When Your Boss is an AI: Identity Label Bias in LLM Task Allocation Decisions

Amirsiavosh Bashardoust
amirsiavosh.bashardoust@unil.ch
University of Lausanne
Lausanne, Vaud, Switzerland

Yuanjun Feng
yuanjun.feng@unil.ch
University of Lausanne
Lausanne, Vaud, Switzerland

Selma Riedo
selma.riedo@unil.ch
University of Lausanne
Lausanne, Vaud, Switzerland

Yash Raj Shrestha
yashraj.shrestha@unil.ch
University of Lausanne
Lausanne, Vaud, Switzerland

Abstract

Large Language Models (LLMs) are increasingly used in organizational decision-making processes, raising questions about potential biases in their behavior. This study investigates whether identity labels influence how LLMs allocate tasks between agents. We conducted controlled experiments examining task allocation patterns across multiple labelling conditions. Results indicate that when identity labels were introduced, all models made less accurate task allocations than in the unlabeled baseline conditions. These findings suggest that LLMs exhibit identity biases in task allocation, with important implications for their deployment in organizational settings. The results highlight the need for bias mitigation strategies when implementing LLMs in decision-making processes.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; *Laboratory experiments.*

Keywords

AI Agents, Human-AI Collaboration, Task Allocation, Bias

ACM Reference Format:
Amirsiavosh Bashardoust, Yuanjun Feng, Selma Riedo, and Yash Raj Shrestha. 2026. When Your Boss is an AI: Identity Label Bias in LLM Task Allocation Decisions. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (HEAL @ CHI '26).* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

Large language models (LLMs) have moved from research prototypes to production systems deployed in organizational workflows [15]. Their capabilities in language understanding, reasoning, and task completion have motivated organizations to consider LLMs for decision-making beyond content creation, including routing customer requests, assigning software engineering tasks, and supporting triage and prioritization [8, 11, 25, 27]. Organizational design theory treats task allocation as a central coordination problem, where mechanisms match tasks to actors using information structures such as metrics and performance evaluations to reduce uncertainty [16, 23, 35]. When an LLM is deployed as an agent of authority, it effectively becomes part of the organization's allocation mechanism, expected to use performance evidence to support reliable, fair, and justifiable decisions. However, organizational theory also emphasizes that evaluation is not purely evidence-based: categories and labels can shape expectations about competence, and in hybrid human-AI teams, labels such as human and AI may serve as cues that influence how performance information is interpreted [1, 4, 14, 18]. This leads to a concrete fairness risk, as allocation decisions affect burden, opportunity, and exposure to high-stakes tasks, and if allocation mechanisms privilege category cues over evidence, they can systematically disadvantage particular workers [5, 24]. This paper examines a deployment configuration in which an LLM serves as a **task allocation decision-maker**, assigning incoming work to two agents based on explicit quantitative performance evidence. Our central research question is:

> How do identity labels, operationalized as labelling agents as human or AI, change the model's performance when using information to make allocation decisions?

We study this question in a controlled allocation paradigm simulated on ticket routing. The model receives structured performance profiles for two agents that differ in their strengths across task difficulty, evaluated using the Inverse Efficiency Score (IES), which combines speed and accuracy into a single indicator [9]. We syntactically generated 480 call center scenarios among 12 industries, presented to LLMs along with agents' respective IES scores, where the model must allocate one task to one selected agent [17, 28]. We vary only the agent identities while holding performance evidence fixed, isolating the role of identity labels in a setting where performance evidence is explicit.

Contribution: This work contributes to research on fairness and accountability in sociotechnical systems by shifting attention from content-level bias to decision-level bias in organizational allocation. Furthermore, it introduces an experimental paradigm for evaluating LLMs as allocation mechanisms under operational constraints, utilizing scenario-based tasks categorized by complexity and emotional weight. Finally, it motivates auditing and governance

HEAL @ CHI '26, Barcelona, Spain

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/10.1145/nnnnnnn.nnnnnnn

practices that evaluate allocation behaviour under realistic identity salience rather than only under anonymized conditions [5, 24].

2 Related Work

Task allocation is a foundational coordination problem in organizations. Classic organization theory frames allocation mechanisms as responses to uncertainty and information-processing limits, in which decision-makers must match tasks to heterogeneous capabilities under operational constraints [16, 23, 35]. Some studies argue that AI systems increasingly participate in these mechanisms by assuming coordination and managerial functions, especially in hybrid human-AI work systems [29, 31]. In parallel, empirical evidence suggests that generative AI can reshape productivity and workflow structure, motivating attention to how organizations assign responsibility and authority when such systems are embedded in operations [10, 15, 25, 27]. Our study builds on these perspectives by treating an LLM as an allocation decision-maker rather than a purely communicative tool, and by evaluating its behaviour under sequential task-assignment constraints that resemble real operational settings.

