Evaluation of bias and robustness in search and conversational systems

Abolghasemi, A.

Stellingen

Behorende bij het proefschrift

Evaluation of Bias and Robustness in Search and Conversational Systems

1. Lexical richness does not necessarily guarantee improved performance in language model–based retrieval and may even diminish the relative benefit of contextualization compared to traditional term-based retrieval (Chapter 2).

2. Large language models can be used to systematically generate satisfaction-focused counterfactual dialogues, that is, dialogues for which the user has the opposite satisfaction label (satisfactory versus dissatisfactory) (Chapter 3).

3. Measuring societal bias in a ranked list of documents based on term-based group representations can capture only a partial and sometimes misleading view of societal bias (Chapter 4).

4. There can be a discrepancy between the fairness observed in the ranked outputs of a ranking model and the underlying bias embedded in the ranking model itself (Chapter 4).

5. Source attribution reliability in retrieval-augmented large language models depends on factors beyond document relevance (Chapter 5).

6. In studying biases in retrieval-augmented large language models, careful analysis of confounding factors in the experimental setup is essential, as the outputs of these models are affected by numerous retrieval/generation configuration choices.

7. Systematically exploring alternative what-if scenarios can be used to evaluate and identify potential biases in large language models.

8. Evaluation methodologies themselves must be continuously evaluated and refined; otherwise, systems might be optimized toward suboptimal objectives.

9. As interactions between humans and large language models continue to increase, their biases should be continuously evaluated.

10. Domain experts who use large language models within their own fields may overestimate model performance based on plausible but not necessarily correct outputs observed in a small number of interactions. Such users should have a basic understanding of how performance of a large language model on a specific task is evaluated.

Amin Abolghasemi
Leiden, March 6, 2026