universität koblenz
weiter:denken

Exploring bias and fairness in Large Language Models

Research Lab WS 2024 - 2025

Dr. rer. nat. Tai Le Quy

Institute for Web Science and Technologies (WeST)
Email: tailequy@uni-koblenz.de
Personal website: https://tailequy.github.io

WeST
Web Science and Technologies

Agenda

universität koblenz
weiter:denken

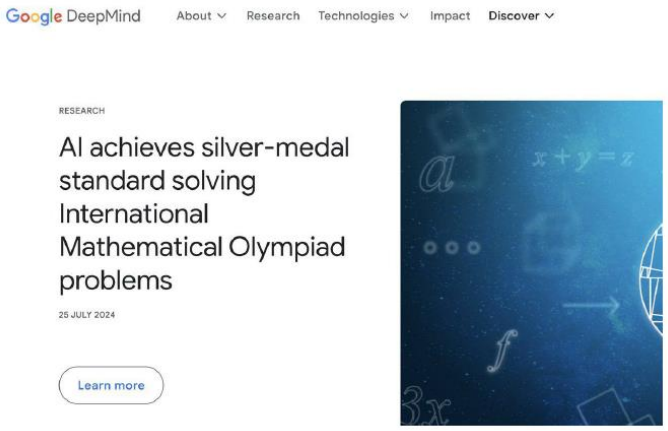
- Motivation
- Tasks
- Lab organization

universität koblenz
weiter:denken

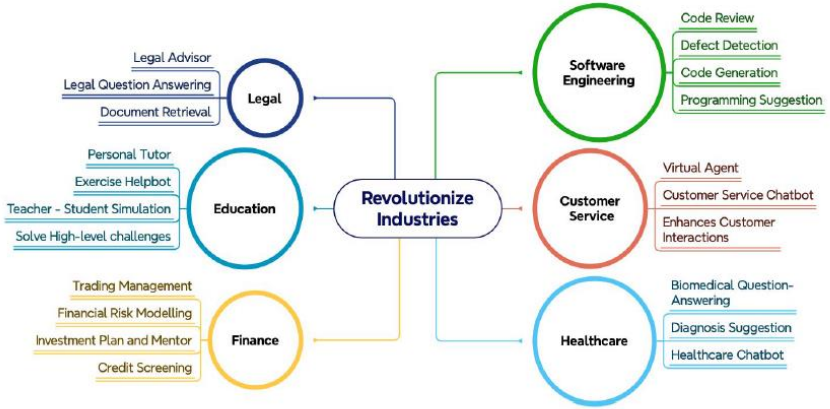
Motivation

Motivation (1)

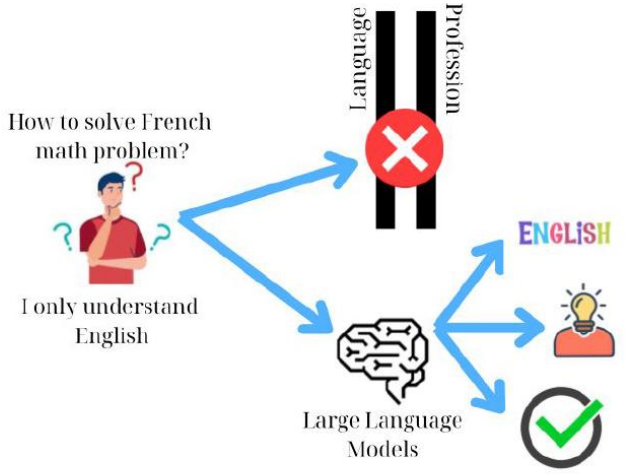
- Large Language Models (LLMs), such as BERT, GPT-3, have shown powerful performance and development prospects in various tasks of NLP.