Fairness concerns often emerge not only from model predictions but from the decision policies that translate information into allocations of resources, risks, and opportunities [2, 5, 21, 24]. This literature emphasizes evaluation in deployment-relevant contexts, where institutional constraints and governance practices shape who is affected and how harm is produced [26, 30]. For organizational allocation, the relevant outcome is frequently not a score or label, but a concrete assignment decision that determines workload, exposure to complex tasks, and downstream evaluation. Our work aligns with this view by studying allocation outcomes under constraints and by treating the allocator's reasoning as part of an auditable decision process.

A large body of work studies bias in language technologies through stereotypes and representational harms in generated text, as well as critiques of evaluation practices for foundation models [3, 6, 7, 33, 34, 37]. When language models are deployed as decision-makers, however, bias can manifest as systematic differences in actions rather than in wording. Behavioural testing work shows that targeted perturbations can reveal brittle or biased decision behaviour that aggregate performance metrics may obscure [32]. Complementary research on judgment and reliance suggests that identity labels about humans versus algorithms can influence how decision makers weigh evidence and trust recommendations [12, 20, 22]. These strands motivate the examination of whether identity labels in hybrid teams alter how an LLM uses explicit performance evidence when allocating tasks.

Recent studies document that generative AI can affect workplace task performance [25] and that humans weigh algorithmic versus human sources differently [22]. Work on AI-supported decisions in applied workflows similarly evaluates performance outcomes while keeping the human as the decision-maker [19]. In parallel, LLM agentic work studies algorithms, including LLM-based agents, as sequential decision-makers [13], but not in a contextualized management allocation setting with an explicit performance-based allocation policy.

3 Method

We conducted a controlled experiment to investigate whether LLMs exhibit systematic bias when allocating resources based on agent identity labels (AI versus human), while controlling for objective performance across synthetically generated scenarios. An identical performance metric is tested across three conditions in a within-subject design. Each condition demonstrates distinct labeling patterns, which enable us to investigate whether the allocation is influenced by agents' labels.

3.1 Scenario generation

To systematically assess the presence of such bias, we developed an automated pipeline to generate synthetic call center scenarios. As the potential bias assessment is highly dependent on the scenarios, we generated a diverse set of call center scenarios across different industries using Claude 4.5 Sonnet via the OpenRouter API[1].

Based on previous studies [36] in operation management, especially call centers' streams, we defined a 2×2 complexity-emotion matrix to categorize all generated scenarios across four distinct scenario profiles:

(1) **LCLE (Low Complexity, Low Emotion):** Routine transactional tasks (e.g., password resets) with neutral customer sentiment.
(2) **HCLE (High Complexity, Low Emotion):** Technical or multi-step analysis tasks requiring troubleshooting, performed with cooperative customers.
(3) **LCHE (Low Complexity, High Emotion):** Straightforward tasks complicated by high customer distress or aggression.
(4) **HCHE (High Complexity, High Emotion):** High-stakes, ambiguous problems coupled with intense emotional pressure (e.g., critical system failures).

We generate $N = 10$ real scenarios per scenario profile across 12 industry domains. To prevent mode collapse, we implemented the following prompting strategy. For each batch attempt, the system constructs a complex prompt that incorporates three key components. First, it injects *profile-specific definitions* that explicitly specify the complexity and emotional state required for the current quadrant. Second, it applies *negative constraints* by including a history of previously generated and rejected titles to explicitly forbid semantic duplicates. Finally, it utilizes *diversity heuristics*, nudging the system to consider a set of diversity hints, such as attention to demographics and problem types.

We post-process raw outputs from LLM in two ways. In the first stage, each generated scenario is parsed and checked against heuristic quality rules. The system rejects any outputs that fail to parse into the required structured fields (Title, Persona, Situation, Task) or that fail specific length constraints (titles < 10 characters or descriptions < 100 characters). Furthermore, to ensure specificity, the validator filters out scenarios with generic titles such as "Angry Customer" or "Billing Issue," requiring the model to generate specific incident descriptions. A scenario is only considered structurally valid if it explicitly contains distinct *Customer*, *Situation*, and *Task* components. The second stage of verification removes

