Evaluation of bias and robustness in search and conversational systems

Abolghasemi, A.

6

CONCLUSIONS

This chapter wraps up the dissertation by highlighting the key findings and outlining potential directions for future research. In Section 6.1, we first reflect on the research questions we asked in Chapter 1 based on the experimental results and findings of the previous chapters. Then, in Section 6.2, we identify potential directions for future research that could build upon the work presented in this dissertation.

6.1. MAIN FINDINGS

In this section, we present our key findings by revisiting the research questions introduced in Chapter 1.

> **RQ1** *How generalizable is contextualized term-based ranking to retrieval settings with lexically rich queries?*

To answer **RQ1**, in Chapter 2, we studied the generalizability of two contextualized term-based ranking models, TILDE and TILDEv2, within the query-by-example (QBE) retrieval setting. In contrast to ad-hoc retrieval, QBE typically involves significantly longer queries which brings more lexical richness for performing retrieval. Our aim was to assess whether the relative performance of these models (compared to both traditional term-based approaches and the strong cross-encoder BERT ranker) extends to these lexically-rich contexts.

Our findings in Chapter 2 reveal that, consistent with the original studies [210, 211], the two contextualized term-based ranking models, TILDE and TILDEv2 perform worse than the BERT cross-encoder ranker in the QBE setting, despite the presence of longer queries that could provide richer context. However, unlike those earlier studies, where TILDE and TILDEv2 outperformed the BM25 baseline, our evaluation shows that BM25 maintains competitive effectiveness in QBE, and, in some instances, even surpasses the performance of the two contextualized term-based ranking models.

This observation is significant for two main reasons: (1) it highlights the unique challenges posed by retrieval settings that deviate from widely used benchmarks

such as MSMARCO and the TREC DL Track, and (2) it raises important questions about the applicability of other contextualized term-based models in such scenarios. Overall, our results suggest that QBE retrieval, as a retrieval setup with lexical richness, is structurally distinct from traditional IR tasks and thus requires specific development of retrieval models/methods.

In addition, we explored the effect of interpolating BM25 scores with those of TILDE and TILDEv2. We found that linear interpolation leads to enhanced ranking performance, indicating that the relevance signals from these contextualized models are complementary to those captured by BM25. Our further analysis using oracle interpolation supports this finding, which suggests that more nuanced combination strategies could yield even greater improvements by leveraging the strengths of both types of models.

RQ2 *How robust are user satisfaction estimators in task-oriented dialogue systems with more dissatisfactory user experiences?*

To address **RQ2**, in Chapter 3, we first extended two widely used benchmarks for user satisfaction estimation in task-oriented dialogue systems, MultiWoZ [52] and SGD [142], by incorporating a larger set of dissatisfactory dialogue samples. To generate these dissatisfactory dialogue samples, we introduced satisfaction-oriented counterfactual dialogue generation with LLMs: given a dialogue sample with a specific satisfaction label (e.g., satisfactory), we generate a corresponding counterpart (e.g., dissatisfactory), in which the user satisfaction is deliberately altered. We then conducted human annotation on the resulting generated dialogues to ensure the quality of satisfaction labels for these generated dialogues. Using these augmented test collections, we demonstrated a notable discrepancy in the performance of satisfaction estimators between the original datasets and those containing a higher proportion of dissatisfaction cases. We examined model robustness under varying class distributions by gradually increasing the proportion of dissatisfaction dialogue samples in the test sets. Specifically, while fine-tuned state-of-the-art models, BERT and ASAP [75, 194], performed strongly on the original, imbalanced test sets, their performance dropped sharply as dissatisfaction samples increased. In contrast, few-shot in-context learning with LLMs demonstrated greater sensitivity to dissatisfaction: LLMs often surpassed or matched fine-tuned models as the class distribution became more balanced, i.e., as test sets included more dissatisfactory dialogue samples. This highlighted LLMs' potential for reliably detecting user dissatisfaction, a critical factor for deploying dialogue systems. Moreover, the discrepancy in the performance of various user satisfaction estimators under different class distributions of dialogue samples highlighted the limitations in their generalizability and robustness across alternative evaluation setups.

In summary, our findings in Chapter 3 exposed a key gap in prior work: the lack of attention to the robustness of satisfaction estimators, especially in identifying user dissatisfaction. Furthermore, our results highlighted the importance of data augmentation strategies to improve the training of such estimators. We hypothesized that incorporating more balanced training data can enhance model robustness. In

addition, Chapter 3 illustrated the potential of large language models in generating high-quality counterfactual dialogue examples, which suggests a promising direction for augmenting training data in satisfaction estimation tasks.

RQ3 *How to effectively measure the societal bias in a ranked list of documents based on group-representative term sets?*

To address **RQ3**, in Chapter 4, we first identified a key limitation in the widely used group fairness metric NFaiRR [146], which assesses fairness based on the individual unbiasedness scores of documents within a ranked list. This approach to fairness calculation results in the effects of different documents not being able to cancel each other out. For example, if the top-ranked document is biased toward female groups for a given query and the second-ranked document is biased toward male groups, these opposing biases do not offset one another. To address this issue, we introduced a new metric, TExFAIR, which extends the previously proposed AWRF metric [51, 141, 153] by incorporating two components: (1) term-based associations, which link documents to societal groups through predefined sets of representative terms, with each set serving as a proxy for the presence of a particular societal group within the retrieved content; and (2) a rank-biased discounting factor that accounts for the reduced influence of non-representative documents (i.e., documents that do not include any group representative terms) in the ranked list. Due to these structural differences, TExFAIR captures a distinct dimension of fairness compared to NFaiRR. Consequently, when fairness is considered during model selection (for example, when a combined metric of fairness and effectiveness is used) TExFAIR and NFaiRR may lead to different model choices.

In Chapter 4, we also carried out a counterfactual evaluation to estimate the inherent group biases – specifically gender-related – present in ranking models. This analysis revealed a discrepancy between the fairness observed in the ranked outputs (as measured by NFaiRR or TExFAIR) and the underlying bias embedded in the ranking models themselves. However, due to the limitations of term-based fairness evaluation, exploring more semantically grounded approaches is required to better understand the relationship between model-level biases and the fairness of the rankings they generate. Furthermore, the limitations of relying on term-based group representations, which may not align with real users' perceptions of fairness, necessitate more user-centered methodologies for assessing societal fairness in ranked lists of documents.

RQ4 *How sensitive and biased are LLMs to the generators of source documents in attributive retrieval-augmented generation?*

To address **RQ4**, in Chapter 5, we introduced and examined the concepts of attribution sensitivity and bias in retrieval-augmented LLMs in relation to the authorship metadata of their source documents. We proposed a structured evaluation framework based on counterfactual evaluation of the effect of authorship metadata

in source documents. Our findings in Chapter 5 showed that including authorship information in the source documents of attributive retrieval-augmented LLMs can significantly affect their attribution behavior: LLMs cited different documents for their generated answers when informed about the author (generator) of the input source documents. Additionally, experiments across three LLMs revealed a consistent bias toward documents with explicit human authorship, which competes with prior research suggesting that LLMs often favor AI-generated content over human-written material.

This behavior in LLMs could be attributed to different factors such as training cues that LLMs could pick up during their pretraining over large scale data. Also, safeguard fine-tuning of LLMs could have an effect. However, deeper investigation into the causes of this sensitivity and bias would require access to the implementation, training, and fine-tuning of these models, which is beyond the scope of our work in Chapter 5. Our results in Chapter 5 underscore an important vulnerability in how LLMs attribute content. This brittleness in attribution can be exploited in both beneficial and harmful ways; for instance, a user might manipulate LLM outputs in their favor by embedding authorship cues in their documents.

6.2. FUTURE DIRECTIONS

In this section, we discuss the limitations of the research presented in this thesis and suggest possible directions for future work.

