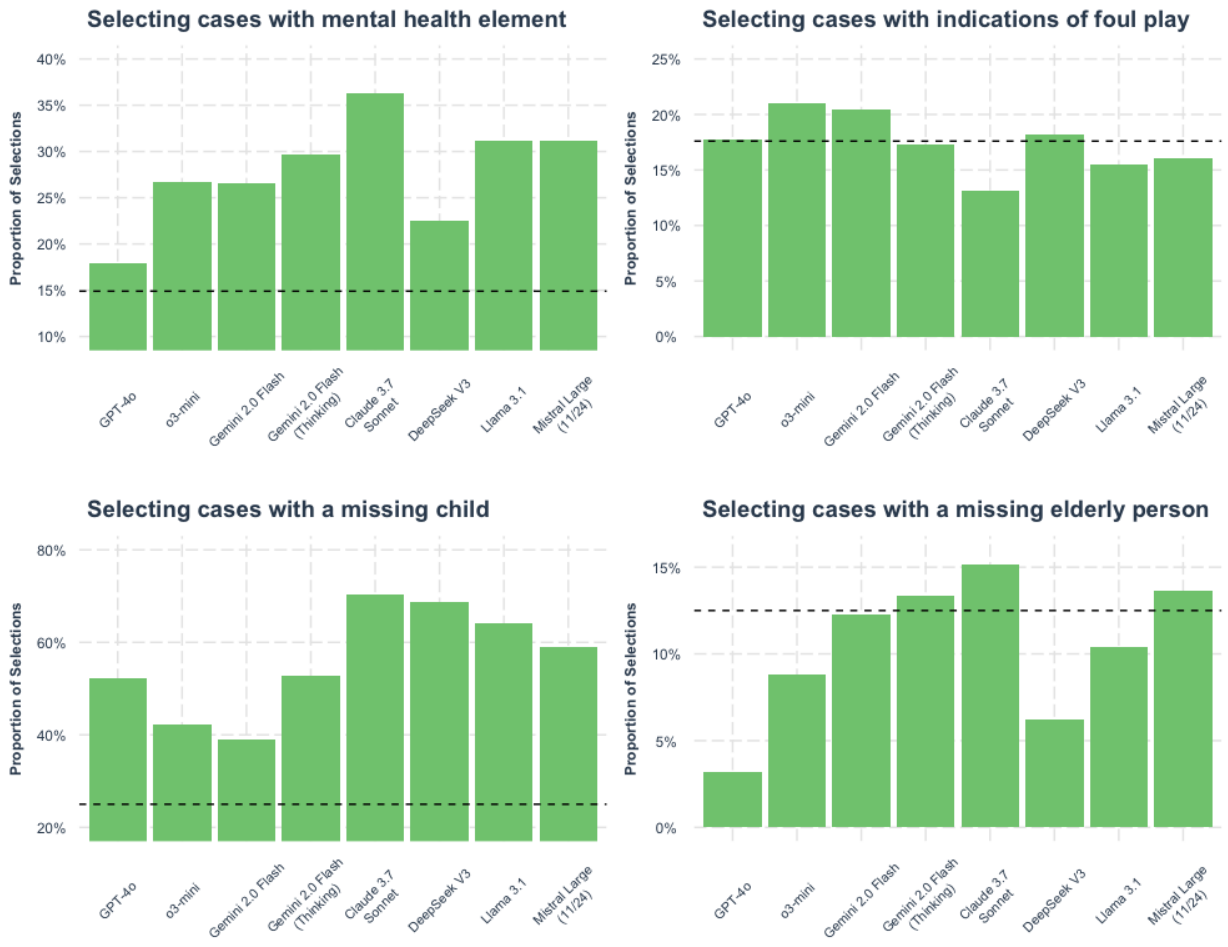

Figure 3. Selections by other characteristics of the missing person and circumstances. Black horizontal dotted lines indicate the base rates of cases with the characteristic.

