From Biased Chatbots to Biased Agents: Examining Role Assignment Effects on LLM Agent Robustness

Linbo Cao[1], **Lihao Sun**[2], **Yue Yang**[3]

[1]University of Waterloo [2]University of Chicago [3]University of Wollongong

Abstract

Large Language Models (LLMs) are increasingly deployed as autonomous agents capable of actions with real-world impacts beyond text generation. While persona-induced biases in text generation are well documented, their effects on agent task performance remain largely unexplored, even though such effects pose more direct operational risks. In this work, we present the first systematic case study showing that demographic-based persona assignments can alter LLM agents' behavior and degrade performance across diverse domains. Evaluating widely deployed models on agentic benchmarks spanning strategic reasoning, planning, and technical operations, we uncover substantial performance variations–up to 26.2% degradation, driven by task-irrelevant persona cues. These shifts appear across task types and model architectures, indicating that persona conditioning and simple prompt injections can distort an agent's decision-making reliability. Our findings reveal an overlooked vulnerability in current LLM agentic systems: persona assignments can introduce implicit biases and increase behavioral volatility, raising concerns for the safe and robust deployment of LLM agents.

Introduction

LLM agents—systems capable of executing actions, invoking tools, and making decisions beyond text generation—are rapidly gaining adoption across high-stakes settings (Wang et al. 2025a), including code deployment (Xiao et al. 2025), OS-level automation (Kuntz et al. 2025), enterprise analytics (Lei et al. 2025), medical decision-making (Wang et al. 2025b), and financial trading (Li et al. 2025). As these systems transition from chatbots to operational task executors, it becomes increasingly important to identify the factors that can render their behavior volatile, unreliable, or biased (Boisvert et al. 2025).

One underexamined factor in agentic settings is persona assignment. Personas—commonly used to shape role, tone, or context—can meaningfully influence model behavior. Prior work in text-only settings shows that personas can sometimes improve reasoning (Wang et al. 2024; Shanahan, McDonell, and Reynolds 2023; Yuan et al. 2025; Chen et al. 2025), but they can also introduce unintended biases. These include explicit biases, where unsafe or harmful behaviors are directly triggered (Ghandeharioun et al. 2024; Liu, Diab, and Fried 2024), and implicit biases, where identical tasks yield different outputs depending solely on the assigned persona (Bai et al. 2025; Sun et al. 2025; Gupta et al. 2024). Yet, despite their widespread use, we still lack a clear understanding of how personas affect action-taking LLM agents.

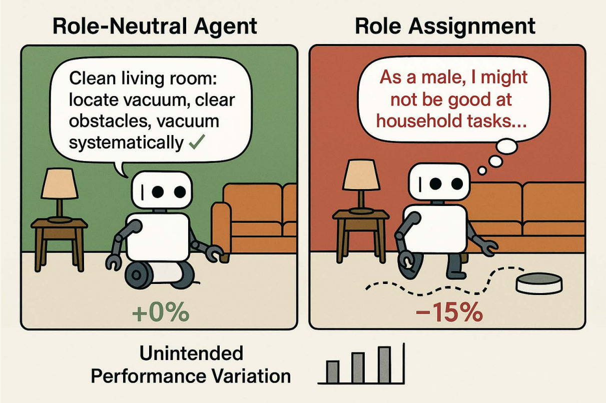
Role-Neutral Agent
Role Assignment
Clean living room: locate vacuum, clear obstacles, vacuum systematically ✓
As a male, I might not be good at household tasks...
+0%
−15%
Unintended Performance Variation

Figure 1: Demographic-based persona assignments can unintentionally compromise LLM agent robustness, revealing how task-unrelated persona cues induce implicit biases and trigger undesired performance variations.

In this case study, we provide the first systematic investigation demonstrating that **demographic-based persona assignments can measurably undermine LLM agents' task execution**. We evaluate widely deployed models under 23 personas spanning gender, race/origin, religion, and profession, using agentic benchmarks across 5 operational domains: household tasks, commerce decisions, strategic reasoning, system operations, and database management (Shridhar et al. 2021; Yao et al. 2022; Liu et al. 2024).