6.2.1. EVALUATING CONTEXTUALIZED LEXICAL MODELS IN QUERY-BY-EXAMPLE RETRIEVAL (CHAPTER 2)

In query-by-example (QBE) retrieval, the lexical richness of queries creates conditions that differ substantially from generic ad hoc retrieval, where user queries are typically short and less diverse in vocabulary. Our findings in Chapter 2 showed that this abundance of lexical relevance signals may diminish the added value of contextualization for models such as BM25, raising questions about the generalizability of contextualized approaches. However, other retrieval models, including dense retrieval model, may still benefit from contextualization in QBE. Future research should therefore examine the generalizability of such methods to QBE. This is particularly important, as the long query contexts in QBE introduce additional semantic complexities that further distinguish it from standard retrieval tasks. Prior work [15] has already shown that developing effective QBE methods with dense retrieval models is highly task-specific, and that ranking models cannot be applied off the shelf to this setting. These observations underscore the need for task-specific evaluation setups and model development tailored to scenarios with high lexical richness.

6.2.2. ROBUST USER SATISFACTION ESTIMATION IN TASK-ORIENTED DIALOGUE SYSTEMS (CHAPTER 3)

In Chapter 3, we demonstrated the potential of LLMs to generate high-quality counterfactual dialogue samples, which we used to augment the current benchmarks with a more balanced distribution of satisfactory and dissatisfactory dialogue samples. However, the focus of our study was on the generation of evaluation test samples, and we did not explore how adding the generated dialogues to the training sets would affect the performance of user satisfaction estimators. As such, augmenting the training data for user satisfaction estimators in task-oriented dialogue (TOD) systems is an important direction that needs to be explored in future studies.

Additionally, Chapter 3 exclusively focused on turn-level satisfaction estimation, we recognize the importance of dialogue-level satisfaction estimation which requires more advanced methods. In the meantime, we acknowledge that generating dialogue-level counterfactuals may require more complex methods. Lastly, the scope of our work in Chapter 3 was limited to task-oriented dialogue systems, whereas user satisfaction estimation has also been explored in other domains, such as conversational recommender systems [164]. One possible direction to extend our counterfactual dialogue generation approach is to broader applications of satisfaction estimation in various dialogue system settings.

6.2.3. MEASURING SOCIETAL BIAS IN RANKED LISTS OF DOCUMENTS (CHAPTER 4)

In Chapter 4, we studied societal bias in a ranked list documents with a particular focus on gender representation in ranked lists of documents using term-based group representations. Evaluating bias with term-based group representations, however, has clear limitations compared to real-world user evaluations. Depite this, such evaluation is still useful given the importance of societal fairness and the risks of unfair ranking systems. Future work should look into more semantic approaches that better match user perceptions. Our current method using counterfactual data substitution may also miss some learned gender biases, since some of such association of terms to societal groups often exist along a spectrum in models. Additionally, our Counterfactually-estimated Rank-biased Overlap (CRBO) estimation is currently based on the divergence between results from the original collection and a single counterfactual collection. Future research could explore more stratified counterfactual collection setups (instead of a single counterfactual collection) to better capture nuanced bias patterns.

6.2.4. ATTRIBUTION SENSITIVITY AND BIAS IN RAG (CHAPTER 5)

In Chapter 5, we explored attribution sensitivity and bias in retrieval-augmented generation (RAG) systems. In that study, we examined only human versus AI authorship as the metadata of source documents. However, the proposed systematic evaluation approach can also be applied to assess sensitivity and bias toward other metadata attributes, such as the author's gender or race, or even the source from

which a document originates. In addition, the methodology could be incorporated into existing LLM trustworthiness benchmarks. The framework is flexible with respect to attribution quality metrics, meaning that measures other than precision and recall can be used in our proposed equations for quantifying attribution sensitivity and bias.

There are also limitations to this research. We do not propose or assess methodologies for mitigating the identified attribution bias; rather, our focus is on revealing the brittleness of LLMs when used in attributive retrieval-augmented generation. Our experiments were conducted with three LLMs, two of which are open-source and one closed-source. Applying the same sensitivity and bias analysis to a broader range of models is of interest for future work. Additionally, in our experimental setup, we used queries where there was only one relevant document that contains the ground-truth answer in the top-k retrieved documents. While this design supports more precise attribution traceability, it limits the ability to measure fine-grained attribution contributions from multiple relevant sources. Exploring more semantic evaluation of attribution in generated answers is a promising direction for future work. Finally, the scope of our evaluation was restricted to English-language datasets and prompts. An obvious next step would be to extend the analysis to other languages. In particular, it would be valuable to investigate whether similar biases exist across other languages in LLMs.

6.2.5. FINAL THOUGHTS: TOWARDS EVALUATING AGENTIC SYSTEMS

Recently, the design and implementation of agentic solutions have gained popularity as LLMs have shown to perform well when being employed as decision making end points [85, 185]. At their core, these solutions delegate decision-making to several specialized LLMs, granting them agency in determining the next action. Applications of agentic solutions cover a broad range, from tool calling [160] to agentic retrieval-augmented generation [48].

However, LLMs have been also shown to be prone to errors in their decision-making processes [111]. This susceptibility has reached a point where implementing guardrails for the actions and decisions made by LLMs has become a necessary and integral component of agentic systems in practice.

Consequently, each new deployment of agentic solutions calls for the robust evaluation of their performance. Robust evaluation should address a broad spectrum of factors, from the accuracy of agents in selecting actions (i.e., making decisions) to beyond-accuracy considerations such as their reliability, fairness and trustworthiness [58]. Our line of research in this thesis can pave the way for designing proper evaluation frameworks for measuring the reliability of agentic systems. Specifically, our perspective on designing evaluation setups and exploring how a system works in what-if scenarios can help and inspire future work on developing task-specific experimental setups and/or evaluation metrics for agentic systems. More precisely, the use of counterfactual thinking in this thesis (the systematic exploration of “what if” scenarios, by considering alternative inputs and conditions) can inspire future research on ensuring the comprehensiveness and generalizability of both agentic systems and their evaluation.

BIBLIOGRAPHY

[1] A. Abolghasemi, A. Askari, and S. Verberne. “On the Interpolation of Contextualized Term-based Ranking with BM25 for Query-by-Example Retrieval”. In: *Proceedings of the 2022 ACM SIGIR International Conference on Theory of Information Retrieval*. 2022, pp. 161–170.

[2] A. Abolghasemi, L. Azzopardi, A. Askari, M. de Rijke, and S. Verberne. “Measuring Bias in a Ranked List Using Term-Based Representations”. In: *European Conference on Information Retrieval*. Springer. 2024, pp. 3–19.

[3] A. Abolghasemi, L. Azzopardi, S. H. Hashemi, M. de Rijke, and S. Verberne. “PAttriEval: A Python Library for the Evaluation of Attribution in Retrieval-Augmented Large Language Models”. In: *R3AG: The First Workshop on Refined and Reliable Retrieval Augmented Generation*. ACM, Dec. 2024.

[4] A. Abolghasemi, L. Azzopardi, S. H. Hashemi, M. de Rijke, and S. Verberne. “Evaluation of Attribution Bias in Generator-Aware Retrieval-Augmented Large Language Models”. In: *Findings of the Association for Computational Linguistics: ACL 2025*. Vienna, Austria: Association for Computational Linguistics, July 2025, pp. 21105–21124.

[5] A. Abolghasemi, Z. Ren, A. Askari, M. Aliannejadi, M. Rijke, and S. Verberne. “CAUSE: Counterfactual Assessment of User Satisfaction Estimation in Task-Oriented Dialogue Systems”. In: *Findings of the Association for Computational Linguistics ACL 2024*. Bangkok, Thailand and virtual meeting: Association for Computational Linguistics, Aug. 2024, pp. 14623–14635.

[6] A. Abolghasemi, S. Verberne, A. Askari, and L. Azzopardi. “Retrievability Bias Estimation Using Synthetically Generated Queries”. In: *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. 2023, pp. 3712–3716.

[7] A. Abolghasemi, S. Verberne, and L. Azzopardi. “Improving BERT-based Query-by-Document Retrieval with Multi-Task Optimization”. In: *Advances in Information Retrieval, 44th European Conference on IR Research, ECIR 2022*. 2022.

[8] A. Abolghasemi, S. Verberne, L. Azzopardi, and M. de Rijke. “On the Explainability of Exposing Query Identification”. In: *6th FAccTRec Workshop on Responsible Recommendation at RecSys*. 2023.

[9] A. Asai, Z. Wu, Y. Wang, A. Sil, and H. Hajishirzi. “Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection”. In: *The Twelfth International Conference on Learning Representations*. 2023.