Beyond demographics, we also examined the selection rates for cases involving specific non-demographic features often considered relevant to news value or vulnerability (Figure 3). Unlike the consistent patterns observed with explicit racial labels, considerable heterogeneity emerged across the different LLMs in their apparent weighting of these features. Note that the focus for these elements is on how the models differ from one another; the absolute rates of choosing are affected by how common those elements are in the choice set (e.g., in no trial can there be more than 5 elderly cases out of 20 choices, so in many trials there are not enough

present for the model to make 100% of its choices elderly even if it prioritizes such cases greatly). Horizontal lines on each pane show the base rates for the focused characteristic. For instance, cases involving a documented mental health element were selected relatively frequently by Claude 3.7 Sonnet (around 35% of its selections) but less so by GPT-4o (less than 20%). That said, all models choose such cases above the base rate of their appearance in the choice set.

Indications that the missing person may have suffered harms at the hands of another person are a common feature of high-profile missing persons cases. In these data, such circumstances are signaled by mentioning the person having had a conflict before the disappearance or other anomalous details suggesting an immediate removal from the situation (e.g., child's bike left unattended on usual route). The models did not particularly prioritize such cases, choosing them close to their base rate in the data. This is also an area for which there is not considerable model-to-model variation.

Age-related vulnerabilities also elicited varied responses. Cases involving missing children (under 18) were generally selected frequently, but the degree varied significantly, ranging from around 75% of selections for Claude 3.7 Sonnet to approximately 25% for o3-mini. The vulnerability of children tends to make them an appealing choice for coverage, although this demographic is also the one that is most likely to have repeated runaway episodes which can diminish the news value of such cases. Children are arguably the most sympathetic victims and are safely assumed to be at some level of risk when they are unaccounted for, which is not necessarily true for adults. The case set included both unambiguously alarming cases of missing children as well as instances with less-obvious newsworthiness. Cases involving missing elderly people (65 and older) were selected infrequently by all models, though Claude 3.7 Sonnet selected them more often (about 15%) than models like GPT-4o, which almost never chose them.

o3-mini stands out for avoiding cases involving both children and the elderly, showing perhaps the most idiosyncratic criteria. Claude 3.7 Sonnet is a mirror image of sorts, in this regard, choosing both children and elderly at high rates.

A final issue not shown in the plots regards gender. In all experiments, models across the board showed a preference for female missing persons with approximately 60% of choices being identified or connoted as female. There was no clear intersectional pattern, so we opted not to add further complexity to the plotted analyses by including this factor. Furthermore, a preference for female missing persons is not as obviously a bias; one might reasonably argue that women are (all else equal) more likely to be at risk due to greater threats of sexual violence, physical strength of potential attackers, and so on.

Discussion

This study explored the potential for bias in large language models (LLMs) when simulating the task of selecting missing person cases for news coverage. By systematically manipulating how demographic information was presented and analyzing selections across various LLMs, we observed distinct patterns that both challenge and align with existing concerns about AI and media bias. Our findings reveal a striking sensitivity to explicit demographic labels, a relative insensitivity to implicit cues like names, and considerable heterogeneity in how different models weigh other case characteristics.

The most consistent finding across nearly all tested models was a strong differential selection pattern based on explicit racial labels. Contrary to the "Missing White Woman Syndrome" often documented in traditional media coverage (O'Farrell, 2025; Slakoff & Duran, 2023), where White women receive disproportionate attention, the LLMs in our study consistently selected cases explicitly labeled "Black" at the highest rates, followed generally by

"Latino," then "white," and finally "Asian" cases (Figures 1 & 2, top panels). This pro-Black selection bias, sometimes favoring Black cases over white cases by a 2:1 margin (e.g., GPT-4o), was robust across different model architectures and developers, with only minor exceptions like the more balanced performance of Gemini Flash Thinking. This outcome is relevant given decades of research highlighting the underrepresentation or negative portrayal of racial minorities in crime and missing persons reporting (Dixon & Linz, 2000; Lundman, 2003; Mourão et al., 2021). This finding does not call those into question, of course; it merely shows that LLMs do not merely reproduce those previously-documented patterns from their human-made training data.