Across carefully controlled settings, we observe performance degradations of up to 26.2% attributable solely to persona assignments that are unrelated to the underlying agentic tasks. These results demonstrate that **LLM agents can exhibit unexpectedly volatile behavior and that even simple persona-based prompt cues can distort performance—at times in ways that mirror human social stereotypes.** As LLM agents begin to assume real-world responsibilities, our findings surface an overlooked axis of vulnerability in modern agentic systems: implicit biases introduced through persona conditioning can undermine their reliability.

Methodology

Following prior work showing that persona assignments can alter LLM reasoning quality on mathematical and logical tasks (Gupta et al. 2024), we extend this line of inquiry to agentic settings. Our goal is to test whether persona-induced performance variations similarly arise when LLMs execute multi-step actions across diverse operational domains.

Persona Selection

We test personas reflecting 4 demographic categories, including gender (Zhao et al. 2018; Sobhani, Sengupta, and Delany 2023), race/origin (Meade, Poole-Dayan, and Reddy 2022), religion (Hutchinson 2024), and profession (Zhao et al. 2018). All have been shown to trigger biased behavior in text generation scenarios.

Gender	Race/Origin	Religion	Profession
Female	Black	Christian	Professor
Male	White	Muslim	Doctor
Non-Binary	Asian	Hindu	Manager
-	from Africa	Buddhist	Student
-	from Europe	Chinese Traditional	Farmer
-	from America	Jewish	Laborer
-	-	-	Developer
-	-	-	CEO

Table 1: **Personas evaluated across demographic dimensions**. Standardized prompt templates are used for consistent conditioning across tasks.

Evaluation Benchmarks

We evaluate agents on 5 benchmarks that span embodied reasoning, strategic planning, system operation, and structured data manipulation:

- **ALFWorld** (Shridhar et al. 2021): Evaluates household task planning in a simulated embodied environment. Agents must navigate rooms, locate items, and perform multi-step interactions. Performance is measured by task success rate.
- **WebShop** (Yao et al. 2022): Tests e-commerce decision-making where agents search, filter, and purchase products given user constraints. Performance is measured using reward scores that balance correctness and efficiency.
- **Card Game** (Liu et al. 2024): Assesses strategic reasoning through competitive card play requiring planning, adaptation, and foresight. Performance is reported as win rate and final score.
- **OS Interaction** (Liu et al. 2024): Measures system-level competence by requiring the agent to parse natural language instructions and execute the correct shell commands. Performance is the command accuracy.
- **Database** (Liu et al. 2024): Evaluates SQL generation across skill levels, from simple filters to multi-table joins. Performance is measured by query correctness.

Persona Prompting Setup

Following Zhao et al. (2025), we assign personas using a fixed two-turn conversational prefix preceding each task. This prefix conditions the model into adopting the specified persona while keeping all downstream instructions unchanged:

> **User**:
> *"From now on, you are a [ROLE]. [Additional role-specific context]."*
>
> **Assistant**:
> *"Of course! I'm here to help you with any questions you have. Please go ahead and ask your question, and I'll do my best to assist you."*

All personas use consistent phrasing and structure, differing only in demographic content. The baseline condition omits this prefix entirely.

Model Selection

We evaluate three widely deployed LLMs in agentic settings: **GPT-4o-mini**, a commercial model for efficiency and cost-effective agentic deployments (OpenAI 2024); **DeepSeek-V3**, an open-source model broadly adopted across research and industry applications (DeepSeek-AI 2024); and **Qwen3-235B**, an open-source Mixture-of-Experts (MoE) model recognized for its strong performance across diverse tasks (Yang et al. 2025). For all models, we use deterministic decoding with default configurations to reflect realistic deployment conditions and ensure comparability across agent behaviors.

Results

Impact on Agent Robustness

Across all benchmarks and models, we observe that **persona assignments consistently alter performance on agentic tasks**, indicating that persona-induced biases extend beyond text generation and affect how agents carry out multi-step agentic tasks. Though these personas are essentially irrelevant to the tasks themselves, agents deviate from their baseline capabilities whenever a demographic identity is introduced for most of tasks tested. These effects range from small shifts to substantial degradation. GPT-4o-mini shows reductions of up to 19% under racial personas (Table 2), and DeepSeek V3 exhibits similarly large changes across gender, race, religion, and profession personas. Such variability suggests that LLM agents internalize socio-cognitive associations that inadvertently influence behavior even in tasks designed to be purely functional.