[10] A. Askari and S. Verberne. “Combining Lexical and Neural Retrieval with Longformer-Based Summarization for Effective Case Law Retrieval”. In: *Proceedings of the Second International Conference on Design of Experimental Search & Information Retrieval Systems*. CEUR. 2021, pp. 162–170.

[11] A. Askari, A. Abolghasemi, G. Pasi, W. Kraaij, and S. Verberne. “Injecting the BM25 Score as Text Improves BERT-Based Re-rankers”. In: *Advances in Information Retrieval*. Ed. by J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, and A. Caputo. Cham: Springer Nature Switzerland, 2023, pp. 66–83.

[12] A. Askari, M. Aliannejadi, A. Abolghasemi, E. Kanoulas, and S. Verberne. “CLosER: Conversational Legal Longformer with Expertise-Aware Passage Response Ranker for Long Contexts”. In: *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. CIKM ’23. Birmingham, United Kingdom: Association for Computing Machinery, 2023, pp. 25–35.

[13] A. Askari, R. Petcu, C. Meng, M. Aliannejadi, A. Abolghasemi, E. Kanoulas, and S. Verberne. “Self-Seeding and Multi-Intent Self-Instructing LLMs for Generating Intent-Aware Information-Seeking Dialogs”. In: *arXiv preprint arXiv:2402.11633* (2024).

[14] A. Askari, R. Petcu, C. Meng, M. Aliannejadi, A. Abolghasemi, E. Kanoulas, and S. Verberne. “SOLID: Self-seeding and Multi-intent Self-instructing LLMs for Generating Intent-aware Information-Seeking Dialogs”. In: *Findings of the Association for Computational Linguistics: NAACL 2025*. Ed. by L. Chiruzzo, A. Ritter, and L. Wang. Albuquerque, New Mexico: Association for Computational Linguistics, Apr. 2025, pp. 6375–6395.

[15] A. Askari, S. Verberne, A. Abolghasemi, W. Kraaij, and G. Pasi. “Retrieval for Extremely Long Queries and Documents with RPRS: A Highly Efficient and Effective Transformer-Based Re-Ranker”. In: *ACM Transactions on Information Systems* 42.5 (2024), pp. 1–32.

[16] I. Augenstein, T. Baldwin, M. Cha, T. Chakraborty, G. L. Ciampaglia, D. Corney, R. DiResta, E. Ferrara, S. Hale, A. Halevy, *et al.* “Factuality Challenges in the Era of Large Language Models”. In: *arXiv preprint arXiv:2310.05189* (2023).

[17] P. Bajaj, D. Campos, N. Craswell, L. Deng, J. Gao, X. Liu, R. Majumder, A. McNamara, B. Mitra, T. Nguyen, M. Rosenberg, X. Song, A. Stoica, S. Tiwary, and T. Wang. “MS MARCO: A Human Generated Machine Reading Comprehension Dataset”. In: *arXiv preprint arXiv:1611.09268* (2016).

[18] I. Beltagy, K. Lo, and A. Cohan. “SciBERT: A Pretrained Language Model for Scientific Text”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019, pp. 3615–3620.

[19] E. Ben-David, B. Carmeli, and A. Anaby-Tavor. “Improved Goal Oriented Dialogue via Utterance Generation and Look Ahead”. In: *arXiv preprint arXiv:2110.12412* (2021).

[20] A. Berger and J. Lafferty. “Information Retrieval as Statistical Translation”. In: *ACM SIGIR Forum*. Vol. 51. 2. ACM New York, NY, USA. 2017, pp. 219–226.

[21] A. J. Biega, K. P. Gummadi, and G. Weikum. “Equity of Attention: Amortizing Individual Fairness in Rankings”. In: *The 41st international ACM SIGIR Conference on Research & Development in Information Retrieval*. 2018, pp. 405–414.

[22] A. Bigdeli, N. Arabzadeh, S. Seyedsalehi, B. Mitra, M. Zihayat, and E. Bagheri. “De-biasing Relevance Judgements for Fair Ranking”. In: *Advances in Information Retrieval: 45th European Conference on Information Retrieval*. Springer. 2023, pp. 350–358.

[23] A. Bigdeli, N. Arabzadeh, S. Seyedsalehi, M. Zihayat, and E. Bagheri. “On the Orthogonality of Bias and Utility in Ad Hoc Retrieval”. In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2021, pp. 1748–1752.

[24] A. Bigdeli, N. Arabzadeh, S. Seyedsalehi, M. Zihayat, and E. Bagheri. “A Light-Weight Strategy for Restraining Gender Biases in Neural Rankers”. In: *European Conference on Information Retrieval*. Springer. 2022, pp. 47–55.

[25] P. K. Bodigutla, A. Tiwari, S. Matsoukas, J. Valls-Vargas, and L. Polymenakos. “Joint Turn and Dialogue level User Satisfaction Estimation on Multi-Domain Conversations”. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. 2020, pp. 3897–3909.

[26] B. Bohnet, V. Q. Tran, P. Verga, R. Aharoni, D. Andor, L. B. Soares, M. Ciaramita, J. Eisenstein, K. Ganchev, J. Herzig, *et al.* “Attributed question answering: Evaluation and modeling for attributed large language models”. In: *arXiv preprint arXiv:2212.08037* (2022).

[27] I. Bojic, J. Chen, S. Y. Chang, Q. C. Ong, S. Joty, and J. Car. “Hierarchical Evaluation Framework: Best Practices for Human Evaluation”. In: *Proceedings of the 3rd Workshop on Human Evaluation of NLP Systems*. 2023, pp. 11–22.

[28] L. Bonifacio, H. Abonizio, M. Fadaee, and R. Nogueira. “InPars: Data Augmentation for Information Retrieval Using Large Language Models”. In: *arXiv preprint arXiv:2202.05144* (2022).

[29] L. Bottou, J. Peters, J. Quiñonero-Candela, D. X. Charles, D. M. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson. “Counterfactual Reasoning and Learning Systems: The Example of Computational Advertising”. In: *Journal of Machine Learning Research* 14.11 (2013).

[30] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, *et al.* “Language Models Are Few-Shot Learners”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 1877–1901.

[31] W. Cai and L. Chen. “Predicting User Intents and Satisfaction with Dialogue-Based Conversational Recommendations”. In: *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*. 2020, pp. 33–42.

[32] N. Calderon, E. Ben-David, A. Feder, and R. Reichart. “DoCoGen: Domain Counterfactual Generation for Low Resource Domain Adaptation”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by S. Muresan, P. Nakov, and A. Villavicencio. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 7727–7746.

[33] J. Chen, X. Dong, W. Xie, R. Peng, K. Zeng, and T. Hao. “LLM-Enhanced Query Generation and Retrieval Preservation for Task-Oriented Dialogue”. In: *Findings of the Association for Computational Linguistics: ACL 2025*. 2025, pp. 14307–14321.

[34] P. Chen, X.-Y. Guo, Y.-F. Li, X. Zhang, and Z. Feng. “Mitigating Language Bias of LMMs in Social Intelligence Understanding with Virtual Counterfactual Calibration”. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 2024, pp. 1300–1310.

[35] X. Chen, B. He, H. Lin, X. Han, T. Wang, B. Cao, L. Sun, and Y. Sun. “Spiral of Silences: How is Large Language Model Killing Information Retrieval?–A Case Study on Open Domain Question Answering”. In: *arXiv preprint arXiv:2404.10496* (2024).

[36] Z. Cheng, M. Cao, M.-A. Rondeau, and J. C. Cheung. “Stochastic Chameleons: Irrelevant Context Hallucinations Reveal Class-Based (Mis)Generalization in LLMs”. In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar. Vienna, Austria: Association for Computational Linguistics, July 2025, pp. 30187–30214.

[37] C.-H. Chiang and H.-Y. Lee. “Can Large Language Models Be an Alternative to Human Evaluations?” In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2023, pp. 15607–15631.

[38] C. L. Clarke, A. Vtyurina, and M. D. Smucker. “Assessing Top-preferences”. In: *ACM Transactions on Information Systems (TOIS)* 39.3 (2021), pp. 1–21.

[39] A. Cohan, S. Feldman, I. Beltagy, D. Downey, and D. Weld. “SPECTER: Document-level Representation Learning using Citation-informed Transformers”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, July 2020, pp. 2270–2282.