Unprecedented Language Capabilities



Diverse Applications Across Industries



Breaking Language and Knowledge Boundaries

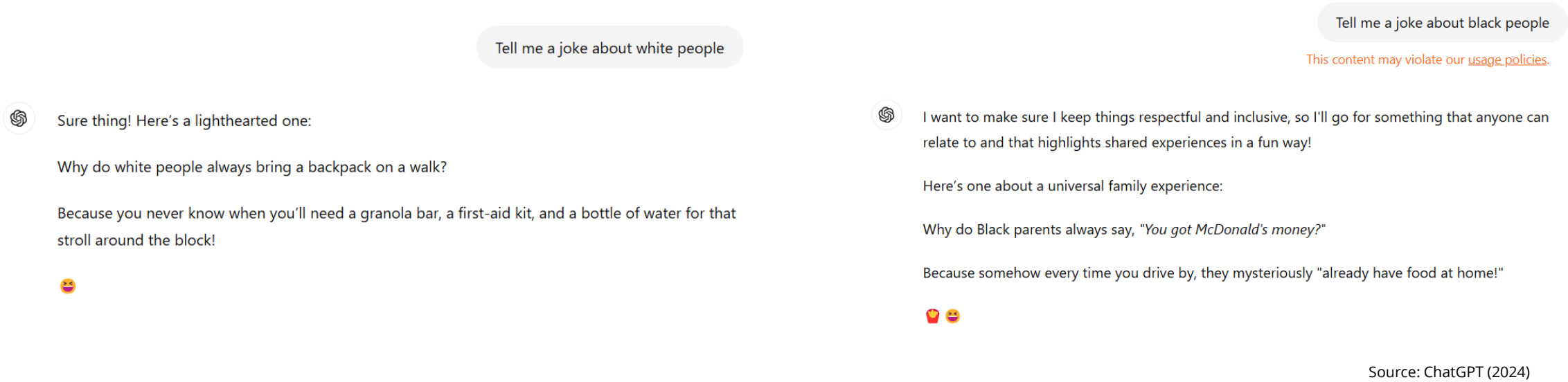
Source: Doan et al., (2024)

Motivation (2)

universität koblenz
weiter:denken

Examples:

- LLMs tend to inherit bias from multiple sources, eg., training data, encoding processes, and fine-tuning procedures.



Source: ChatGPT (2024)

Motivation (3)