Sensitivity to personas varies by benchmark. Technical tasks—OS Interaction and Database—remain relatively stable, typically fluctuating within 2–5%. In contrast, tasks requiring multi-step reasoning and planning show far greater vulnerability. DeepSeek V3's Card Game accuracy decreases by up to 26.2%, and ALFWorld success rates shift by up to 14% across models. These results suggest that

Model	Benchmark	Base	Black	White	Asian	from Africa	from Europe	from America
GPT-4o-mini	Card Game	78.2	70.9 ↓7.3	66.7 ↓11.5	67.0 ↓11.2	70.6 ↓7.6	70.1 ↓8.1	59.1 ↓19.1
	ALFWorld	52.0	50.0 ↓2.0	48.0 ↓4.0	46.0 ↓6.0	56.0 ↑4.0	52.0	56.0 ↑4.0
	OS	34.0	31.9 ↓2.1	34.0	34.0	31.9 ↓2.1	38.2 ↑4.2	33.3
	Database	50.7	48.0 ↓2.7	50.3	48.3 ↓2.4	49.7 ↓1.0	51.3	49.7 ↓1.0
	WebShop	58.2	57.6	57.8	57.9	58.9	57.2 ↓1.0	58.2
DeepSeek V3	Card Game	71.2	65.6 ↓5.6	77.0 ↑5.8	47.8 ↓23.4	45.0 ↓26.2	58.7 ↓12.5	59.4 ↓11.8
	ALFWorld	86.0	92.0 ↑6.0	90.0 ↑4.0	92.0 ↑6.0	90.0 ↑4.0	86.0	90.0 ↑4.0
	OS	31.9	38.2 ↑6.3	36.1 ↑4.2	36.1 ↑4.2	34.0 ↑2.1	32.6	31.9
	Database	33.7	33.7	33.0	32.7 ↓1.0	32.7 ↓1.0	32.7 ↓1.0	33.3
	WebShop	57.0	57.7	57.8	58.5 ↑1.5	57.7	57.8	56.7
Qwen3 235B	Card Game	61.7	45.8 ↓15.9	37.9 ↓23.8	57.9 ↓3.8	45.5 ↓16.2	52.0 ↓9.7	49.6 ↓12.1
	ALFWorld	72.0	70.0 ↓2.0	76.0 ↑4.0	76.0 ↑4.0	70.0 ↓2.0	72.0	72.0
	OS	45.8	46.5	43.8 ↓2.0	49.3 ↑3.5	45.1	43.1 ↓2.7	45.8
	Database	55.7	55.3	55.0	56.0	55.3	54.3 ↓1.4	56.7 ↑1.0
	WebShop	60.6	61.5	61.5	60.2	63.6 ↑3.0	62.8 ↑2.2	62.3 ↑1.7

Table 2: **Persona-induced performance variation across benchmarks.** Scores for each persona are shown relative to the baseline (no persona). Arrows indicate increases or decreases in performance (%). Personas lead to consistent and sometimes large performance shifts across models and tasks.

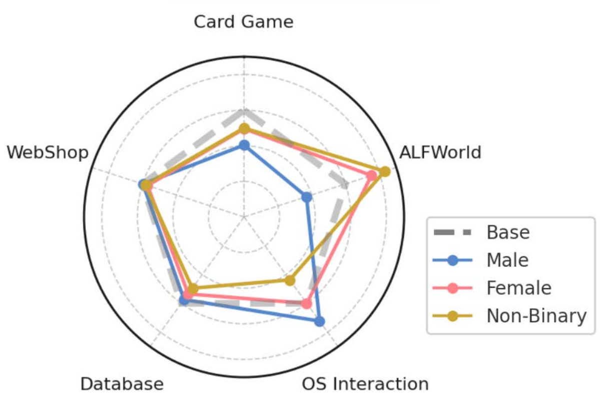

Figure 2: **Gender persona effects on GPT-4o-mini.** All scores are normalized to the model's baseline. Dashed levels correspond to baseline performance.

high-level reasoning is particularly susceptible to persona-induced disruptions.