[40] N. Craswell, B. Mitra, E. Yilmaz, and D. Campos. “Overview of the TREC 2020 deep learning track”. In: *Proceedings of the Twenty-Ninth Text REtrieval Conference. NIST Special Publication*. 2021.

[41] N. Craswell, B. Mitra, E. Yilmaz, D. Campos, E. M. Voorhees, and I. Soboroff. “TREC Deep Learning Track: Reusable Test Collections in the Large Data Regime”. In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2021, pp. 2369–2375.

[42] P. Czarnowska, Y. Vyas, and K. Shah. “Quantifying Social Biases in NLP: A Generalization and Empirical Comparison of Extrinsic Fairness Metrics”. In: *Transactions of the Association for Computational Linguistics* 9 (2021), pp. 1249–1267.

[43] S. Dai, Y. Zhou, L. Pang, W. Liu, X. Hu, Y. Liu, X. Zhang, G. Wang, and J. Xu. “Neural Retrievers Are Biased Towards LLM-Generated Content”. In: *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2024, pp. 526–537.

[44] Y. Deng, W. Zhang, W. Lam, H. Cheng, and H. Meng. “User Satisfaction Estimation with Sequential Dialogue Act Modeling in Goal-oriented Conversational Systems”. In: *Proceedings of the ACM Web Conference 2022*. 2022, pp. 2998–3008.

[45] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186.

[46] F. Diaz, B. Mitra, M. D. Ekstrand, A. J. Biega, and B. Carterette. “Evaluating Stochastic Rankings with Expected Exposure”. In: *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2020, pp. 275–284.

[47] H. Djeddal, P. Erbacher, R. Toukal, L. Soulier, K. Pinel-Sauvagnat, S. Katrenko, and L. Tamine. “An Evaluation Framework for Attributed Information Retrieval using Large Language Models”. In: *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM 2024)*. Boise, Idaho, USA: Association for Computing Machinery, 2024.

[48] G. Dong, J. Jin, X. Li, Y. Zhu, Z. Dou, and J.-R. Wen. “RAG-Critic: Leveraging Automated Critic-Guided Agentic Workflow for Retrieval Augmented Generation”. In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2025, pp. 3551–3578.

[49] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, *et al.* “The Llama 3 Herd of Models”. In: *arXiv preprint arXiv:2407.21783* (2024).

[50] M. D. Ekstrand, A. Das, R. Burke, and F. Diaz. “Fairness in Information Access Systems”. In: *Foundations and Trends in Information Retrieval* 16.1-2 (2022), pp. 1–177.

[51] M. D. Ekstrand, G. McDonald, A. Raj, and I. Johnson. “Overview of the TREC 2021 Fair Ranking Track”. In: *The Thirtieth Text REtrieval Conference (TREC 2021) Proceedings*. 2022.

[52] M. Eric, R. Goel, S. Paul, A. Sethi, S. Agarwal, S. Gao, A. Kumar, A. Goyal, P. Ku, and D. Hakkani-Tur. "MultiWOZ 2.1: A Consolidated Multi-Domain Dialogue Dataset with State Corrections and State Tracking Baselines". In: *Proceedings of the Twelfth Language Resources and Evaluation Conference*. 2020, pp. 422–428.

[53] D. Esiobu, X. Tan, S. Hosseini, M. Ung, Y. Zhang, J. Fernandes, J. Dwivedi-Yu, E. Presani, A. Williams, and E. Smith. "ROBBIE: Robust Bias Evaluation of Large Generative Language Models". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 2023, pp. 3764–3814.

[54] Y. Feng, G. Lampouras, and I. Iacobacci. "Topic-Aware Response Generation in Task-Oriented Dialogue with Unstructured Knowledge Access". In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. 2022, pp. 7199–7211.

[55] T. Formal, C. Lassance, B. Piwowarski, and S. Clinchant. "SPLADE v2: Sparse lexical and expansion model for information retrieval". In: *arXiv preprint arXiv:2109.10086* (2021).

[56] T. Formal, B. Piwowarski, and S. Clinchant. "SPLADE: Sparse Lexical and Expansion Model for First Stage Ranking". In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2021, pp. 2288–2292.

[57] A. Fujii, M. Iwayama, and N. Kando. "Overview of the Patent Retrieval Task at the NTCIR-6 Workshop." In: *NTCIR*. 2007.

[58] G. A. Gabison and R. P. Xian. "Inherent and Emergent Liability Issues in LLM-based Agentic Systems: a Principal-Agent Perspective". In: *Proceedings of the 1st Workshop for Research on Agent Language Models (REALM 2025)*. Ed. by E. Kamalloo, N. Gontier, X. H. Lu, N. Dziri, S. Murty, and A. Lacoste. Vienna, Austria: Association for Computational Linguistics, July 2025, pp. 109–130.

[59] L. Gao, Z. Dai, and J. Callan. "COIL: Revisit Exact Lexical Match in Information Retrieval with Contextualized Inverted List". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2021, pp. 3030–3042.

[60] R. Gao and C. Shah. "Toward Creating a Fairer Ranking in Search Engine Results". In: *Information Processing & Management* 57.1 (2020), p. 102138.

[61] T. Gao, H. Yen, J. Yu, and D. Chen. "Enabling Large Language Models to Generate Text with Citations". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by H. Bouamor, J. Pino, and K. Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 6465–6488.

[62] M. Gardner, Y. Artzi, V. Basmov, J. Berant, B. Bogin, S. Chen, P. Dasigi, D. Dua, Y. Elazar, A. Gottumukkala, *et al.* "Evaluating Models' Local Decision Boundaries via Contrast Sets". In: *Findings of Empirical Methods in Natural Language Processing* (2020).

[63] S. Garg, V. Perot, N. Limtiaco, A. Taly, E. H. Chi, and A. Beutel. “Counterfactual Fairness in Text Classification through Robustness”. In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 2019, pp. 219–226.

[64] A. Ghosh, R. Dutt, and C. Wilson. “When Fair Ranking Meets Uncertain Inference”. In: *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*. 2021, pp. 1033–1043.

[65] S. Goldfarb-Tarrant, A. Lopez, R. Blanco, and D. Marcheggiani. “Bias Beyond English: Counterfactual Tests for Bias in Sentiment Analysis in Four Languages”. In: *Findings of the Association for Computational Linguistics: ACL 2023*. 2023, pp. 4458–4468.

[66] R. Han, Y. Zhang, P. Qi, Y. Xu, J. Wang, L. Liu, W. Y. Wang, B. Min, and V. Castelli. “RAG-QA Arena: Evaluating Domain Robustness for Long-form Retrieval Augmented Question Answering”. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 2024, pp. 4354–4374.

[67] M. Heuss, D. Cohen, M. Mansoury, M. de Rijke, and C. Eickhoff. “Predictive Uncertainty-Based Bias Mitigation in Ranking”. In: *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. CIKM ’23. Birmingham, United Kingdom, 2023, pp. 762–772.

[68] M. Heuss, F. Sarvi, and M. de Rijke. “Fairness of Exposure in Light of Incomplete Exposure Estimation”. In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2022, pp. 759–769.

[69] D. Hiemstra. “A Linguistically Motivated Probabilistic Model of Information Retrieval”. In: *International Conference on Theory and Practice of Digital Libraries*. Springer. 1998, pp. 569–584.

[70] S. Hofstätter, S. Althammer, M. Schröder, M. Sertkan, and A. Hanbury. “Improving Efficient Neural Ranking Models with Cross-architecture Knowledge Distillation”. In: *arXiv preprint arXiv:2010.02666* (2020).

[71] S. Hofstätter, S.-C. Lin, J.-H. Yang, J. Lin, and A. Hanbury. “Efficiently Teaching an Effective Dense Retriever with Balanced topic Aware Sampling”. In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2021, pp. 113–122.

[72] G. Hong, A. P. Gema, R. Saxena, X. Du, P. Nie, Y. Zhao, L. Perez-Beltrachini, M. Ryabinin, X. He, C. Fourrier, *et al.* “The Hallucinations Leaderboard–An Open Effort to Measure Hallucinations in Large Language Models”. In: *arXiv preprint arXiv:2404.05904* (2024).

[73] P. Howard, A. Madasu, T. Le, G. L. Moreno, A. Bhiwandiwalla, and V. Lal. “SocialCounterfactuals: Probing and Mitigating Intersectional Social Biases in Vision-Language Models with Counterfactual Examples”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 11975–11985.