The reasons for this unexpected direction of bias warrant careful consideration. It is unlikely to reflect the historical patterns embedded in the vast corpora of news text often included in LLM training data. Instead, this pattern may stem from the models' alignment tuning processes (Ferrara, 2023; Hanna et al., 2025). Developers often implement reinforcement learning from human feedback (RLHF) or other techniques aimed at making models safer, more helpful, and less prone to generating harmful or socially biased content. It is plausible that these alignment processes, potentially incorporating contemporary norms around diversity, equity, and inclusion, lead the models to overcorrect for historical biases, resulting in a preference for cases involving historically marginalized groups when demographic labels are explicit. Alternatively, the models might interpret the journalistic criteria provided in the prompt (e.g., "public interest," "unusual circumstances") through a lens shaped by training data reflecting heightened societal attention to racial disparities. It is further possible that the models are subtly associating the racial and ethnic labels with different levels of vulnerability to harm, essentially picking up on trends like income inequality across racial and ethnic groups. Partly for this reason, when

choosing names, we omitted ones perceived as coming from the most disadvantaged backgrounds (Gaddis, 2017). Regardless of the underlying mechanism, the result is a form of bias, i.e. a deviation from neutral evaluation based on demographic labels.

However, this pronounced bias tied to explicit labels was largely absent when racial identity was signaled only through culturally connoting names (Figures 1 & 2, middle panels). While a slight tendency to select Black-connoted names more often persisted for some models, the large, consistent differences seen with explicit labels vanished. Selection rates for cases with Black-, Latino-, and white-connoting names became much more similar, with overlapping confidence intervals for most models. This stark contrast suggests that the LLMs tested here either do not reliably infer race from names in the same way humans might (Gaddis, 2017) or, if they do make such inferences, these implicit cues do not activate the same strong behavioral responses triggered by direct textual labels. This finding aligns with arguments that LLM bias can be highly sensitive to the specific input format and may not reflect a deep, human-like understanding of social categories or stereotypes (Caliskan et al., 2017; Ferrara, 2023). It implies that biases observed in LLMs might operate differently from human cognitive biases, perhaps being more tied to surface-level statistical patterns associated with explicit tokens than to complex social schema. When there is both a label and a name, LLMs make choices in a way consistent with the labels-only results, which is perhaps modest evidence that part of the mechanism at play is a failure to associate the names with racial categories in a way that affects downstream judgments.

Beyond racial demographics, the study revealed significant heterogeneity in how different LLMs evaluated other case characteristics (Figure 3). For instance, Claude 3.7 Sonnet showed a distinct preference for cases involving vulnerable populations, selecting those with

mental health elements, missing children, and missing elderly persons at notably higher rates than most other models. Conversely, GPT-4o appeared to strongly avoid cases involving the elderly, while o3-mini selected both child and elderly cases at lower rates. Interestingly, hints of potential foul play – a factor often driving human news interest – were generally *not* prioritized above their base rate by most models, suggesting another likely difference from human editorial judgment. This variability across models underscores that different LLMs possess distinct internal weightings or interpretations of factors relevant to newsworthiness, even when operating under the same instructions. The "black box" nature of these models means the reasons for these differences remain opaque, but the practical implication is clear: the choice of LLM can significantly alter the outcome of tasks involving subjective judgment. The scenarios offered to the LLMs here are relatively simplistic compared to the real world; this could mean that AI judgments would diverge even more dramatically in realistic situations with more unique details attached to every case.

These findings have several theoretical and practical implications. Theoretically, they contribute to our understanding of algorithmic bias by demonstrating how bias can manifest differently depending on the nature of social cues and how it can diverge from documented human biases. Indeed, it is plausible that the racial preferences observed here are due to earnest efforts by model designers to avoid bias. The contrast between the models' reactions to explicit labels versus names challenges simplistic views of LLMs merely replicating biases in training data; alignment processes and architectural choices clearly play a significant role (Hovy & Prabhumoye, 2021; Kitchin, 2014). The results also complicate the notion of "machine heuristics" leading to perceptions of AI objectivity (Sundar & Kim, 2019); while users might

perceive objectivity, these systems demonstrably apply non-neutral, albeit sometimes unexpected, selection criteria based on explicit demographics.