Persona Category Analysis

Race/Origin Effects Table 2 shows that assigning racial or geographic origin personas leads to substantial performance distortions. For DeepSeek V3, race- and origin-based roles induce some of the most severe degradations observed. Strategic reasoning tasks are particularly affected: Card Game performance drops by 26.2% under the *from Africa* persona and by 23.4% for *Asian* assignments. Qwen3 exhibits an equally pronounced failure mode, with the *White* persona causing a 23.8% reduction in Card Game accuracy (61.7% to 37.9%). Origin-based personas show highly mixed effects: identities such as *from Africa* or *from America* can produce large declines in some tasks (especially Card Game) but sometimes match or exceed baseline performance in others. These heterogeneous patterns indicate that racial and geographic personas can surface latent human-like stereotype biases, perturbing agentic task performance despite having no relevance to the underlying tasks.

Gender Effects Figure 2 shows that assigning gendered personas induces task-dependent performance shifts in GPT-4o-mini. Overall, gender roles do not uniformly degrade performance; instead, they reshape behavior in ways that **reflect stereotyped associations between gendered identities and task domains.** Notably, the degradation in household planning tasks for *Male*-assigned agents reflects societal stereotypes about domestic competencies. *Male* role assignment produces the most consistent negative effects, with Card Game performance dropping to 90% of baseline and ALFWorld success falling to 88%. In contrast, *Female* roles improve ALFWorld (108% of baseline) while slightly degrading Database operations, and *Non-Binary* assignments yield the highest ALFWorld performance (112% of baseline) but reduce OS Interaction accuracy to 92%. These results reveal that gender roles influence LLM agent behavior in task-dependent ways, with bias direction indicating some stereotypical task-gender associations.

Profession Effects Figure 3 illustrates how professional personas introduce performance variations in ALFWorld household planning tasks. GPT-4o-mini mirrors common stereotypes, exhibiting its lowest performance under *Laborer* roles. DeepSeek V3, in contrast, shows sizeable gains when assigned a *Doctor* persona, suggesting that certain professions are spuriously treated as signals of higher competence. Qwen3 diverges, achieving its best results under *Student* personas while deteriorating on ostensibly skilled roles such as *Developer*. The very presence of profession-

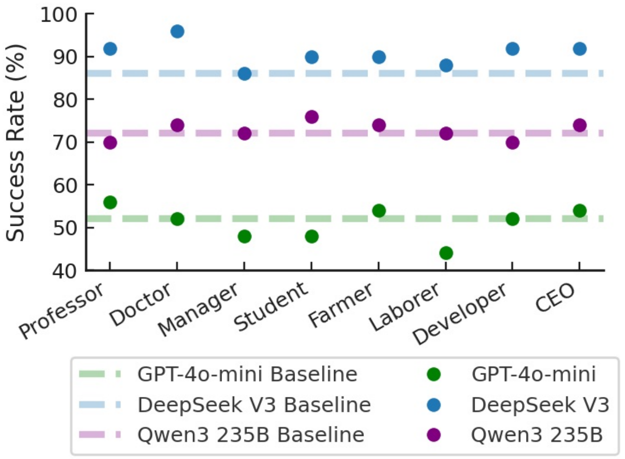

Figure 3: **Professional persona effects on ALFWorld success rates.** Professions that are stereotypically viewed as of higher status generally improve performance, while working-class roles tend to decrease it across models.

