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.

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

Cover designed by Amin Abolghasemi

Contents

1 INTRODUCTION

Search and conversational systems have become integral to daily life, and have transformed how individuals access information, communicate, and perform tasks [157]. Search engines like Google and conversational AI systems, such as ChatGPT, Siri, and Alexa, facilitate seamless interactions by understanding and responding to natural language. These systems also provide instant access to knowledge, which enables fast decision-making and learning on an unprecedented scale: by reducing search costs and cognitive load through efficient retrieval and summarization, they allow users to focus on higher-level tasks rather than the mechanics of locating information in the search space [157, 166].

Previously, information systems mainly revolved around retrieval pipelines consisting of one or more ranking stages [108, 213]. The goal in these retrieval pipelines is to retrieve information that is relevant to a user query and addresses their information needs. These multi-stage pipelines can include different types of rankers, such as initial retrievers, re-rankers, or hybrid retrieval models [108]. Moreover, retrieval models vary widely, including sparse, dense, and learned sparse approaches, among others [70, 106, 109, 149]. The output of retrieval pipelines is typically presented to the user as a ranked list of items.

With the advent of large language models (LLMs), there has been a shift in information systems from purely retrieval-based towards generation-based approaches [82, 97, 98]. However, the knowledge encoded in LLMs is static and may quickly become outdated [129]. This may be attributed to various factors. For instance, certain information cannot be included in LLM training because of privacy restrictions, while other information may not yet exist at the time of training. Furthermore, LLMs are prone to generating plausible but factually incorrect output, commonly referred to as hallucinations [36, 72, 77, 168].

Retrieval-augmented generation (RAG) [98] has emerged as an effective solution to overcome the limitations of generation-only approaches, by enhancing factual accuracy, ensuring grounding in external knowledge, and maintaining up-to-date information access [96, 129, 168]. A RAG system typically consists of a stack that begins with a user query, followed by retrieval from a set of items (e.g., documents), which are then used as context for generating an answer to the given user query [96, 98].

However, like any other multi-stage system, each component of a RAG pipeline (whether retrieval or generation) can fail, making their evaluation both necessary and critical. Specifically, the widespread use of such multi-stage pipelines behind search and conversational systems raises challenges related to algorithmic bias and reliability in long-tail or unpredictable scenarios [104, 151, 173, 176]. Further, as these systems grow in sophistication and ubiquity, it becomes increasingly important to ensure their continuous and robust evaluation. Evaluation can span across multiple dimensions, including:

- **Generalizability and robustness of models to long-tail scenarios**: Settings in which search and conversational systems operate can be highly diverse and it may be difficult to anticipate low-frequent (long-tail) events during model development [66, 159, 195] for these systems. It is therefore important to scrutinize the performance of the models across as many potential scenarios as possible to ensure robust and reliable outcomes. For instance, long-tail retrieval scenarios pose challenges that ranking models are typically not designed to address. One such setting is retrieval with lexically rich queries, which often arise in the query-by-example retrieval task where documents themselves are used as queries. In Chapter 2, we investigate the generalizability of contextualized term-based ranking models, which combine the contextualization power of language models with the efficiency of lexical models. This evaluation highlights whether language-model-based retrievers can generalize effectively to scenarios where queries contain rich lexical information and provide abundant lexical signals about user intent.

  Additionally, in Chapter 3, we investigate the robustness of language models for user satisfaction estimation in task-oriented dialogue systems. Specifically, we examine the performance of user satisfaction estimators under evaluation settings with different distributions of satisfactory and dissatisfactory dialogue samples, a scenario that had not been previously explored. This evaluation highlights whether user satisfaction estimators in task-oriented dialogue systems can generalize to alternative evaluation settings applied to commonly-used benchmarks of this task.
- **Potential biases and trustworthiness of models**: Bias in information systems that rely on language models can arise from multiple sources [34, 124, 173, 176]. At training time, biases may be inherited from the data used for pre-training, fine-tuning, or post-training. At inference time, biases may also emerge from the data on which these models are applied. Moreover, such biases can manifest across diverse use cases of language models in the development of information systems. This thesis examines these issues in two contexts: retrieval (Chapter 4) and generation (Chapter 5). In both chapters, we propose evaluation metrics and methodologies for detecting and quantifying biases in information retrieval and generation systems, with a specific focus on retrieval and generation with language models. Studying the ways in which language models introduce and propagate bias is critical, as these biases can have downstream effects on real-world information systems [35, 174].

These evaluation scenarios, however, require devising task-specific experimental setups and/or evaluation metrics. To this aim, we use counterfactual thinking, which enables the systematic exploration of "what-if" scenarios. This perspective helps to ensure the comprehensivenesss and generalizability of both evaluation methods [2] and models [1, 4, 5] by covering hypothetical conditions. Specifically, we employ counterfactual evaluation which can be used to assess how a model's predictions change when a specific feature or set of features is altered while keeping everything else constant. By simulating these scenarios, we can evaluate and enhance the robustness of search and conversational systems by identifying potential brittleness in ranking and generative models utilized behind these systems.