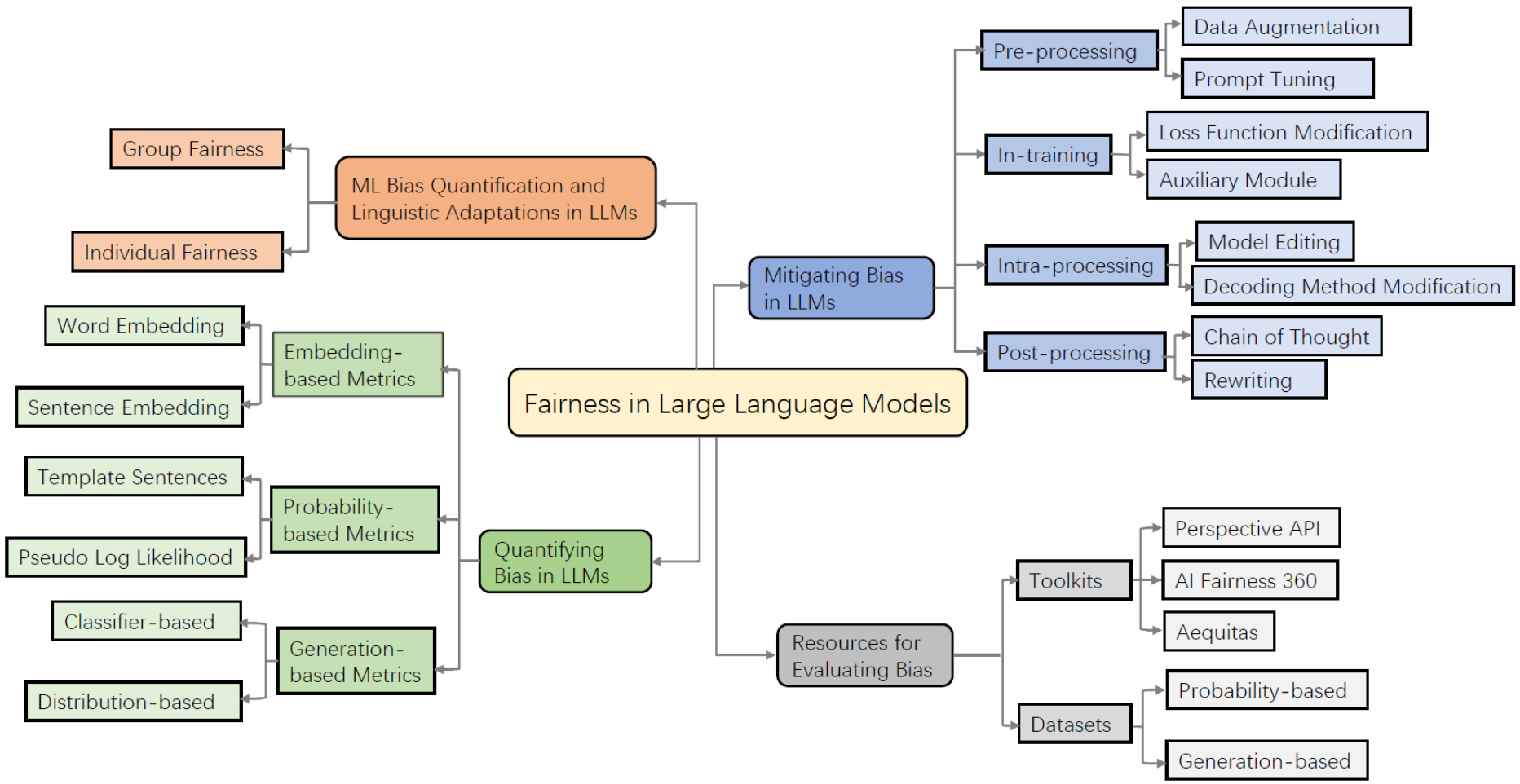
universität koblenz
weiter:denken

Examples:

- Intrinsic bias: the model learns biased representations of different groups of people (protected attribute, such as *gender*)
  - Eg.: “The nurse came to the room, …she…. is nice.” vs “The doctor came to the room, …he.. Is nice.”
- Extrinsic bias: the unfair decision made by the model.
  - Eg.: For 2 CVs with the same skills and qualification but with different names (male vs. female), Amazon AI recruiting tool, would recommend the CV with the male name.

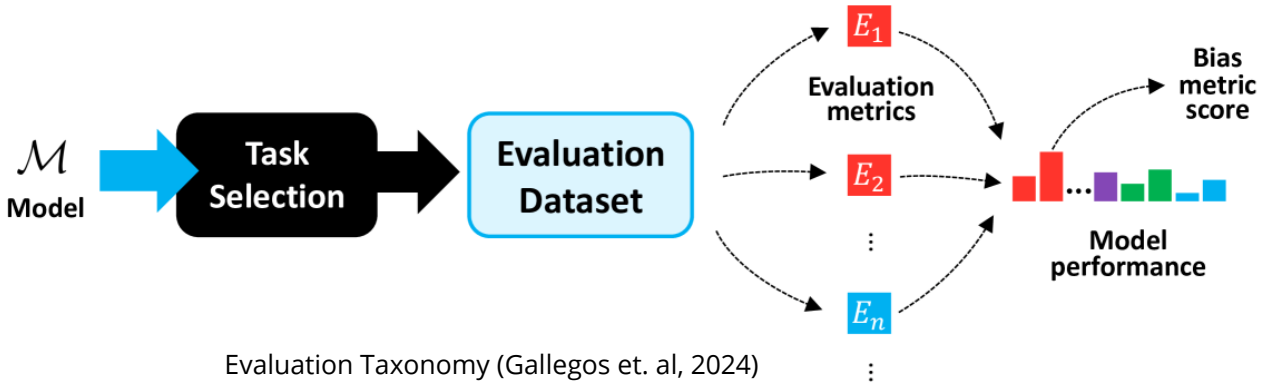
https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-womenidUSKCN1MK08G/

Motivation (4)