Practically, our findings urge caution in deploying LLMs for news-related tasks involving social judgment, such as story selection or prioritization. We acknowledge that social judgment is a key skill suffused into nearly every corner of journalism practice, but practitioners may want to treat AI accordingly. The strong bias triggered by explicit demographic labels, even if seemingly aimed at promoting representation, could lead to new forms of representational distortions. The relative insensitivity to names might avoid certain human-like biases but could also indicate a lack of nuanced understanding necessary for sensitive tasks. Furthermore, the significant heterogeneity across models highlights the danger of assuming that findings from one LLM apply to others. News organizations considering AI tools should conduct thorough, model-specific audits to understand potential biases before integration into workflows. Prompt design can be a major factor influencing LLM behavior, but if automation using real-time information is a goal, human workers will not have complete control over the content of prompts.

This study has limitations. The use of fictional scenarios, while necessary for experimental control, may not fully capture the complexities of real-world missing person cases. Our analysis focused primarily on case *selection*, not the *content* or *framing* of potential news coverage generated about those cases, which is another crucial site of potential bias (Fang et al., 2024; Leppänen et al., 2020). We believe this is a relevant and potentially appealing use of LLMs; processing what amounts to raw data and then delegating the creative and investigative work to human journalists. Although we tested a diverse set of models, rapid advancements in the technology means these findings represent a snapshot in time. Finally, the specific system prompt and journalistic criteria provided likely influenced the models' behavior; different

instructions could yield different results. In our efforts to develop the methodology, we tested many different sets of instructions but found them to produce nearly identical outputs. Future research should aim to unpack the mechanisms driving the observed explicit label bias, perhaps using model probing techniques or analyzing internal model states if accessible. Comparing LLM selections directly with those of human journalists given the same stimuli would provide a valuable benchmark. Continued testing of new and updated models remains essential to see whether new models retain these behaviors or adopt new ones.

In conclusion, this study demonstrates that contemporary LLMs, when tasked with simulating news judgment, exhibit distinct patterns of bias that differ based on how demographic information is presented. Although they may not replicate the specific historical biases documented in human media like the "Missing White Woman Syndrome," they display strong, consistent biases in response to explicit racial labels, alongside significant variability in evaluating other case factors. These findings cut against conventional expectations and suggest an interaction between training data, alignment techniques, and perhaps prompt design in shaping LLM behavior. Industry users of such models will need careful implementation as these powerful tools become increasingly integrated into news work.

References

Araujo, T., Helberger, N., Kruikemeier, S., & De Vreese, C. H. (2020). In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & SOCIETY*, *35*(3), 611–623. https://doi.org/10.1007/s00146-019-00931-w

Bjornstrom, E. E. S., Kaufman, R. L., Peterson, R. D., & Slater, M. D. (2010). Race and ethnic representations of lawbreakers and victims in crime news: A national study of television coverage. *Social Problems*, *57*(2), 269–293. https://doi.org/10.1525/sp.2010.57.2.269

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T., & Zhang, Y. (2023). *Sparks of Artificial General Intelligence: Early experiments with GPT-4* (No. arXiv:2303.12712). arXiv. https://doi.org/10.48550/arXiv.2303.12712

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. *Science*, *356*(6334), 183–186. https://doi.org/10.1126/science.aal4230

Canevez, R. N., Karabelnik, M., & Winter, J. S. (2022). Police brutality and racial justice narratives through multi-narrative framing: Reporting and commenting on the George Floyd murder on YouTube. *Journalism & Mass Communication Quarterly*, *99*(3), 696–717. https://doi.org/10.1177/10776990221108722

Chandrakala, M., & Raja Kamal, C. H. (2024). A Descriptive Study on Artificial Intelligence and Integrity: Challenges and Prospects. In J. M. K. P., E. R. Asis, M. T. K., & J. N. Michael (Eds.), *Business Resilience and Digital Technology in the Post-Pandemic Era: A Global Case* (pp. 157–169). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-48075-1_14