[74] N. Hu, J. Chen, Y. Wu, G. Qi, S. Bi, T. Wu, and J. Z. Pan. “Benchmarking Large Language Models in Complex Question Answering Attribution Using Knowledge Graphs”. In: *arXiv preprint arXiv:2401.14640* (2024).

[75] Z. Hu, Y. Feng, A. T. Luu, B. Hooi, and A. Lipani. “Unlocking the Potential of User Feedback: Leveraging Large Language Model as User Simulators to Enhance Dialogue System”. In: *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*. CIKM ’23. <conf-loc>, <city>Birmingham</city>, <country>United Kingdom</country>, </conf-loc>: Association for Computing Machinery, 2023, pp. 3953–3957.

[76] J. Huang and K. Chang. “Citation: A Key to Building Responsible and Accountable Large Language Models”. In: *Findings of the Association for Computational Linguistics: NAACL 2024*. 2024, pp. 464–473.

[77] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, *et al.* “A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions”. In: *ACM Transactions on Information Systems* 43.2 (2025), pp. 1–55.

[78] P.-S. Huang, H. Zhang, R. Jiang, R. Stanforth, J. Welbl, J. Rae, V. Maini, D. Yogatama, and P. Kohli. “Reducing Sentiment Bias in Language Models via Counterfactual Evaluation”. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. 2020, pp. 65–83.

[79] Y. Huang, J. Feng, X. Wu, and X. Du. “Counterfactual Matters: Intrinsic Probing for Dialogue State Tracking”. In: *The First Workshop on Evaluations and Assessments of Neural Conversation Systems*. 2021, pp. 1–6.

[80] J. Hutter, D. Rau, M. Marx, and J. Kamps. “Lost But not Only in the Middle: Positional Bias in Retrieval Augmented Generation”. In: *European Conference on Information Retrieval*. Springer. 2025, pp. 247–261.

[81] A. Jain, P. Aggarwal, R. Sahay, C. Dong, and A. Saladi. “AutoEval-ToD: Automated Evaluation of Task-oriented Dialog Systems”. In: *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 2025, pp. 10133–10148.

[82] S. Jeong, J. Baek, S. Cho, S. J. Hwang, and J. C. Park. “Adaptive-RAG: Learning to Adapt Retrieval-Augmented Large Language Models through Question Complexity”. In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 2024, pp. 7029–7043.

[83] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung. “Survey of Hallucination in Natural Language Generation”. In: *ACM Computing Surveys* 55.12 (2023), pp. 1–38.

[84] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, *et al.* “Mistral 7B”. In: *arXiv preprint arXiv:2310.06825* (2023).

[85] S. Jiang, D. JU, A. Cohen, S. Mitts, A. Foss, J. T. Kao, X. Li, and Y. Tian. “Towards Full Delegation: Designing Ideal Agentic Behaviors for Travel Planning”. In: *arXiv preprint arXiv:2411.13904* (2024).

[86] X. Jiao, Y. Yin, L. Shang, X. Jiang, X. Chen, L. Li, F. Wang, and Q. Liu. “TinyBERT: Distilling BERT for Natural Language Understanding”. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. 2020, pp. 4163–4174.

[87] E. Kamalloo, A. Jafari, X. Zhang, N. Thakur, and J. Lin. “HAGRID: A Human-LLM Collaborative Dataset for Generative Information-Seeking with Attribution”. In: *arXiv preprint arXiv:2307.16883* (2023).

[88] J. Kamps, N. Kondylidis, and D. Rau. “Impact of Tokenization, Pretraining Task, and Transformer Depth on Text Ranking”. In: *TREC*. 2020.

[89] M. Kay, C. Matuszek, and S. A. Munson. “Unequal Representation and Gender Stereotypes in Image Search Results for Occupations”. In: *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 2015, pp. 3819–3828.

[90] M. Khalifa, D. Wadden, E. Strubell, H. Lee, L. Wang, I. Beltagy, and H. Peng. “Source-Aware Training Enables Knowledge Attribution in Language Models”. In: *ICLR 2024 Workshop on Secure and Trustworthy Large Language Models*.

[91] O. Khattab and M. Zaharia. “ColBERT: Efficient and effective passage search via contextualized late interaction over BERT”. In: *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*. 2020, pp. 39–48.

[92] T. E. Kim and A. Lipani. “A multi-task based neural model to simulate users in goal oriented dialogue systems”. In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2022, pp. 2115–2119.

[93] D. P. Kingma and J. Ba. “Adam: A Method for Stochastic Optimization”. In: *ICLR (Poster)*. 2015.

[94] A. Klasnja, N. Arabzadeh, M. Mehrvarz, and E. Bagheri. “On the Characteristics of Ranking-Based Gender Bias Measures”. In: *14th ACM Web Science Conference 2022*. 2022, pp. 245–249.

[95] T. Kwiatkowski, J. Palomaki, O. Redfield, M. Collins, A. Parikh, C. Alberti, D. Epstein, I. Polosukhin, J. Devlin, K. Lee, *et al.* “Natural Questions: A Benchmark for Question Answering Research”. In: *Transactions of the Association for Computational Linguistics* 7 (2019), pp. 453–466.

[96] M.-C. Lee, Q. Zhu, C. Mavromatis, Z. Han, S. Adeshina, V. N. Ioannidis, H. Rangwala, and C. Faloutsos. “HybGRAG: Hybrid Retrieval-Augmented Generation on Textual and Relational Knowledge Bases”. In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by W. Che, J. Nabende, E. Shutova, and

M. T. Pilehvar. Vienna, Austria: Association for Computational Linguistics, July 2025, pp. 879–893.

[97] M. Lee, S. An, and M.-S. Kim. "PlanRAG: A Plan-then-Retrieval Augmented Generation for Generative Large Language Models as Decision Makers". In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 2024, pp. 6537–6555.

[98] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, *et al.* "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 9459–9474.

[99] S. Li, S. Yavuz, K. Hashimoto, J. Li, T. Niu, N. Rajani, X. Yan, Y. Zhou, and C. Xiong. "CoCo: Controllable Counterfactuals for Evaluating Dialogue State Trackers". In: *International Conference on Learning Representations*. 2020.

[100] S. Li, S. Park, I. Lee, and O. Bastani. "TRAQ: Trustworthy Retrieval Augmented Question Answering via Conformal Prediction". In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 2024, pp. 3799–3821.

[101] X. Li, Y. Cao, L. Pan, Y. Ma, and A. Sun. "Towards Verifiable Generation: A Benchmark for Knowledge-aware Language Model Attribution". In: *Findings of the Association for Computational Linguistics ACL 2024*. Ed. by L.-W. Ku, A. Martins, and V. Srikumar. Bangkok, Thailand and virtual meeting: Association for Computational Linguistics, Aug. 2024, pp. 493–516.

[102] Y. Li, X. Yue, Z. Liao, and H. Sun. "AttributionBench: How Hard is Automatic Attribution Evaluation?" In: *Findings of the Association for Computational Linguistics ACL 2024*. Ed. by L.-W. Ku, A. Martins, and V. Srikumar. Bangkok, Thailand and virtual meeting: Association for Computational Linguistics, Aug. 2024, pp. 14919–14935.

[103] Y. Li, M. Xu, X. Miao, S. Zhou, and T. Qian. "Large Language Models as Counterfactual Generator: Strengths and Weaknesses". In: *arXiv preprint arXiv:2305.14791* (2023).

[104] Y. Li, X. Guo, J. Gao, G. Chen, X. Zhao, J. Zhang, Q. Liu, H. Wu, X. Yao, and X. Wei. "LLMs Trust Humans More, That'sa Problem! Unveiling and Mitigating the Authority Bias in Retrieval-Augmented Generation". In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2025, pp. 28844–28858.

[105] C.-Y. Lin. "ROUGE: A Package for Automatic Evaluation of Summaries". In: *Text Summarization Branches Out*. 2004, pp. 74–81.

[106] J. Lin and X. Ma. "A Few Brief Notes on DeepImpact, COIL, and a Conceptual Framework for Information Retrieval Techniques". In: *arXiv preprint arXiv:2106.14807* (2021).

[107] J. Lin, X. Ma, S.-C. Lin, J.-H. Yang, R. Pradeep, and R. Nogueira. “Pyserini: A Python Toolkit for Reproducible Information Retrieval Research with Sparse and Dense Representations”. In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2021, pp. 2356–2362.