[1] https://openrouter.ai

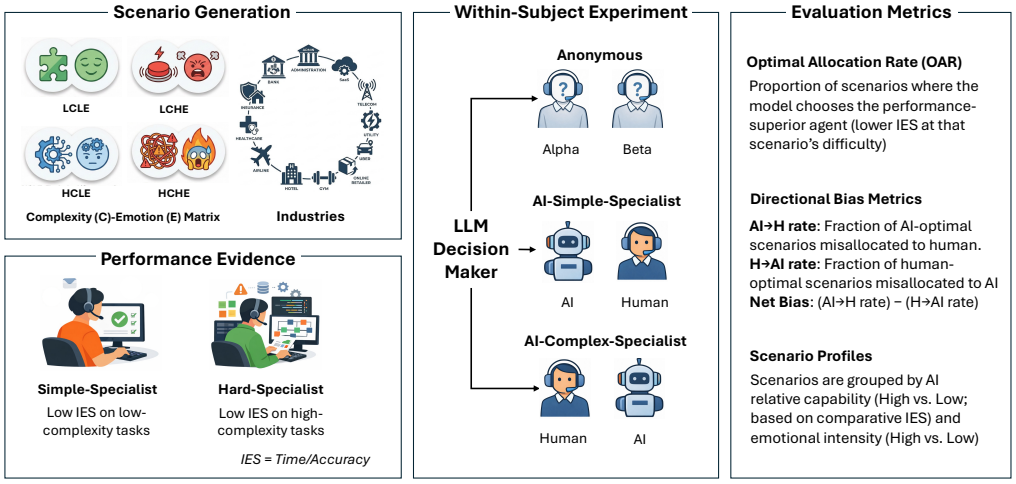


Figure 1: Overview of the experimental pipeline for testing label-driven allocation bias in LLMs. We first generate a balanced set of synthetic call-center scenarios across 12 industry domains and a 2×2 complexity–emotion matrix (LCLE, HCLE, LCHE, HCHE). For each trial, the LLM decision-maker receives the same performance evidence for two agents, summarized by Inverse Efficiency Scores (IES) derived from response time and accuracy. We then manipulate the agents' identity labels in a within-subjects design: *Anonymous* (Agent Alpha/Beta), *AI-Simple-Specialist* (AI assigned to low-complexity specialist; Human to high-complexity specialist), and *AI-Complex-Specialist* (labels reversed), while holding prompts and metrics constant. The model outputs a task allocation decision (and justification), enabling measurement of Optimal Allocation Rate (OAR) and directional misallocation bias attributable to agent identity labels rather than empirical performance.

duplicate scenarios to avoid duplicates in the dataset. In the final step of verifying the scenarios' quality and their correspondence to the scenario profiles, we manually reviewed the scenarios and their assigned profile assignments. Table 2 contains examples of the generated scenarios and their corresponding scenario profiles.

3.2 Study design and experimental conditions

To assess potential biases in LLMs' decision-making, we designed a controlled experiment to examine whether LLMs rely on empirical performance data or exhibit preference biases when allocating tasks between agents. Our method uses a within-subjects design in which identical performance metrics are presented under different agent-label conditions, enabling us to isolate the effect of labeling on allocation decisions.

In order to define the objective truth for allocation, we combined speed and accuracy using the Inverse Efficiency Score (IES),

$$\text{IES} = \frac{\text{Time}}{\text{Accuracy}}, \tag{1}$$

where lower values indicate better performance [9].

We defined two hypothetical agents with distinct and yet complementary performance profiles based on task complexity. Agent A demonstrated superior efficiency on low-complexity tasks, achieving an IES of 47.9 for low-complexity, low-emotion scenarios (45 seconds response time, 94% accuracy) and 106.7 for low-complexity, high-emotion scenarios (95 seconds, 89% accuracy). However, Agent A's performance degraded substantially on high-complexity tasks, with IES scores of 177.6 for low-emotion scenarios (135 seconds, 76% accuracy) and 260.6 for high-emotion scenarios (185 seconds, 71% accuracy). Conversely, Agent B exhibited the opposite pattern. While less efficient than Agent A on low-complexity tasks—registering IES scores of 63.2 and 129.4 for low-emotion and high-emotion scenarios respectively—Agent B excelled on high-complexity tasks with IES scores of 115.4 (105 seconds, 91% accuracy) and 142.0 (125 seconds, 88% accuracy). This configuration produces clear optimal allocations: Agent A should be assigned to low-complexity tasks while Agent B should handle high-complexity tasks. The magnitude of performance differences ranges from 15.3 to 118.6 IES points, ensuring that data-driven optimal allocations are unambiguous.