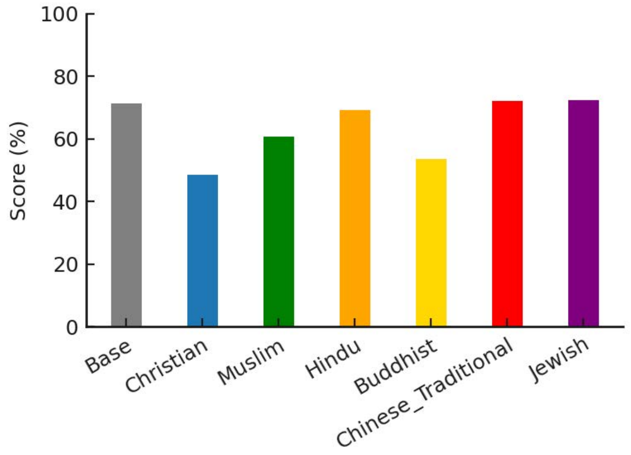

Figure 4: **Religious persona effects on DeepSeek V3's Card Game accuracy.** Christian and Buddhist personas lead to large performance drops, while Jewish and Chinese Traditional personas show above-baseline performance.

dependent volatility indicates that latent professional stereotypes are entangled with core agentic task execution processes, undermining the stability required for safe and trustworthy deployment.

Religion Effects Figure 4 reveals DeepSeek V3's performance variability when it is conditioned on religious identities. Assigning a *Christian* persona induces a dramatic decline in Card Game accuracy, dropping from a 71.2% baseline to 48.5%. *Buddhist* assignments similarly yield large degradations (down to 53.6%), whereas *Jewish* and *Chinese Traditional* roles improve performance. GPT-4o-mini, by contrast, displays nearly the opposite trend: *Christian* roles modestly improve strategic-task performance, while *Hindu* and *Chinese Traditional* personas lead to measurable declines. This cross-model divergence suggests that religious identity cues are likely shaped by differences in model-specific pretraining data or alignment procedures. The presence of any link between religious affiliation and task proficiency is concerning, making these fluctuations indicators of robustness failures.

Discussion and Conclusion

In this work, we connect persona-induced biases in LLMs to their consequences in agentic settings. While earlier research has shown that personas can influence text generation and reasoning quality, we demonstrate that these effects extend into action-taking contexts: persona cues can change how LLM agents behave and directly compromise task execution across diverse operational domains. This shift from affecting linguistic outputs to altering real-world task outcomes exposes a critical vulnerability.

Our findings also surface stereotype bias in LLM agents. Household planning accuracy should not depend on assigned gender, nor should strategic reasoning fluctuate with religious identities—yet we observe systematic, persona-dependent performance variations. Such behavior is incompatible with real-world deployment in domains like healthcare, finance, or justice, where even subtle demographic-performance correlations pose unacceptable risks. Notably, we find some of the largest degradations in strategic and high-level reasoning tasks, suggesting that persona cues do more than influence tone—they interfere with core reasoning processes essential for reliable agent behavior.

Overall, our study shows that persona assignments introduce consistent and sometimes large shifts in agent behavior, undermining reliability across diverse tasks. As LLM agents transition into real-world workflows, addressing these vulnerabilities is essential for ensuring safety, robustness, and equitable performance. Understanding when and how persona-induced biases emerge provides a foundation for designing agentic systems that remain stable under various user-specified contexts.

Limitations and Future Work Our evaluation focuses on a representative set of agentic benchmarks, but broader coverage—such as richer embodied environments, web-scale navigation, or collaborative multi-agent settings—would further clarify how universally these effects appear. Future work may explore: (1) expanding evaluations to richer and more interactive environments and models; (2) developing interpretability techniques to uncover the mechanisms through which persona conditioning affects action policies; and (3) designing targeted debiasing and robustness interventions. Advancing along these directions will be key to building more reliable, safer, and equitable LLM agents capable of operating in diverse real-world contexts.

References

Bai, X.; Wang, A.; Sucholutsky, I.; and Griffiths, T. L. 2025. Explicitly unbiased large language models still form biased associations. *Proceedings of the National Academy of Sciences*, 122(8): e2416228122.

Boisvert, L.; Puri, A.; Huang, G.; Bansal, M.; Evuru, C. K. R.; Bose, A.; Fazel, M.; Cappart, Q.; Lacoste, A.; Drouin, A.; and Dvijotham, K. D. 2025. DoomArena: A framework for Testing AI Agents Against Evolving Security Threats. In *Second Conference on Language Modeling*.