[108] J. Lin, R. Nogueira, and A. Yates. “Pretrained Transformers for Text Ranking: BERT and Beyond”. In: *Synthesis Lectures on Human Language Technologies* 14.4 (2021), pp. 1–325.

[109] S.-C. Lin, J.-H. Yang, and J. Lin. “Distilling Dense Representations for Ranking using Tightly-coupled Teachers”. In: *arXiv preprint arXiv:2010.11386* (2020).

[110] K. Lu, P. Mardziel, F. Wu, P. Amancharla, and A. Datta. “Gender Bias in Neural Natural Language Processing”. In: *Logic, Language, and Security*. Springer, 2020, pp. 189–202.

[111] A. Maharana, D.-H. Lee, S. Tulyakov, M. Bansal, F. Barbieri, and Y. Fang. “Evaluating Very Long-Term Conversational Memory of LLM Agents”. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by L.-W. Ku, A. Martins, and V. Srikumar. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 13851–13870.

[112] C. Malaviya, S. Lee, S. Chen, E. Sieber, M. Yatskar, and D. Roth. “ExpertQA: Expert-Curated Questions and Attributed Answers”. In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 2024, pp. 3025–3045.

[113] A. Mallia, O. Khattab, T. Suel, and N. Tonellotto. “Learning Passage Impacts for Inverted Indexes”. In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2021, pp. 1723–1727.

[114] R. H. Maudslay, H. Gonen, R. Cotterell, and S. Teufel. “It’s All in the Name: Mitigating Gender Bias with Name-based Counterfactual Data Substitution”. In: *arXiv preprint arXiv:1909.00871* (2019).

[115] G. McDonald, C. Macdonald, and I. Ounis. “Search Results Diversification for Effective Fair Ranking in Academic Search”. In: *Information Retrieval Journal* 25.1 (2022), pp. 1–26.

[116] H. Mei and J. M. Eisner. “The Neural Hawkes Process: A Neurally Self-Modulating Multivariate Point Process”. In: *Advances in Neural Information Processing Systems* 30 (2017).

[117] J. Menick, M. Trebacz, V. Mikulik, J. Aslanides, F. Song, M. Chadwick, M. Glaese, S. Young, L. Campbell-Gillingham, G. Irving, *et al.* “Teaching Language Models to Support Answers with Verified Quotes”. In: *arXiv preprint arXiv:2203.11147* (2022).

[118] X. Miao, Y. Li, and T. Qian. “Generating Commonsense Counterfactuals for Stable Relation Extraction”. In: *The 2023 Conference on Empirical Methods in Natural Language Processing*. 2023.

[119] S. Min, M. Lewis, L. Zettlemoyer, and H. Hajishirzi. “MetaICL: Learning to Learn In Context”. In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2022, pp. 2791–2809.

[120] M. Morik, A. Singh, J. Hong, and T. Joachims. “Controlling Fairness and Bias in Dynamic Learning-to-rank”. In: *Proceedings of the 43rd international ACM SIGIR Conference on Research and Development in Information Retrieval*. 2020, pp. 429–438.

[121] B. Muller, J. Wieting, J. H. Clark, T. Kwiatkowski, S. Ruder, L. Soares, R. Aharoni, J. Herzig, and X. Wang. “Evaluating and Modeling Attribution for Cross-Lingual Question Answering”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 2023, pp. 144–157.

[122] S. Mysore, A. Cohan, and T. Hope. “Multi-Vector Models with Textual Guidance for Fine-Grained Scientific Document Similarity”. In: *arXiv preprint arXiv:2111.08366* (2021).

[123] S. Mysore, T. O’Gorman, A. McCallum, and H. Zamani. “CSFCube-A Test Collection of Computer Science Research Articles for Faceted Query by Example”. In: *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*. 2021.

[124] H. Nghiem, J. Prindle, J. Zhao, and H. Daumé Iii. “You Gotta be a Doctor, Lin: An Investigation of Name-Based Bias of Large Language Models in Employment Recommendations”. In: *arXiv preprint arXiv:2406.12232* (2024).

[125] T. Nguyen, M. Rosenberg, X. Song, J. Gao, S. Tiwary, R. Majumder, and L. Deng. “MS MARCO: A Human Generated Machine Reading Comprehension Dataset”. In: *CoCo@ NIPS*. 2016.

[126] R. Nogueira and K. Cho. “Passage Re-ranking with BERT”. In: *arXiv preprint arXiv:1901.04085* (2019).

[127] R. Nogueira, W. Yang, J. Lin, and K. Cho. “Document Expansion by Query Prediction”. In: *arXiv preprint arXiv:1904.08375* (2019).

[128] OpenAI. “GPT-4 Technical Report”. In: *OpenAI* (2023).

[129] J. Ouyang, T. Pan, M. Cheng, R. Yan, Y. Luo, J. Lin, and Q. Liu. “HoH: A Dynamic Benchmark for Evaluating the Impact of Outdated Information on Retrieval-Augmented Generation”. In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar. Vienna, Austria: Association for Computational Linguistics, July 2025, pp. 6036–6063.

[130] K. Ozeki, R. Ando, T. Morishita, H. Abe, K. Mineshima, and M. Okada. “Exploring Reasoning Biases in Large Language Models Through Syllogism: Insights from the NeuBAROCO Dataset”. In: *Findings of the Association for Computational Linguistics ACL 2024*. Ed. by L.-W. Ku, A. Martins, and V. Srikumar. Bangkok, Thailand and virtual meeting: Association for Computational Linguistics, Aug. 2024, pp. 16063–16077.

[131] Y. Pan, M. Ma, B. Pflugfelder, and G. Groh. “User Satisfaction Modeling with Domain Adaptation in Task-oriented Dialogue Systems”. In: *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 2022, pp. 630–636.

[132] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. “BLEU: A Method for Automatic Evaluation of Machine Translation”. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. 2002, pp. 311–318.

[133] N. Patel, S. Subramanian, S. Garg, P. Banerjee, and A. Misra. “Towards Improved Multi-Source Attribution for Long-Form Answer Generation”. In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 2024, pp. 3906–3919.

[134] J. Pearl. “Causal Inference in Statistics: An Overview”. In: *Statistics Surveys* 3 (2009), pp. 96–146.

[135] E. Perez, D. Kiela, and K. Cho. “True Few-Shot Learning with Language Models”. In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 11054–11070.

[136] F. Petroni, A. Piktus, A. Fan, P. Lewis, M. Yazdani, N. De Cao, J. Thorne, Y. Jernite, V. Karpukhin, J. Maillard, *et al.* “KILT: A Benchmark for Knowledge Intensive Language Tasks”. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2021, pp. 2523–2544.

[137] F. Piroi and A. Hanbury. “Multilingual Patent Text Retrieval Evaluation: CLEF–IP”. In: *Information Retrieval Evaluation in a Changing World*. Springer, 2019, pp. 365–387.

[138] F. Piroi, M. Lupu, A. Hanbury, and V. Zenz. “CLEF-IP 2011: Retrieval in the Intellectual Property Domain.” In: *CLEF (notebook papers/labs/workshop)*. Citeseer. 2011.

[139] J. M. Ponte and W. B. Croft. “A Language Modeling Approach to Information Retrieval”. In: *ACM SIGIR Forum*. Vol. 51. 2. ACM New York, NY, USA. 2017, pp. 202–208.

[140] A. Raj and M. D. Ekstrand. “Measuring Fairness in Ranked Results: An Analytical and Empirical Comparison”. In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2022, pp. 726–736.

[141] A. Raj, C. Wood, A. Montoly, and M. D. Ekstrand. “Comparing Fair Ranking Metrics”. In: *arXiv preprint arXiv:2009.01311* (2020).

[142] A. Rastogi, X. Zang, S. Sunkara, R. Gupta, and P. Khaitan. “Towards Scalable Multi-Domain Conversational Agents: The Schema-Guided Dialogue Dataset”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 05. 2020, pp. 8689–8696.

[143] D. Rau, H. Déjean, N. Chirkova, T. Formal, S. Wang, S. Clinchant, and V. Nikoulina. “BERGEN: A Benchmarking Library for Retrieval-Augmented Generation”. In: *Findings of the Association for Computational Linguistics: EMNLP 2024*. 2024, pp. 7640–7663.