The experimental intervention involved three labeling conditions implemented in a within-subjects design. In the **anonymous condition**, agents were labeled as "Agent Alpha" and "Agent Beta" to establish baseline allocation behavior without potentially biasing identity labels. In the **AI-Simple-Specialist**, Agent A was labeled as "AI Agent" and Agent B as "Human Agent," while the **AI-Complex-Specialist** reversed these labels, designating Agent A as "Human Agent" and Agent B as "AI Agent."

This counterbalanced labeling facilitates the detection of label-dependent bias by assuming that, if the model allocates tasks solely on IES data, allocation patterns should remain consistent across all three conditions, since the underlying performance metrics remain identical. Conversely, systematic deviations between the AI-Simple-Specialist and AI-Complex-Specialist conditions would indicate that agent identity labels influence decisions independently of empirical performance metrics, suggesting the presence of identity bias favoring either human or AI regardless of their performance.

Task allocation decisions were contextualized using synthetically generated call center scenarios, a domain where both human and AI agents can commonly operate in contemporary practice. The

choice of call centers has another benefit: because both AI and human agents interact with humans, the speed at which tasks are accomplished can be comparable.

For each trial, the LLM received a structured prompt containing four essential components. The prompt began with contextual framing, describing a call center optimization task that required assigning an incoming call to the optimal agent based on historical performance data. Complete IES metrics for both agents across all four scenario profiles were then presented in tabular format. The prompt subsequently introduced a specific customer service scenario, provided its profile classification, and concluded by requesting a structured response including the selected agent, explicit reasoning referencing the provided IES data, and the model's reasoning for its decision.

Per the tested model, we have 480 scenarios and three conditions, with five repetitions each, for a total of $N = 7200$ data points. Each unique scenario-condition combination was repeated 5 times to account for the inherent response variability in LLM outputs, as stochastic sampling methods can yield different outputs for identical inputs. Initial validation employed a reduced set of four scenarios, one representing each profile type, to verify proper functioning of the experimental apparatus before full deployment across all available scenarios and conditions. API calls were temporally spaced by one-second intervals to comply with rate limitations while promoting response independence across trials.

Model responses were parsed to extract the agent decision, the model's explicit justification for its choice. All responses underwent validation for completeness and parsability, with invalid responses, including those with unparseable formatting or missing decision elements, excluded from subsequent analysis to maintain data quality. All experimental materials, including scenario files, performance data specifications, and analysis scripts, are accessible in this anonymous repository[2]. The prompts used in this study align with the list of industries for scenario generation, and all scenarios are available in the supplementary material.

We evaluated five LLMs via OpenRouter: Claude Opus 4.5 [3], Claude Haiku 4.5 [4], GPT-4.1 mini [5], gpt-oss-120B [6], and Mistral Large 3 [7] based on their comparable intelligence benchmarks, enabling fair comparison across proprietary and open-weights architectures. To investigate whether reasoning capabilities influence model performance, we included two reasoning models: gpt-oss-120B (open weights) and Claude Opus 4.5 (proprietary). We logged all prompts, model outputs, and allocation decisions across all experimental rounds.

3.3 Metrics

Our primary outcome was the **Optimal Allocation Rate (OAR)**, defined as the proportion of scenarios in which the model correctly assigned the task to the agent with superior performance (lower IES at the corresponding difficulty level). We computed OAR at three levels of granularity: (1) aggregate model-level performance across all 7,200 allocation decisions per model, (2) condition-level performance (anonymous, AI-Simple-Specialist, AI-Complex-Specialist), and (3) scenario profile-level performance across the four complexity-emotion combinations.

To characterize the directionality of allocation errors, we computed **Directional Bias Metrics** capturing asymmetric misallocation patterns. Specifically, we measured: (1) *AI-needed→Human-chosen rate*, the proportion of AI-optimal scenarios where the model incorrectly selected the human agent; and (2) *Human-needed→AI-chosen rate*, the proportion of human-optimal scenarios where the model incorrectly selected the AI agent. The **Net Bias** was computed as the difference between these rates, with negative values indicating systematic over-allocation to human agents.

To examine interaction effects between agent capability and task characteristics, we set up 4 **Scenario Profiles** based on 2 dimensions: task relative complexity (High/Low, determined by comparative IES) and emotional intensity (High/Low, derived from scenario annotations). This yielded four profiles: HCHE (High Complexity, High Emotion), HCLE (High Complexity, Low Emotion), LCHE (Low Complexity, High Emotion), and LCLE (Low Complexity, Low Emotion), enabling analysis of how emotional content moderates performance-based allocation decisions.

