GENDER BIASES IN LARGE LANGUAGE MODELS

"AI-based systems often perpetuate (and even scale and amplify) human, structural, and social biases. These biases are not only difficult to mitigate, but can also cause harm at the individual, collective, or societal level."

Sources of bias in AI

- Biases in data
- Biases in algorithm selection
- Biases in implementation

"Nearly 9 out of 10 men and women hold fundamental biases towards women," and these biases are reflected in AI systems.

Ways AI perpetuates biases

In the job dispolacement

- Disproportionate unemployment in sectors where women are an important part of the workforce

In the AI development process

- Underrepresentation of women
- Weak regulation facilitates the misuse of AI

In decision making

- Personnel selection
- Finance and insurance

Social implications

- Damage to social cohesion
- Gender-based violence
- Homogenization of vulnerable populations

"The increasing complexity of artificial intelligence systems demands more rigorous efforts to achieve fairness in AI-driven decisions and interactions."

How to detect biases in large language models (LLM)

- By analyzing how an LLM associates different concepts in interaction
- By observing how an LLM improvises text around a specific topic in practice

Advantages of open source models

- Transparency and auditability
- Facilitates independent research and benchmarking
- Community-driven improvement

"It is essential to take measures in the early stages of the artificial intelligence development cycle to prevent biases."

Target audiences: 📚 Librarians, 👩‍🏫 Professors, 👨‍🎓 Students

REFERENCES

UNESCO (2024). *Challenging systematic prejudices: an investigation into bias against women and girls in large language models.* https://unesdoc.unesco.org/ark:/48223/pf0000388971

GEDIS project context

This resource is part of the European project GEDIS (Gender Diversity in Information Science), which promotes open educational tools to tackle gender inequalities in higher education, with an emphasis on disciplines related to Information and Library Science.

GEDIS - Gender Diversity in Information Science: Challenges in Higher Education. https://ub.edu/gedis

Citation: Villarroya, Anna, Juan-José Boté-Vericad, Maddalena Fedele, Mari Vállez. 2026. *Biaixos de gènere en els models de llenguatge externs*. DOI: 10.5281/zenodo.18171512 *Gender biases in large language models*. Translation by: Anna Villarroya & Juan-José Boté-Vericad. DOI: 10.5281/zenodo.18188657

Co-funded by the European Union. The views and opinions expressed are solely those of the author(s) and do not necessarily reflect those of the European Union or the Spanish Service for the Internationalisation of Education (SEPIE). Neither the EU nor the granting authority can be held responsible.