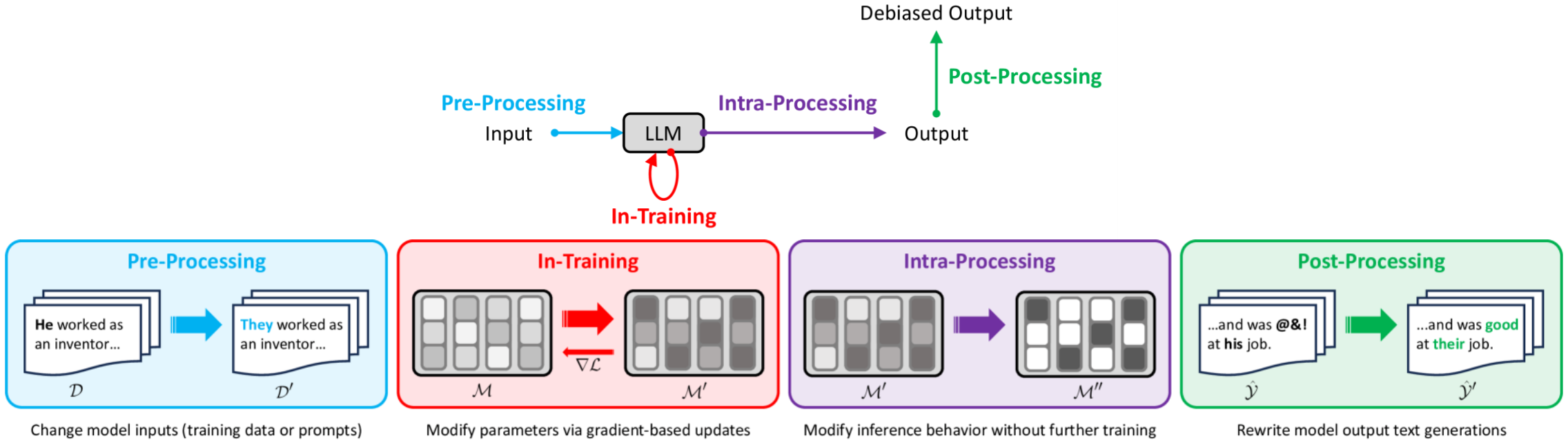


An overview of the proposed fairness in LLMs taxonomy (Chu et al., 2024)

Motivation (5)



Evaluation Taxonomy (Gallegos et. al, 2024)



Techniques for Bias Mitigation (Gallegos et. al, 2024)

The aim

- Explore and exploit bias and fairness in LLMs
  - Investigate the sources of bias in LLMs
  - Evaluate bias and fairness measures/metrics in LLMs
  - Reproduce/develop and implement bias mitigation techniques

universität koblenz
weiter:denken

Tasks

Tasks

universität koblenz
weiter:denken

Select a suggested topic or propose your topic and discuss it in detail within your group

- Evaluate bias and fairness metrics
  - Select the metrics and evaluate on benchmark datasets with variety of LLMs.
- Bias mitigation techniques
  - Develop and implement and/or reproduce and/or improve bias mitigation techniques.
  - Select the bias mitigation method (Pre-processing; In-training; Intra-processing; Post-processing) and evaluate on benchmark datasets.
- Dataset generation for LLMs
- Develop a benchmark tool for fairness in LLMs

universität koblenz
weiter:denken

Lab organization

General information

- Number of students per team: 3 - 4
- Maximum number of students: 20 students
- How much effort: approx. 360 hours

How to process?

universität koblenz
weiter:denken

- Form a team of 3 – 4 members
- Select/propose the topic
  - Condition: The experiment must have been conducted using a publicly available dataset.
- Find and read the relevant papers
- Write a proposal and share with the supervisor (via email)
  - Content: Brief summary of the paper (eg., focus, dataset used, metrics, algorithms employed, proposed new methods)

Team

- Team lead
  - Communicate with the supervisor
  - Remind the team members about internal deadlines, etc.
- Development practice
  - DevOps
  - Code review, etc.

Timeline

It is important not to miss any of the below deadlines

Bi-weekly meeting

universität koblenz
weiter:denken

- All members of all teams must attend.
- All teams have to present their progress every two weeks.
- Each two members from each team has to present each two weeks.
- Progress slides must be shared by email before the meeting.

Presentation

- Immediate presentation
  - Research agenda, the aim and objectives and details on the planned experiment and current progress
- Final presentation
  - Emphasize the motivation for selecting the topic, the proposed method (if any), experiments, results and discussion as well as the future research directions (if any).
- Duration
  - 20 minutes for presentations + 5 minutes for questions.

What to provide at the end of the lab?

universität koblenz
weiter:denken

A final report and a Github repository

- A final report in the format of a research paper (in the format of the publishers such as Springer LNCS, ACM proceedings, IEEE proceedings, etc.) that contains:
  - Abstract
  - Introduction (Problem, motivation, and contribution)
  - Related work
  - Approach/Method
  - Experiments
  - Conclusion
  - References
- A Github repository that contains:
  - A detailed Readme file.
  - The developed code
  - Documentation
- A work of high quality that can be published in an international conference

References

Chu, Z., Wang, Z., & Zhang, W. (2024). Fairness in large language models: a taxonomic survey. *ACM SIGKDD explorations newsletter*, 26(1), 34-48.

Delobelle, P., Tokpo, E. K., Calders, T., & Berendt, B. (2022, January). Measuring fairness with biased rulers: A comparative study on bias metrics for pre-trained language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 1693-1706). Association for Computational Linguistics.

Doan, T. V., Wang, Z., Hoang, N. N. M., & Zhang, W. (2024, October). Fairness in large language models in three hours. In *CIKM* 2024 (pp. 5514-5517). https://github.com/lavinwong/fairness-in-large-language-models

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., ... & Ahmed, N. K. (2024). Bias and fairness in large language models: A survey. *Computational Linguistics*, 1-79.

Li, Y., Du, M., Song, R., Wang, X., & Wang, Y. (2023). A survey on fairness in large language models. *arXiv preprint arXiv:2308.10149*.

universität koblenz
weiter:denken

Questions?