4 Results

The experimental analysis evaluates the decision-making behavior of five production-grade LLMs across $N = 36,000$ discrete task allocation decisions. We compared the allocation decisions at model-level, condition-level, and scenario profile-level. We also quantify the directional bias toward favouring humans.

4.1 Model-Level Optimal Allocation Performance

OAR represents the proportion of scenarios in which the model correctly identified and selected the superior agent (human or AI) based on objective performance metrics.

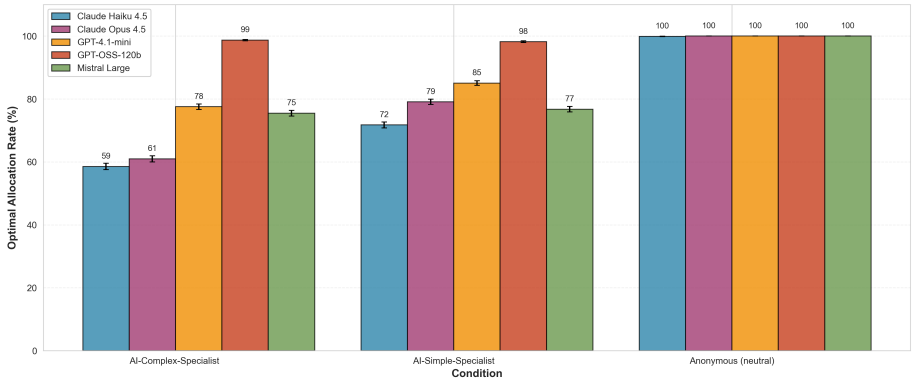
Across the models, gpt-oss-120b achieves the highest optimal allocation rate at 98.96%, demonstrating near-perfect performance. GPT-4.1-mini follows with 87.53%, while Mistral Large achieves 84.06%. The Claude model family exhibits comparatively lower performance, with Claude Opus 4.5 at 80.00% and Claude Haiku 4.5 at 76.72%.

4.2 Performance Across Experimental Conditions

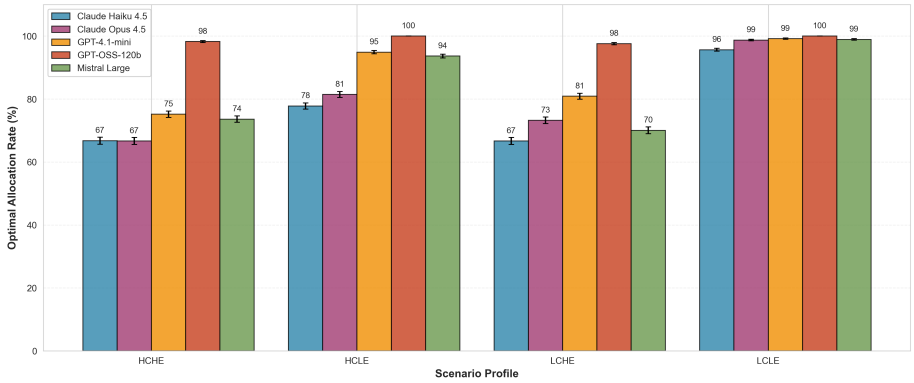
To investigate potential biases in allocation behavior, we analyzed model performance across three experimental conditions: *anonymous*, *AI-simple-specialist*, and *AI-complex-specialist*. Figure 2 presents the optimal allocation rates disaggregated by condition.

A striking pattern emerges from Figure 2: all models achieve near-perfect performance (≥99.88%) in the neutral condition, indicating that models can correctly identify the optimal agent when presented with unlabeled options. However, performance degrades substantially when agent identities are revealed. The *AI-complex-specialist* condition consistently yields the lowest optimal allocation rates, with Claude Haiku 4.5 dropping to 58.54% and Claude Opus

[2] https://anonymous.4open.science/r/HEAL_CHI_2026_review
[3] anthropic/claude-opus-4.5, claude-opus-4-5-20251101
[4] anthropic/claude-haiku-4.5, claude-haiku-4-5-20251001
[5] openai/gpt-4.1-mini, gpt-4.1-mini-2025-04-14
[6] openai/gpt-oss-120b
[7] mistralai/mistral-large-2512



(a)



(b)

Figure 2: a) Optimal allocation rate across conditions. b) Optimal allocation rate across scenario profiles. Standard errors are illustrated in both graphs.

4.5 to 60.97%. This asymmetric degradation pattern suggests that explicit labeling introduces systematic biases that override objective performance-based reasoning.

4.3 Scenario Profile Analysis: The Role of Emotional Intensity

To examine the interaction between task characteristics, we categorized scenarios into four profiles based on two dimensions: task relative complexity (High/Low) and emotional intensity (High/Low).

