



Universiteit
Leiden

The Netherlands

Evaluation of bias and robustness in search and conversational systems

Abolghasemi, A.

Citation

Abolghasemi, A. (2026, March 6). *Evaluation of bias and robustness in search and conversational systems*. Retrieved from <https://hdl.handle.net/1887/4296728>

Version: Publisher's Version

License: [Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden](#)

Downloaded from: <https://hdl.handle.net/1887/4296728>

Note: To cite this publication please use the final published version (if applicable).

Evaluation of Bias and Robustness in Search and Conversational Systems

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof.dr. S. de Rijcke,
volgens besluit van het college voor promoties
te verdedigen op vrijdag 6 maart 2026
klokke 11:30 uur

door

Amin Abolghasemi

geboren te Teheran

Promotores:

Prof. dr. S. Verberne
Prof. dr. M. de Rijke University of Amsterdam

Co-promotor:

Dr. L. Azzopardi Microsoft

Promotiecommissie:

Prof. dr. M. M. Bonsangue
Prof. dr. ir. W. Kraaij
Prof. dr. A. Hanbury TU Wien
Prof. dr. E. Kanoulas University of Amsterdam
Dr. M. Maistro University of Copenhagen



**Universiteit
Leiden**
The Netherlands

The work in this dissertation was funded by European Union's Horizon 2020 research and innovation program, Marie Skłodowska-Curie grant agreement No. 860721.

The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

Copyright © 2026 by A. Abolghasemi

An electronic copy of this dissertation is available at
<https://theses.liacs.nl>.

Cover designed by Amin Abolghasemi

CONTENTS

1	Introduction	1
1.1	Research Outline and Questions	3
1.1.1	Evaluating Contextualized Lexical Models in Query-by-Example Retrieval	3
1.1.2	Robust User Satisfaction Estimation in Task-Oriented Dialogue Systems	4
1.1.3	Measuring Societal Bias in Ranked Lists of Documents	4
1.1.4	Attribution Sensitivity and Bias in RAG	5
1.2	Contributions	6
1.2.1	Methodological Contributions	6
1.2.2	Empirical Contributions	6
1.2.3	Resource Contributions	7
1.3	Thesis Origins	7
2	Contextualized Term-based Ranking	11
2.1	Introduction	12
2.2	Background: Retrieval Models	14
2.2.1	Traditional lexical matching models	14
2.2.2	Term Independent Likelihood Model: TILDE	14
2.2.3	Lexical Exact Matching: TILDEv2	15
2.2.4	Cross-encoder BERT Ranker	16
2.3	Methods and Experimental Settings	16
2.3.1	Evaluation Benchmark	16
2.3.2	BERT-based Tokenization in Traditional Models	17
2.3.3	Interpolation between BM25 and TILDE (TILDEv2) scores	18
2.3.4	Document Expansion with TILDE	18
2.3.5	Domain-Specific BERT in TILDE and TILDEv2	19
2.3.6	Implementation Details	19
2.4	Results	19
2.5	Discussion	24
2.5.1	Interpolation effectiveness	24
2.5.2	Interpolation weight	25
2.6	Conclusion	25
3	CAUSE: Counterfactual Assessment of User Satisfaction Estimation	27
3.1	Introduction	28
3.2	Related Work	30
3.2.1	User Satisfaction Estimation in TODSS	30

3.2.2	Counterfactual Data Generation	31
3.3	User Satisfaction Estimation	31
3.4	Methodology	32
3.4.1	Counterfactual Utterance Generation	32
3.4.2	User Satisfaction Estimation using LLMs	32
3.5	Experimental Setup	33
3.5.1	Benchmarks	33
3.5.2	Evaluation Metrics	34
3.5.3	Baselines	34
3.5.4	Human Annotation	34
3.6	Experimental Results	35
3.6.1	Data Quality	35
3.6.2	User Satisfaction Estimation Results	35
3.7	Conclusion	39
3.A	Appendix	40
3.A.1	Counterfactual Response Generation Prompt	40
3.A.2	Full-dialogue Counterfactual Examples	40
4	Measuring Bias in a Ranked List using Term-based Representations	43
4.1	Introduction	43
4.2	Background	45
4.3	Methodology	47
4.4	Experimental Setup	50
4.5	Results	50
4.6	Discussion	53
4.7	Conclusion	55
5	Evaluation of Attribution Bias in Retrieval-Augmented LLMs	57
5.1	Introduction	57
5.2	Background	60
5.3	Methodology	60
5.3.1	RAG Modes	60
5.3.2	Answer/Attribution Generation	61
5.3.3	Evaluation Metrics	62
5.4	Experimental Settings	64
5.5	Experimental Results	64
5.6	Conclusion and Future Work	70
5.A	Synthetic Document Generation	71
5.B	Authorship Informed Answer/Attribution Generation Prompt	72
5.C	Extended Set of Authorship Labels	72
5.D	Effect of the Number of Source Documents	75
5.E	Effect of the Retriever	75
5.F	Attribution Quality Results	76
5.G	Confidence Results	76
5.H	Average Number of Cited Documents	76
5.I	Mixed RAG Mode Results	76

5.J	Examples	76
6	Conclusions	85
6.1	Main Findings	85
6.2	Future Directions	88
6.2.1	Evaluating Contextualized Lexical Models in Query-by-Example Retrieval (Chapter 2)	88
6.2.2	Robust User Satisfaction Estimation in Task-Oriented Dialogue Systems (Chapter 3)	89
6.2.3	Measuring Societal Bias in Ranked Lists of Documents (Chapter 4)	89
6.2.4	Attribution Sensitivity and Bias in RAG (Chapter 5)	89
6.2.5	Final Thoughts: Towards Evaluating Agentic Systems	90
	Bibliography	91
	Summary	111
	Samenvatting	113
	Acknowledgements	115
	Curriculum Vitæ	117