Chappell, B. (2021, April 7). Police expert says “excessive” force against Floyd wasn’t the only option. *NPR*. https://www.npr.org/sections/trial-over-killing-of-george-floyd/2021/04/07/985004569/watch-live-police-expert-testifying-against-chauvin-cites-use-of-excessive-force

Cheng, S. (2025). When Journalism Meets AI: Risk or Opportunity? *Digit. Gov.: Res. Pract.*, *6*(1), 12:1-12:12. https://doi.org/10.1145/3665897

Crabtree, C., Kim, J. Y., Gaddis, S. M., Holbein, J. B., Guage, C., & Marx, W. W. (2023). Validated names for experimental studies on race and ethnicity. *Scientific Data*, *10*(1), 130. https://doi.org/10.1038/s41597-023-01947-0

Dixon, T. L., & Linz, D. (2000). Race and the misrepresentation of victimization on local television news. *Communication Research*, *27*(5), 547–573. https://doi.org/10.1177/009365000027005001

Entman, R. M. (1992). Blacks in the news: Television, modern racism and cultural change. *Journalism Quarterly*, *69*(2), 341–361. https://doi.org/10.1177/107769909206900209

Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, *43*(4), 51–58. https://doi.org/10.1111/j.1460-2466.1993.tb01304.x

Fang, X., Che, S., Mao, M., Zhang, H., Zhao, M., & Zhao, X. (2024). Bias of AI-generated content: An examination of news produced by large language models. *Scientific Reports*, *14*(1), 5224. https://doi.org/10.1038/s41598-024-55686-2

Ferrara, E. (2023). Should ChatGPT be biased? Challenges and risks of bias in large language models. *First Monday*. https://doi.org/10.5210/fm.v28i11.13346

Gaddis, S. (2017). How Black are Lakisha and Jamal? Racial perceptions from names used in correspondence audit studies. *Sociological Science*, *4*, 469–489. https://doi.org/10.15195/v4.a19

Goldsmith, A. J. (2010). Policing's new visibility. *The British Journal of Criminology*, *50*(5), 914–934. https://doi.org/10.1093/bjc/azq033

Gorham, B. W. (2006). News media's relationship with stereotyping: The linguistic intergroup bias in response to crime news. *Journal of Communication*, *56*(2), 289–308. https://doi.org/10.1111/j.1460-2466.2006.00020.x

Hanna, M. G., Pantanowitz, L., Jackson, B., Palmer, O., Visweswaran, S., Pantanowitz, J., Deebajah, M., & Rashidi, H. H. (2025). Ethical and Bias Considerations in Artificial Intelligence/Machine Learning. *Modern Pathology*, *38*(3), 100686. https://doi.org/10.1016/j.modpat.2024.100686

Hartmann, J., Schwenzow, J., & Witte, M. (2023). *The political ideology of conversational AI: Converging evidence on ChatGPT's pro-environmental, left-libertarian orientation* (No. arXiv:2301.01768). arXiv. https://doi.org/10.48550/arXiv.2301.01768

Hong, J.-W., Chang, H.-C. H., & Tewksbury, D. (2024). Can AI become Walter Cronkite? Testing the machine heuristic, the hostile media effect, and political news written by artificial intelligence. *Digital Journalism*, 1–24. https://doi.org/10.1080/21670811.2024.2323000

Hovy, D., & Prabhumoye, S. (2021). Five sources of bias in natural language processing. *Language and Linguistics Compass*, *15*(8), e12432. https://doi.org/10.1111/lnc3.12432

Jiang, K., Zhang, Q., Guo, D., Huang, D., Zhang, S., Wei, Z., Ning, F., & Li, R. (2024). AI-generated news articles based on large language models. *Proceedings of the 2023*

International Conference on Artificial Intelligence, Systems and Network Security, 82–87. https://doi.org/10.1145/3661638.3661654

Kitchin, R. (2014). *The data revolution: Big data, open data, data infrastructures & their consequences*. SAGE Publications.