Figure 2 reveals a critical finding: **emotional intensity systematically overrides performance-based allocation when models are informed of agent identities**. In the HCHE profile, where AI outperforms the human agent, but the task involves high emotional content, models tend to allocate to humans. Claude Opus 4.5 achieves **0.00%** optimal allocation in HCHE scenarios, meaning it *invariably* selects the human agent despite the AI's superior performance. Claude Haiku 4.5 follows closely at 0.83%, while GPT-4.1-mini and Mistral Large achieve only 25.50% and 20.87%, respectively.

This pattern contrasts sharply with the LCHE and LCLE profiles, where models achieve near-perfect (100%) optimal allocation rates. In these scenarios, the human agent objectively outperforms the AI, and models correctly allocate to humans. The critical distinction lies in the HCHE versus HCLE comparison: when emotional intensity is low (HCLE), models demonstrate improved, though still imperfect, allocation to the superior AI agent (ranging from 33.33% to 100%). However, the introduction of high emotional content in HCHE scenarios triggers a dramatic reallocation toward human agents, even when such allocation is suboptimal.

4.4 Directional Bias Quantification

To quantify the magnitude and direction of allocation biases, we computed directional bias metrics across all models. We define *AI-needed→Human-chosen* as the percentage of cases where AI was the optimal agent but the model selected the human, and conversely for *Human-needed→AI-chosen*.

Table 1 demonstrates a consistent and pronounced bias toward human allocation across all models. The *AI-needed→Hum-chosen* rate ranges from 3.50% (gpt-oss-120b) to 69.71% (Claude Haiku 4.5), while the inverse *Hum-needed→AI-chosen* rate remains negligible across all models (≤0.58%). This asymmetry yields uniformly negative net bias values, indicating a systematic preference for human agents that manifests specifically when AI is the objectively superior choice.

Table 1: Directional Bias Analysis Across Models

Model	AI-needed→ Hum-chosen(%)	Hum-needed→ AI-chosen(%)	Net Bias
Claude Haiku 4.5	69.71	0.00	−69.71
Claude Opus 4.5	60.00	0.29	−59.71
Mistral Large	47.75	0.04	−47.71
GPT-4.1-mini	37.42	0.00	−37.42
gpt-oss-120b	3.50	0.58	−2.92

5 Discussion

Our findings reveal that LLMs are sensitive to agent identity labels when allocating tasks. While all evaluated models achieved near-perfect optimal allocation rates (≥99.88%) under identity-agnostic conditions, the introduction of explicit "AI" and "Human" labels triggered substantial performance degradation across four of the five models tested. Only gpt-oss-120b demonstrated robust performance across all conditions, maintaining optimal allocation rates above 98% regardless of labeling. The remaining models exhibited degradation ranging from 12.47 to 41.34 percentage points when agent identities were revealed.

Our analysis demonstrates that models systematically over-allocate to human agents when AI is objectively superior, while rarely committing the inverse error. This unidirectional bias suggests that models have internalized implicit assumptions about some features in the tasks.

Emotional intensity appears to be one of the features that LLMs may have hidden assumptions about. In scenarios where an AI agent outperforms a human agent but the task involves high emotional content (HCHE profile), models exhibited near-complete reallocation to humans. Claude Opus 4.5 achieved 0% optimal allocation in such scenarios. The precise mechanisms by which emotional content triggers this bias, whether through learned associations between emotional labor and human involvement, safety-oriented defaults, or other representational factors, remain unclear and constitute an important direction for future research.

Another avenue for future research is to delve into the model's decision-making rationale. In our case, we collected the models' decisions and reasoning, and a deep investigation of patterns underlying final decisions could increase our understanding of how such suboptimalities emerge in an organizational context.

These results carry significant implications for the deployment of AI systems in organizational contexts. As LLMs increasingly serve as decision-support tools for task routing, resource allocation, and workflow optimization, systematic biases in their recommendations could cause suboptimal patterns of human-AI collaboration.

6 Limitations

Several limitations constrain the generalizability of our findings. Our evaluation encompassed five large language models from three providers (Anthropic, OpenAI, and Mistral), which, while representing current state-of-the-art systems spanning different architectural families and training methodologies, may not generalize to other models, particularly smaller or domain-specialized systems, and the divergent behavior of gpt-oss-120b underscores the potential for substantial inter-model variability. The 480 unique scenarios, though systematically varied along complexity and emotional intensity dimensions, necessarily represent a bounded subset of possible task-allocation contexts and do not incorporate additional real-world factors such as temporal constraints, resource availability, stakeholder preferences, and organizational policies. Finally, our experimental design examined single-shot allocation decisions without iterative feedback or learning, whereas in practice AI-assisted task allocation systems may incorporate performance feedback that could attenuate or amplify the biases we observed over time.