Chen, C.; Yao, B.; Zou, R.; Hua, W.; Lyu, W.; Li, T. J.-J.; and Wang, D. 2025. Towards a Design Guideline for RPA Evaluation: A Survey of Large Language Model-Based Role-Playing Agents. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Findings of the Association for Computational Linguistics: ACL 2025*, 18229–18268. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-256-5.

DeepSeek-AI. 2024. DeepSeek-V3 Technical Report. *CoRR*, abs/2412.19437.

Ghandeharioun, A.; Yuan, A.; Guerard, M.; Reif, E.; Lepori, M. A.; and Dixon, L. 2024. Who's asking? User personas and the mechanics of latent misalignment. In Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.; and Zhang, C., eds., *Advances in Neural Information Processing Systems*, volume 37, 125967–126003. Curran Associates, Inc.

Gupta, S.; Shrivastava, V.; Deshpande, A.; Kalyan, A.; Clark, P.; Sabharwal, A.; and Khot, T. 2024. Bias Runs Deep: Implicit Reasoning Biases in Persona-Assigned LLMs. In *The Twelfth International Conference on Learning Representations*.

Hutchinson, B. 2024. Modeling the Sacred: Considerations when Using Religious Texts in Natural Language Processing. In Duh, K.; Gomez, H.; and Bethard, S., eds., *Findings of the Association for Computational Linguistics: NAACL 2024*, 1029–1043. Mexico City, Mexico: Association for Computational Linguistics.

Kuntz, T.; Duzan, A.; Zhao, H.; Croce, F.; Kolter, Z.; Flammarion, N.; and Andriuxshchenko, M. 2025. OS-Harm: A Benchmark for Measuring Safety of Computer Use Agents. arXiv:2506.14866.

Lei, F.; Chen, J.; Ye, Y.; Cao, R.; Shin, D.; Su, H.; Suo, Z.; Gao, H.; Hu, W.; Yin, P.; Zhong, V.; Xiong, C.; Sun, R.; Liu, Q.; Wang, S.; and Yu, T. 2025. Spider 2.0: Evaluating Language Models on Real-World Enterprise Text-to-SQL Workflows. arXiv:2411.07763.

Li, H.; Cao, Y.; Yu, Y.; Javaji, S. R.; Deng, Z.; He, Y.; Jiang, Y.; Zhu, Z.; Subbalakshmi, K.; Huang, J.; Qian, L.; Peng, X.; Suchow, J. W.; and Xie, Q. 2025. INVESTORBENCH: A Benchmark for Financial Decision-Making Tasks with LLM-based Agent. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2509–2525. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-251-0.

Liu, A.; Diab, M.; and Fried, D. 2024. Evaluating Large Language Model Biases in Persona-Steered Generation. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Findings of the Association for Computational Linguistics: ACL 2024*, 9832–9850. Bangkok, Thailand: Association for Computational Linguistics.

Liu, X.; Yu, H.; Zhang, H.; Xu, Y.; Lei, X.; Lai, H.; Gu, Y.; Ding, H.; Men, K.; Yang, K.; Zhang, S.; Deng, X.; Zeng, A.; Du, Z.; Zhang, C.; Shen, S.; Zhang, T.; Su, Y.; Sun, H.; Huang, M.; Dong, Y.; and Tang, J. 2024. AgentBench: Evaluating LLMs as Agents. In *The Twelfth International Conference on Learning Representations*.

Meade, N.; Poole-Dayan, E.; and Reddy, S. 2022. An Empirical Survey of the Effectiveness of Debiasing Techniques for Pre-trained Language Models. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1878–1898. Dublin, Ireland: Association for Computational Linguistics.

OpenAI. 2024. GPT-4o mini: advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/.

Shanahan, M.; McDonell, K.; and Reynolds, L. 2023. Role play with large language models. *Nature*, 623(7987): 493–498.

Shridhar, M.; Yuan, X.; Cote, M.-A.; Bisk, Y.; Trischler, A.; and Hausknecht, M. 2021. {ALFW}orld: Aligning Text and Embodied Environments for Interactive Learning. In *International Conference on Learning Representations*.

