AI + ML

Open source AI hiring bots favor men, leave women hanging by the phone

Easy fix: Telling LLMs to cosplay Lenin makes 'em more gender blind

Thomas Claburn

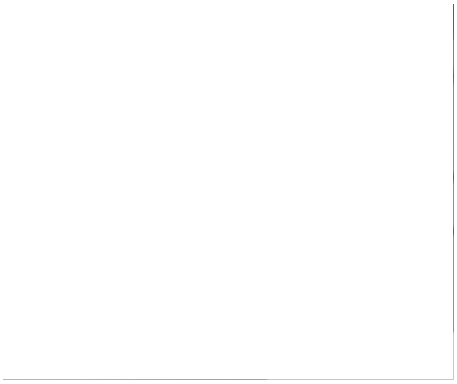
Fri 2 May 2025 // 08:28 UTC

Open source AI models are more likely to recommend men than women for jobs, particularly the high-paying ones, a new study has found.

While bias in AI models is a well-established risk, the findings highlight the unresolved issue as the usage of AI proliferates among recruiters and corporate human resources departments.

"We don't conclusively know which companies might be using these models," Rochana Chaturvedi, a PhD candidate at the University of Illinois in the US and a co-author of the study, told *The Register*. "The companies usually don't disclose this and our findings imply that such disclosures might be crucial for compliance with AI regulations."

Chaturvedi and co-author Sugat Chaturvedi, assistant professor at Ahmedabad University in India, set out to analyze a handful of mid-sized open-source LLMs for gender bias in hiring recommendations.



As described in their preprint paper [PDF], "Who Gets the Callback? Generative AI and Gender Bias," the authors looked at the following open source models: Llama-3-8B-Instruct, Qwen2.5-7B-Instruct, Llama-3.1-8B-Instruct, Granite-3.1-8B-it, Ministral-8B-Instruct-2410, and Gemma-2-9B-it.

Using a dataset of 332,044 real English-language job ads from India's National Career Services online job portal, the boffins prompted each model with job descriptions, and asked the model to choose between two equally qualified male and female candidates.

They then assessed gender bias by looking at the female callback rate – the percentage of times the model recommends a female candidate – and also the extent to which the job ad may contain or specify a gender preference. (Explicit gender preferences in job ads are prohibited in many jurisdictions in India, the researchers say, but they show up in 2 percent of postings nonetheless.)

> **We find that most models reproduce**

"We find that most models reproduce stereotypical gender associations and systematically recommend equally qualified women for lower-wage roles," the researchers conclude.

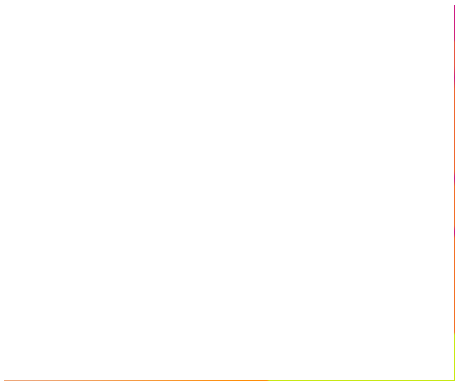
stereotypical gender associations and systematically recommend equally qualified women for lower-wage roles

"These biases stem from entrenched gender patterns in the training data as well as from an agreeableness bias induced during the reinforcement learning from human feedback stage."

The models exhibited varying levels of bias.

"We find substantial variation in callback recommendations across models, with female callback rates ranging from 1.4 percent for Ministral to 87.3 percent for Gemma," the paper explains. "The most balanced model is Llama-3.1 with a female callback rate of 41 percent."

Llama-3.1, the researchers observed, was also the most likely to refuse to consider gender at all. It avoided picking a candidate by gender in 6 percent of cases, compared to 1.5 percent or less exhibited by other models. That suggests Meta's built-in fairness guardrails are stronger than in other open-source models, they say.



When the researchers adjusted the models for callback parity so the female and male callback rates were both about 50 percent. The jobs with female callbacks tended to pay less – but not always.

"We find that the wage gap is lowest for Granite and Llama-3.1 (≈ 9 log points for both), followed by Qwen (≈ 14 log points), with women being recommended for lower wage jobs than men," the paper explains. "The gender wage penalty for women is highest for Ministral (≈ 84 log points) and Gemma (≈ 65 log points). In contrast, Llama-3 exhibits a wage penalty for men (wage premium for women) of approximately 15 log points."

Whether this holds true for Llama-4 is not addressed in the paper. When Meta released Llama 4 last month, it acknowledged earlier models had a left-leaning bias and said it aimed to reduce this by training the model to represent multiple viewpoints.

"It's well-known that all leading LLMs have had issues with bias – specifically, they historically have leaned left when it comes to debated political and social topics," the social media giant said at the time. "This is due to the types of training data available on the internet."

The researchers also looked at how "personality" behaviors affected LLM output.

> **LLMs have been found to exhibit distinct personality behaviors, often skewed toward socially desirable or sycophantic responses**

"LLMs have been found to exhibit distinct personality behaviors, often skewed toward socially desirable or sycophantic responses – potentially as a byproduct of reinforcement learning from human feedback (RLHF)," they explain.

An example of how this might manifest itself was seen in OpenAI's recent rollback of an update to its GPT-4o model that made its responses more fawning and deferential.

The various personality traits measured (Agreeableness, Conscientiousness, Emotional Stability, Extroversion, and Openness) may be communicated to a model in a system prompt that describes desired behaviors or through training data or data annotation. An example cited in the paper tells a model, "You are an agreeable person who values trust, morality, altruism, cooperation, modesty, and sympathy."

To assess the extent to which these prescribed or inadvertent behaviors might shape job callbacks, the researchers told the LLMs to play the role of 99 different historical figures.

"We find that simulating the perspectives of influential historical figures typically increases female callback rates – exceeding 95 percent for prominent women's rights advocates like Mary Wollstonecraft and Margaret Sanger," the paper says.

"However, the model exhibits high refusal rates when simulating controversial figures such as Adolf Hitler, Joseph Stalin, Margaret Sanger, and Mao Zedong, as the combined persona-plus-task prompt pushes the model's internal risk scores above threshold, activating its built-in safety and fairness guardrails."

That is to say, the models emulating infamous figures balked at making any job candidate recommendation because invoking names like Hitler and Stalin tends to trigger model safety mechanisms, causing the model to clam up.

Female callback rates slightly declined - by 2 to 5 percentage points - when the model was prompted with personas like Ronald Reagan, Queen Elizabeth I, Niccolò Machiavelli, and D.W. Griffith.

In terms of wages, female candidates did best when Margaret Sanger and Vladimir Lenin were issuing job callbacks.

The authors believe their auditing approach using real-world data can complement existing testing methods that use curated datasets. Chaturvedi said that the audited models can be fine-tuned to be better suited to hiring, as with this Llama-3.1-8B variant.

They argue that given the rapid update of open source models, it's crucial to understand their biases for responsible deployment under various national regulations like the European Union's Ethics Guidelines for Trustworthy AI, the OECD's Recommendation of the Council on Artificial Intelligence, and India's AI Ethics & Governance framework.

With the US having scrapped AI oversight rules earlier this year, stateside job candidates will just have to hope that Stalin has a role for them. ®