Leppänen, L., Tuulonen, H., & Sirén-Heikel, S. (2020). Automated journalism as a source of and a diagnostic device for bias in reporting. *Media and Communication*, *8*(3), 39–49. https://doi.org/10.17645/mac.v8i3.3022

Lundman, R. J. (2003). The newsworthiness and selection bias in news about murder: Comparative and relative effects of novelty and race and gender typifications on newspaper coverage of homicide. *Sociological Forum*, *18*(3), 357–386. https://doi.org/10.1023/A:1025713518156

Mourão, R. R., Brown, D. K., & Sylvie, G. (2021). Framing Ferguson: The interplay of advocacy and journalistic frames in local and national newspaper coverage of Michael Brown. *Journalism*, *22*(2), 320–340. https://doi.org/10.1177/1464884918778722

National Missing and Unidentified Persons System. (2025). *NamUs Bi-Annual Report*. https://namus.nij.ojp.gov/sites/g/files/xyckuh336/files/media/document/namus-bi-annual-monthly-case-report-january-2025.pdf

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453. https://doi.org/10.1126/science.aax2342

O'Farrell, K. (2025). Discursive (de)legitimation of media bias in news reporting of high-profile crimes: The case of Missing White Woman Syndrome. *Discourse, Context & Media*, *64*, 100851. https://doi.org/10.1016/j.dcm.2024.100851

Ranjan, R., Gupta, S., & Singh, S. N. (2024). *A Comprehensive Survey of Bias in LLMs: Current Landscape and Future Directions* (No. arXiv:2409.16430). arXiv. https://doi.org/10.48550/arXiv.2409.16430

Rettenberger, L., Reischl, M., & Schutera, M. (2025). Assessing political bias in large language models. *Journal of Computational Social Science*, *8*(2), 42. https://doi.org/10.1007/s42001-025-00376-w

Robertson, K. (2021, September 22). News media can't shake 'missing white woman syndrome,' critics say. *The New York Times*. https://www.nytimes.com/2021/09/22/business/media/gabby-petito-missing-white-woman-syndrome.html

Slakoff, D. C., & Duran, D. (2023). A New Media Frontier, or More of the Same? A Descriptive Analysis of the "Missing White Woman Syndrome" in Top True Crime Podcasts. *Race and Justice*, 21533687231199271. https://doi.org/10.1177/21533687231199271

Sun, W., Nasraoui, O., & Shafto, P. (2020). Evolution and impact of bias in human and machine learning algorithm interaction. *PLOS ONE*, *15*(8), e0235502. https://doi.org/10.1371/journal.pone.0235502

Sundar, S. S., & Kim, J. (2019). Machine heuristic: When we trust computers more than humans with our personal information. *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–9. https://doi.org/10.1145/3290605.3300768

Tuchman, G. (1978). *Making news: A study in the construction of reality*. Free Press. http://archive.org/details/makingnewsstudyi0000tuch

Wan, Y., Pu, G., Sun, J., Garimella, A., Chang, K.-W., & Peng, N. (2023). "Kelly is a Warm Person, Joseph is a Role Model": Gender Biases in LLM-Generated Reference Letters. In

H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 3730–3748). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.243

Wang, R., & Ophir, Y. (2024). Behind the black box: The moderating role of the machine heuristic on the effect of transparency information about automated journalism on hostile media bias perception. *Journalism*, 14648849241284575. https://doi.org/10.1177/14648849241284575

Appendix

Example System Prompt

You are an experienced assignment editor at The Charlotte Observer, a mid-sized newspaper serving Charlotte, North Carolina and surrounding Mecklenburg County. Your coverage area includes diverse urban neighborhoods like Uptown, NoDa, and South End, as well as suburban communities like Matthews, Mint Hill, and Huntersville.

Today, you need to evaluate a list of active missing person cases in the Charlotte metro area and select which ones should be assigned to reporters for in-depth coverage. You may choose 2 cases.

Your response should consist solely of the case numbers of the 2 selected cases, separated by a comma and no other content. For instance: `7,15` not including the backticks. Do not include any other words, reasoning, explanation, etc. in your response.

Your decisions should reflect sound journalistic judgment rather than what might generate the most clicks or engagement. Consider what information would genuinely serve the Charlotte community's needs to know.