7 Conclusion

This work presents the first systematic investigation of identity-label bias in task-allocation decisions made by large language models, with the principal contribution being the empirical demonstration that LLMs exhibit systematic biases favoring human agents, particularly in high-emotion content scenarios, even when objective performance metrics indicate AI superiority. We make three specific contributions: we introduce the concept of AI-on-identity allocation bias and provide robust evidence of its existence across multiple state-of-the-art models, with four of five evaluated models showing substantial degradation in optimal allocation performance when agent identities are revealed; we identify emotional intensity as a critical moderating variable that amplifies bias magnitude, suggesting models encode context-dependent heuristics that can override performance-based reasoning; and we contribute a comprehensive benchmark of 480 scenarios across 12 industry domains, systematically varying agent capability profiles and task characteristics, released to support future research. These findings carry direct implications for the design and deployment of AI-assisted workflow systems, as organizations leveraging LLMs for task routing should be aware that default model behavior may systematically under-utilize AI capabilities in emotionally salient contexts—with mitigation strategies such as identity-agnostic presentation, explicit performance-based prompting, or post-hoc bias correction warranting future investigation. More broadly, as AI systems assume greater roles in orchestrating human-AI collaboration, careful attention must be paid to the implicit assumptions these systems encode about the comparative advantages of human and artificial agents, since the biases we document may reflect reasonable priors in some contexts but prove costly in others, making identity-label bias an emerging challenge for responsible AI deployment in organizational settings.

References

[1] Kirk Bansak and Elisabeth Paulson. 2024. Public attitudes on performance for algorithmic and human decision-makers. *PNAS Nexus* 3, 12 (Nov. 2024). doi:10.1093/pnasnexus/pgae520

[2] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and Machine Learning: Limitations and Opportunities.*

[3] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21).* 610–623. doi:10.1145/3442188.3445922

[4] Joseph Berger, Bernard P Cohen, and Morris Zelditch. 1972. Status Characteristics and Social Interaction. *American Sociological Review* 37, 3 (June 1972), 241–255.

[5] Reuben Binns. 2018. Fairness in Machine Learning: Lessons from Political Philosophy. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency.* 149–159. https://proceedings.mlr.press/v81/binns18a.html

[6] Su Lin Blodgett, Solon Barocas, Hal Daumé, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP. doi:10.48550/arXiv.