[144] V. Rawte, S. Chakraborty, A. Pathak, A. Sarkar, S. T. I. Tonmoy, A. Chadha, A. Sheth, and A. Das. “The Troubling Emergence of Hallucination in Large Language Models-An Extensive Definition, Quantification, and Prescriptive Remediations”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 2023, pp. 2541–2573.

[145] N. Reimers and I. Gurevych. “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks”. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Nov. 2019.

[146] N. Rekabsaz, S. Kopeinik, and M. Schedl. “Societal Biases in Retrieved Contents: Measurement Framework and Adversarial Mitigation of BERT Rankers”. In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2021, pp. 306–316.

[147] N. Rekabsaz and M. Schedl. “Do Neural Ranking Models Intensify Gender Bias?” In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2020, pp. 2065–2068.

[148] S. E. Robertson, S. Walker, S. Jones, and M. Hancock-Beaulieu. “Okapi at TREC-3”. In: *Proceedings of the Third Text Retrieval Conference (TREC-3)*. 1995.

[149] S. E. Robertson and S. Walker. “Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval”. In: *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1994, pp. 232–241.

[150] G. M. Rosa, R. C. Rodrigues, R. Lotufo, and R. Nogueira. “Yes, BM25 Is a Strong Baseline for Legal Case Retrieval”. In: *arXiv preprint arXiv:2105.05686* (2021).

[151] J. J. Ross, E. Khramtsova, A. van der Vegt, B. Koopman, and G. Zuccon. “RARR Unraveled: Component-Level Insights into Hallucination Detection and Mitigation”. In: *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2025, pp. 3286–3295.

[152] C. Rus, J. Luppes, H. Oosterhuis, and G. H. Schoenmacker. “Closing the Gender Wage Gap: Adversarial Fairness in Job Recommendation”. In: *The 2nd Workshop on Recommender Systems for Human Resources, in conjunction with the 16th ACM Conference on Recommender Systems* (2022).

[153] P. Sapiezynski, W. Zeng, R. E. Robertson, A. Mislove, and C. Wilson. “Quantifying the Impact of User Attention on Fair Group Representation in Ranked Lists”. In: *Companion Proceedings of the 2019 World Wide Web Conference*. 2019, pp. 553–562.

[154] S. M. Sarwar and J. Allan. “Query by Example for Cross-Lingual Event Retrieval”. In: *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2020, pp. 1601–1604.

[155] P. Sen, X. Wang, R. Xu, and E. Yilmaz. “Task2KB: a public task-oriented knowledge base”. In: *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*. AAAI’23/IAAI’23/EAAI’23. AAAI Press, 2023.

[156] S. Seyedsalehi, A. Bigdeli, N. Arabzadeh, B. Mitra, M. Zihayat, and E. Bagheri. “Bias-aware Fair Neural Ranking for Addressing Stereotypical Gender Biases.” In: *EDBT*. 2022, pp. 2–435.

[157] C. Shah and R. W. White. “From To-Do to Ta-Da: Transforming Task-Focused IR with Generative AI”. In: *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2025, pp. 3911–3921.

[158] J. Shen, T. Zhou, Y. Chen, K. Liu, and J. Zhao. “CiteLab: Developing and Diagnosing LLM Citation Generation Workflows via the Human-LLM Interaction”. In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*. 2025, pp. 490–501.

[159] X. Shen, R. Blloshmi, D. Zhu, J. Pei, and W. Zhang. “Assessing “Implicit” Retrieval Robustness of Large Language Models”. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 2024, pp. 8988–9003.

[160] Z. Shi, S. Gao, L. Yan, Y. Feng, X. Chen, Z. Chen, D. Yin, S. Verberne, and Z. Ren. “Tool Learning in the Wild: Empowering Language Models as Automatic Tool Agents”. In: WWW ’25. Sydney NSW, Australia: Association for Computing Machinery, 2025, pp. 2222–2237.

[161] A. Singh and T. Joachims. “Fairness of Exposure in Rankings”. In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2018, pp. 2219–2228.

[162] S. Siriwardhana, R. Weerasekera, E. Wen, T. Kaluarachchi, R. Rana, and S. Nanayakkara. “Improving the Domain Adaptation of Retrieval Augmented Generation (RAG) Models for Open Domain Question Answering”. In: *Transactions of the Association for Computational Linguistics* 11 (2023), pp. 1–17.

[163] C. Siro, M. Aliannejadi, and M. de Rijke. “Understanding User Satisfaction with Task-Oriented Dialogue Systems”. In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2022, pp. 2018–2023.

[164] C. Siro, M. Aliannejadi, and M. de Rijke. “Understanding and Predicting User Satisfaction with Conversational Recommender Systems”. In: *ACM Transactions on Information Systems* 42.2 (2023), pp. 1–37.

[165] K. Song, Y. Kang, J. Liu, X. Li, C. Sun, and X. Liu. “A Speaker Turn-Aware Multi-Task Adversarial Network for Joint User Satisfaction Estimation and Sentiment Analysis”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 37. 11. 2023, pp. 13582–13590.

[166] M. Stadler, M. Bannert, and M. Sailer. “Cognitive Ease at a Cost: LLMs Reduce Mental Effort But Compromise Depth in Student Scientific Inquiry”. In: *Computers in Human Behavior* 160 (2024), p. 108386.

[167] A. Stolfo. “Groundedness in Retrieval-augmented Long-form Generation: An Empirical Study”. In: *Findings of the Association for Computational Linguistics: NAACL 2024*. 2024, pp. 1537–1552.

[168] Y. Sui, Y. He, Z. Ding, and B. Hooi. “Can Knowledge Graphs Make Large Language Models More Trustworthy? An Empirical Study Over Open-ended Question Answering”. In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar. Vienna, Austria: Association for Computational Linguistics, July 2025, pp. 12685–12701.

[169] E. Sulem, O. Abend, and A. Rappoport. “BLEU is Not Suitable for the Evaluation of Text Simplification”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Oct. 2018, pp. 738–744.

[170] W. Sun, S. Guo, S. Zhang, P. Ren, Z. Chen, M. de Rijke, and Z. Ren. “Metaphorical User Simulators for Evaluating Task-oriented Dialogue Systems”. In: *ACM Transactions on Information Systems* (2023).

[171] W. Sun, L. Yan, X. Ma, S. Wang, P. Ren, Z. Chen, D. Yin, and Z. Ren. “Is ChatGPT Good at Search? Investigating Large Language Models as Re-Ranking Agents”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by H. Bouamor, J. Pino, and K. Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 14918–14937.

[172] W. Sun, S. Zhang, K. Balog, Z. Ren, P. Ren, Z. Chen, and M. de Rijke. “Simulating User Satisfaction for the Evaluation of Task-Oriented Dialogue Systems”. In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2021, pp. 2499–2506.

[173] S. T.y.s.s and I. Chowdhury. “Fairness Beyond Performance: Revealing Reliability Disparities Across Groups in Legal NLP”. In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar. Vienna, Austria: Association for Computational Linguistics, July 2025, pp. 24376–24390.

[174] H. Tan, F. Sun, W. Yang, Y. Wang, Q. Cao, and X. Cheng. “Blinded by Generated Contexts: How Language Models Merge Generated and Retrieved Contexts for Open-Domain QA?” In: *arXiv preprint arXiv:2401.11911* (2024).

[175] L. Tunstall, E. Beeching, N. Lambert, N. Rajani, K. Rasul, Y. Belkada, S. Huang, L. von Werra, C. Fourrier, N. Habib, *et al.* “Zephyr: Direct distillation of lm alignment”. In: *arXiv preprint arXiv:2310.16944* (2023).

[176] B. Wang, W. Chen, H. Pei, C. Xie, M. Kang, C. Zhang, C. Xu, Z. Xiong, R. Dutta, R. Schaeffer, *et al.* “DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models”. In: *Advances in Neural Information Processing Systems* 36 (2024).

[177] S. Wang, S. Zhuang, and G. Zuccon. “BERT-Based Dense Retrievers Require Interpolation with BM25 for Effective Passage Retrieval”. In: *Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*. 2021, pp. 317–324.

[178] W. Wang, Z. Zhang, J. Guo, Y. Dai, B. Chen, and W. Luo. “Task-Oriented Dialogue System as Natural Language Generation”. In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2022, pp. 2698–2703.