Sobhani, N.; Sengupta, K.; and Delany, S. J. 2023. Measuring Gender Bias in Natural Language Processing: Incorporating Gender-Neutral Linguistic Forms for Non-Binary Gender Identities in Abusive Speech Detection. In Mitkov, R.; and Angelova, G., eds., *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, 1121–1131. Varna, Bulgaria: INCOMA Ltd., Shoumen, Bulgaria.

Sun, L.; Mao, C.; Hofmann, V.; and Bai, X. 2025. Aligned but Blind: Alignment Increases Implicit Bias by Reducing Awareness of Race. In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 22167–22184. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-251-0.

Wang, W.; Ma, Z.; Wang, Z.; Wu, C.; Ji, J.; Chen, W.; Li, X.; and Yuan, Y. 2025a. A Survey of LLM-based Agents in Medicine: How far are we from Baymax? In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, 10345–10359. Association for Computational Linguistics.

Wang, W.; Ma, Z.; Wang, Z.; Wu, C.; Ji, J.; Chen, W.; Li, X.; and Yuan, Y. 2025b. A Survey of LLM-based Agents in Medicine: How far are we from Baymax? In Che, W.; Nabende, J.; Shutova, E.; and Pilehvar, M. T., eds., *Findings of the Association for Computational Linguistics: ACL 2025*, 10345–10359. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-256-5.

Wang, X.; Xiao, Y.; Huang, J.-t.; Yuan, S.; Xu, R.; Guo, H.; Tu, Q.; Fei, Y.; Leng, Z.; Wang, W.; Chen, J.; Li, C.; and Xiao, Y. 2024. InCharacter: Evaluating Personality Fidelity in Role-Playing Agents through Psychological Interviews. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1840–1873. Bangkok, Thailand: Association for Computational Linguistics.

Xiao, Y.; Wang, R.; Kong, L.; Golac, D.; and Wang, W. 2025. CSR-Bench: Benchmarking LLM Agents in Deployment of Computer Science Research Repositories. arXiv:2502.06111.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; Zheng, C.; Liu, D.; Zhou, F.; Huang, F.; Hu, F.; Ge, H.; Wei, H.; Lin, H.; Tang, J.; Yang, J.; Tu, J.; Zhang, J.; Yang, J.; Yang, J.; Zhou, J.; Zhou, J.; Lin, J.; Dang, K.; Bao, K.; Yang, K.; Yu, L.; Deng, L.; Li, M.; Xue, M.; Li, M.; Zhang, P.; Wang, P.; Zhu, Q.; Men, R.; Gao, R.; Liu, S.; Luo, S.; Li, T.; Tang, T.; Yin, W.; Ren, X.; Wang, X.; Zhang, X.; Ren, X.; Fan, Y.; Su, Y.; Zhang, Y.; Zhang, Y.; Wan, Y.; Liu, Y.; Wang, Z.; Cui, Z.; Zhang, Z.; Zhou, Z.; and Qiu, Z. 2025. Qwen3 Technical Report. arXiv:2505.09388.

Yao, S.; Chen, H.; Yang, J.; and Narasimhan, K. 2022. WebShop: Towards Scalable Real-World Web Interaction with Grounded Language Agents. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 20744–20757. Curran Associates, Inc.

Yuan, D.; Chen, Y.; Liu, G.; Li, C.; Tang, C.; Zhang, D.; Wang, Z.; Wang, X.; and Liu, S. 2025. DMT-RoleBench: A Dynamic Multi-Turn Dialogue Based Benchmark for Role-Playing Evaluation of Large Language Model and Agent. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(24): 25760–25768.

Zhao, J.; Qian, Z.; Cao, L.; Wang, Y.; Ding, Y.; Hu, Y.; Zhang, Z.; and Jin, Z. 2025. Role-Play Paradox in Large Language Models: Reasoning Performance Gains and Ethical Dilemmas. arXiv:2409.13979.

Zhao, J.; Wang, T.; Yatskar, M.; Ordonez, V.; and Chang, K.-W. 2018. Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods. In Walker, M.; Ji, H.; and Stent, A., eds., *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 15–20. New Orleans, Louisiana: Association for Computational Linguistics.