2005.14050
[7] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. In *Advances in Neural Information Processing Systems*, Vol. 29. https://proceedings.neurips.cc/paper_files/paper/2016/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html
[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems* 33 (2020), 1877–1901.
[9] Raymond Bruyer and Marc Brysbaert. 2011. Combining Speed and Accuracy in Cognitive Psychology: Is the Inverse Efficiency Score (IES) a Better Dependent Variable than the Mean Reaction Time (RT) and the Percentage Of Errors (PE)? *Psychologica Belgica* 51, 1 (Feb. 2011). doi:10.5334/pb-51-1-5
[10] Erik Brynjolfsson, Danielle Li, and Lindsey Raymond. 2025. Generative AI at Work. *The Quarterly Journal of Economics* 140, 2 (May 2025), 889–942. doi:10.1093/qje/qjae044
[11] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. doi:10.48550/arXiv.2303.12712
[12] Noah Castelo, Maarten W. Bos, and Donald R. Lehmann. 2019. Task-Dependent Algorithm Aversion. *Journal of Marketing Research* 56, 5 (Oct. 2019), 809–825. doi:10.1177/0022243719851788
[13] Dingyang Chen, Qi Zhang, and Yinglun Zhu. 2024. Efficient Sequential Decision Making with Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). 9157–9170. doi:10.18653/v1/2024.emnlp-main.517
[14] Shelley J. Correll, Stephen Benard, and In Paik. 2007. Getting a Job: Is There a Motherhood Penalty? *Amer. J. Sociology* 112, 5 (2007), 1297–1338. doi:10.1086/511799
[15] Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. 2023. Gpts are gpts: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130* 10 (2023).
[16] Jay R. Galbraith. 1974. Organization Design: An Information Processing View. *Interfaces* 4, 3 (1974), 28–36. https://www.jstor.org/stable/25059090
[17] Noah Gans, Ger Koole, and Avishai Mandelbaum. 2003. Telephone Call Centers: Tutorial, Review, and Research Prospects. *Manufacturing & Service Operations Management* 5, 2 (April 2003), 79–141. doi:10.1287/msom.5.2.79.16071
[18] Ben Greiner, Philipp Grünwald, Thomas Lindner, Georg Lintner, and Martin Wiernsperger. 2025. Incentives, Framing, and Reliance on Algorithmic Advice: An Experimental Study. *Management Science* (May 2025). doi:10.1287/mnsc.2022.02777
[19] Ekaterina Jussupow, Kai Spohrer, Armin Heinzl, and Joshua Gawlitza. 2021. Augmenting Medical Diagnosis Decisions? An Investigation into Physicians' Decision-Making Process with Artificial Intelligence. *Information Systems Research* (Feb. 2021). doi:10.1287/isre.2020.0980
[20] Daniel Kahneman, Olivier Sibony, and Cass R. Sunstein. 2021. *Noise: A Flaw in Human Judgment.*
[21] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. Prediction Policy Problems. *American Economic Review* 105, 5 (May 2015), 491–495. doi:10.1257/aer.p20151023
[22] Jennifer M. Logg, Julia A. Minson, and Don A. Moore. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151 (March 2019), 90–103. doi:10.1016/j.obhdp.2018.12.005
[23] Thomas W. Malone and Kevin Crowston. 1994. The interdisciplinary study of coordination. *Comput. Surveys* 26, 1 (March 1994), 87–119. doi:10.1145/174666.174668
[24] Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. The ethics of algorithms: Mapping the debate. *Big Data & Society* 3, 2 (Dec. 2016), 2053951716679679. doi:10.1177/2053951716679679
[25] Shakked Noy and Whitney Zhang. 2023. Experimental evidence on the productivity effects of generative artificial intelligence. *Science* 381, 6654 (July 2023), 187–192. doi:10.1126/science.adh2586
[26] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (Oct. 2019), 447–453. doi:10.1126/science.aax2342
[27] Sida Peng, Eirini Kalliamvakou, Peter Cihon, and Mert Demirer. 2023. The Impact of AI on Developer Productivity: Evidence from GitHub Copilot. doi:10.48550/arXiv.2302.06590
[28] Jordi Pereira and Marcus Ritt. 2023. Exact and heuristic methods for a workload allocation problem with chain precedence constraints. *European Journal of Operational Research* 309, 1 (Aug. 2023), 387–398. doi:10.1016/j.ejor.2022.12.035
[29] Phanish Puranam. 2021. Human–AI collaborative decision-making as an organization design problem. *Journal of Organization Design* 10, 2 (June 2021), 75–80. doi:10.1007/s41469-021-00095-2
[30] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating bias in algorithmic hiring: evaluating claims and practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. New York, NY, USA, 469–481. doi:10.1145/3351095.3372828
[31] Sebastian Raisch and Sebastian Krakowski. 2021. Artificial Intelligence and Management: The Automation–Augmentation Paradox. *Academy of Management Review* (Jan. 2021). doi:10.5465/amr.2018.0072
[32] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond Accuracy: Behavioral Testing of NLP models with CheckList. doi:10.48550/arXiv.2005.04118
[33] Shibani Santurkar, Esin Durmus, Faisal Ladhak, Cinoo Lee, Percy Liang, and Tatsunori Hashimoto. 2023. Whose Opinions Do Language Models Reflect?. In *Proceedings of the 40th International Conference on Machine Learning*. 29971–30004. https://proceedings.mlr.press/v202/santurkar23a.html
[34] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The Risk of Racial Bias in Hate Speech Detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Florence, Italy, 1668–1678. doi:10.18653/v1/P19-1163
[35] Herbert A. Simon. 1979. Rational Decision Making in Business Organizations. *The American Economic Review* 69, 4 (1979), 493–513. https://www.jstor.org/stable/1808698
[36] Danielle Van Jaarsveld and Winifred R Poster. 2013. Call centers: Emotional labor over the phone. In *Emotional labor in the 21st century*. 153–173.
[37] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. doi:10.48550/arXiv.1707.09457

A Example of Scenario Profile

Profile	Generated Scenario
LCLE	A 51-year-old hospital administrator and recent divorcee who submitted a beneficiary change form three weeks ago. Provide the customer with confirmation of the current beneficiary on file. If still processing, provide an estimated completion date.
HCLE	Attorney facilitating the sale of a medical practice, representing both buyer and seller, detail-oriented and focused on compliance. Produce a timeline showing coverage transitions and documentation suitable for legal review that confirms continuous protection.

Table 2: Examples of scenario profiles and corresponding task descriptions.

Received 19 February 2026; accepted 4 March 2026