[179] W. Wang, F. Wei, L. Dong, H. Bao, N. Yang, and M. Zhou. “MiniLM: Deep Delf-attention Distillation for Task-agnostic Compression of Pre-trained Transformers”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 5776–5788.

[180] Z. Wang and A. Culotta. “Robustness to Spurious Correlations in Text Classification via Automatically Generated Counterfactuals”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 16. 2021, pp. 14024–14031.

[181] Z. Wang, J. Araki, Z. Jiang, M. R. Parvez, and G. Neubig. “Learning to Filter Context for Retrieval-Augmented Generation”. In: *arXiv preprint arXiv:2311.08377* (2023).

[182] W. Webber, A. Moffat, and J. Zobel. “A Similarity Measure for Indefinite Rankings”. In: *ACM Transactions on Information Systems (TOIS)* 28.4 (2010), pp. 1–38.

[183] K. Webster, X. Wang, I. Tenney, A. Beutel, E. Pitler, E. Pavlick, J. Chen, E. Chi, and S. Petrov. “Measuring and Reducing Gendered Correlations in Pre-Trained Models”. In: *arXiv preprint arXiv:2010.06032* (2020).

[184] J. Wen, Y. Zhu, J. Zhang, J. Zhou, and M. Huang. “AutoCAD: Automatically Generate Counterfactuals for Mitigating Shortcut Learning”. In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. 2022, pp. 2302–2317.

[185] R. W. White. “Advancing the Search Frontier with AI Agents”. In: *Communications of the ACM* 67.9 (2024), pp. 54–65.

[186] H. Wu, B. Mitra, C. Ma, F. Diaz, and X. Liu. “Joint Multisided Exposure Fairness for Recommendation”. In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2022, pp. 703–714.

[187] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, *et al.* “Google’s Neural Machine Translation System: Bridging the Gap Between Human and Machine Translation”. In: *arXiv preprint arXiv:1609.08144* (2016).

[188] Y. Wu, L. Zhang, and X. Wu. “Counterfactual Fairness: Unidentification, Bound and Algorithm”. In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence*. 2019.

[189] J. Xie, K. Zhang, J. Chen, R. Lou, and Y. Su. “Adaptive Chameleon or Stubborn Sloth: Revealing the Behavior of Large Language Models in Knowledge Conflicts”. In: *The Twelfth International Conference on Learning Representations*.

[190] Z. Xie, V. Kocijan, T. Lukasiewicz, and O.-M. Camburu. “Counter-GAP: Counterfactual Bias Evaluation through Gendered Ambiguous Pronouns”. In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. 2023, pp. 3761–3773.

[191] G. Xiong, Q. Jin, Z. Lu, and A. Zhang. “Benchmarking Retrieval-Augmented Generation for Medicine”. In: *Findings of the Association for Computational Linguistics ACL 2024*. Ed. by L.-W. Ku, A. Martins, and V. Srikumar. Bangkok, Thailand and virtual meeting: Association for Computational Linguistics, Aug. 2024, pp. 6233–6251.

[192] L. Xiong, C. Xiong, Y. Li, K.-F. Tang, J. Liu, P. Bennett, J. Ahmed, and A. Overwijk. “Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval”. In: *arXiv preprint arXiv:2007.00808* (2020).

[193] K. Yang and J. Stoyanovich. “Measuring Fairness in Ranked Outputs”. In: *Proceedings of the 29th International Conference on Scientific and Statistical Database Management*. 2017, pp. 1–6.

[194] F. Ye, Z. Hu, and E. Yilmaz. “Modeling User Satisfaction Dynamics in Dialogue via Hawkes Process”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by A. Rogers, J. Boyd-Graber, and N. Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 8875–8889.

[195] J. Ye, Y. Wu, S. Gao, C. Huang, S. Li, G. Li, X. Fan, Q. Zhang, T. Gui, and X.-J. Huang. “RoTBench: A Multi-Level Benchmark for Evaluating the Robustness of Large Language Models in Tool Learning”. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 2024, pp. 313–333.

[196] X. Ye, R. Sun, S. Arik, and T. Pfister. “Effective Large Language Model Adaptation for Improved Grounding and Citation Generation”. In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 2024, pp. 6237–6251.

[197] Z. Yin, Q. Sun, Q. Guo, J. Wu, X. Qiu, and X. Huang. “Do Large Language Models Know What They Don’t Know?” In: *Findings of the Association for Computational Linguistics: ACL 2023*. Ed. by A. Rogers, J. Boyd-Graber, and N. Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 8653–8665.

[198] W. Yu, D. Iter, S. Wang, Y. Xu, M. Ju, S. Sanyal, C. Zhu, M. Zeng, and M. Jiang. “Generate rather than Retrieve: Large Language Models are Strong Context Generators”. In: *The Eleventh International Conference on Learning Representations*. 2022.

[199] X. Yue, B. Wang, Z. Chen, K. Zhang, Y. Su, and H. Sun. “Automatic Evaluation of Attribution by Large Language Models”. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. 2023, pp. 4615–4635.

[200] Z. Yue, H. Zeng, Y. Lu, L. Shang, Y. Zhang, and D. Wang. “Evidence-Driven Retrieval Augmented Response Generation for Online Misinformation”. In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 2024, pp. 5628–5643.

[201] M. Zehlike, F. Bonchi, C. Castillo, S. Hajian, M. Megahed, and R. Baeza-Yates. “Fa*ir: A Fair Top-k Ranking Algorithm”. In: *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 2017, pp. 1569–1578.

[202] M. Zehlike and C. Castillo. “Reducing Disparate Exposure in Ranking: A Learning to Rank Approach”. In: *Proceedings of the Web Conference 2020*. 2020, pp. 2849–2855.

[203] M. Zehlike, K. Yang, and J. Stoyanovich. “Fairness in Ranking, Part I: Score-based Ranking”. In: *ACM Computing Surveys* 55.6 (2022), pp. 1–36.

[204] W. Zeng, K. He, Y. Wang, C. Zeng, J. Wang, Y. Xian, and W. Xu. “FutureTOD: Teaching Future Knowledge to Pre-trained Language Model for Task-Oriented Dialogue”. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2023, pp. 6532–6546.

[205] G. Zerveas, N. Rekabsaz, D. Cohen, and C. Eickhoff. “Mitigating Bias in Search Results Through Contextual Document Reranking and Neutrality Regularization”. In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2022, pp. 2532–2538.

[206] C. Zhai and J. Lafferty. “A Study of Smoothing Methods for Language Models Applied to Information Retrieval”. In: *ACM Transactions on Information Systems (TOIS)* 22.2 (2004), pp. 179–214.

[207] M. Zhang, T. Qian, T. Zhang, and X. Miao. “Towards Model Robustness: Generating Contextual Counterfactuals for Entities in Relation Extraction”. In: *Proceedings of the ACM Web Conference 2023*. 2023, pp. 1832–1842.

[208] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. “BERTScore: Evaluating Text Generation with BERT”. In: *International Conference on Learning Representations*. 2019.

[209] Z. Zhao, E. Wallace, S. Feng, D. Klein, and S. Singh. “Calibrate Before Use: Improving Few-Shot Performance of Language Models”. In: *International Conference on Machine Learning*. PMLR. 2021, pp. 12697–12706.

[210] S. Zhuang and G. Zuccon. “Fast Passage Re-Ranking with Contextualized Exact Term Matching and Efficient Passage Expansion”. In: *arXiv preprint arXiv:2108.08513* (2021).

[211] S. Zhuang and G. Zuccon. “TILDE: Term Independent Likelihood Model for Passage Re-Ranking”. In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2021, pp. 1483–1492.

[212] C. Ziems, W. Held, J. Dwivedi-Yu, and D. Yang. “Measuring and Addressing Indexical Bias in Information Retrieval”. In: *Findings of the Association for Computational Linguistics ACL 2024*. Ed. by L.-W. Ku, A. Martins, and V. Srikumar. Bangkok, Thailand and virtual meeting: Association for Computational Linguistics, Aug. 2024, pp. 12860–12877.

[213] G. Zuccon, S. Zhuang, and X. Ma. “R2LLMs: Retrieval and Ranking with LLMs”. In: *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2025, pp. 4106–4109.