In summary, this dissertation presents four interrelated studies (in Chapters 2 through 5) that examine how modern retrieval and generative models – particularly language models (LMs) – behave in nuanced, real-world information-seeking contexts. As mentioned above, our investigations span multiple areas, including attributive retrieval-augmented generation, bias/fairness in ranking, retrieval effectiveness in query-by-example settings, and robustness in user satisfaction estimation within task-oriented dialogue systems. Although each chapter addresses a distinct challenge, collectively, they contribute to a deeper understanding of, and improvements in, the robustness and fairness of AI systems under realistic and structurally challenging conditions. They also show how evaluation frameworks can be improved to better reflect the performance of retrieval and generative models in complex scenarios.

1.1. Research Outline and Questions

Each of the four research chapters in this dissertation addresses a specific aspect of our overall investigation. We outline each of these chapters in detail in the following.

1.1.1. Evaluating Contextualized Lexical Models in Query-by-Example Retrieval

In Chapter 2, we investigate the generalizability of contextualized term-based ranking to retrieval settings with lexically rich queries. Contextualized term-based ranking has been shown to bring the power of contextualization into the efficiency of lexical (term-based) ranking in ad hoc retrieval [210, 211]. However, having lexically rich queries in a retrieval setting means that there is an abundance of lexical relevance signals for a term-based ranking model such as BM25. As such, the generalizability of the added value of contextualization to retrieval settings with lexically rich queries remains unexplored. To study this generalizability, we evaluate the performance of two contextualized lexical ranking models (TILDE and TILDEv2) [210, 211] in query-by-example (QBE) retrieval tasks, where documents are used as queries to retrieve other similar documents [122, 123, 154]. This retrieval setting is common in domain-specific applications such as scientific literature search and legal case retrieval, where queries are substantially longer and more semantically complex than typical keyword-based queries. This chapter frames QBE retrieval as a distinct and underexplored lexically rich retrieval setup and highlights the generalizability of contextualized term-based ranking to this setup. In summary, in this chapter we address the following research question:

1

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

1.1.2. Robust User Satisfaction Estimation in Task-Oriented Dialogue Systems

In Chapter 3, we look into the evaluation of user satisfaction estimation (USE) in task-oriented dialogue (TOD) systems [33, 54, 81, 178]. USE is a critical task for ensuring high-quality and responsive conversational agents [44, 75, 172]. A key limitation in this area is the imbalance in existing evaluation datasets, which are heavily skewed toward satisfactory interactions (dialogues) between users and dialogue systems. This imbalance means that the impact of a more balanced set of satisfaction labels on the performance of USE models remains unknown. Put another way, it is not clear how robust and generalizable the performance of current user satisfaction estimators is to evaluation scenarios with more dissatisfactory dialogue samples. This type of robustness is particularly important, as its absence prevents the reliable detection of interactions in which users are dissatisfied. This is a capability essential for real-world deployment, especially in customer-facing or support-oriented applications. Therefore, we address the following research question:

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

To address this question, there is a need to balance the data with more dissatisfactory dialogue samples, which demands further dialogue collection and human annotation, which is a costly and time-consuming task. Therefore, to address **RQ2**, we first explore the use of counterfactual data augmentation as a strategy for enriching evaluation datasets with more dissatisfactory dialogues. By using large language models, we propose a framework for generating dialogue samples that reflect alternative user experiences (satisfactory versus dissatisfactory) while preserving the original task structure. This approach aims to support the creation of more balanced test collections, which enable a more accurate evaluation of user satisfaction estimation models.

This chapter outlines the limitations of current benchmarks, motivates the need for more representative dialogue samples, and presents a direction for augmenting the current benchmarks to be more representative of dialogue scenarios that could occur between a user and a dialogue system. In doing so, it contributes to making dialogue system evaluation more robust and reflective of real-world user behavior, particularly in capturing negative or dissatisfactory user experiences.

1.1.3. Measuring Societal Bias in Ranked Lists of Documents

In Chapter 4, we study societal bias in ranked lists of documents, with a particular focus on gender representation in ranked lists of documents [23, 146]. Document ranking models, often used in web search and other information retrieval systems, can reinforce or amplify existing societal inequalities when certain groups are systematically underrepresented or misrepresented in the retrieved results [50]. One prominent form of this issue is gender bias, where search results disproportionately favor content associated with one gender over another [24, 147, 205]. Prior work has introduced fairness metrics to evaluate the extent of such bias using term-based representations for different societal groups. The presence of group-representative terms in a document can be used to define the association of a document with a group, e.g., female-representative terms such as *she, her, mother* and male-representative terms such as *he, him, father*. Existing metrics often fall short in capturing nuanced representational disparities or in handling documents that do not explicitly reference any gender group [2], i.e., documents that do not include any group-representative terms. In this chapter, we investigate and propose a novel evaluation metric for more effective detection and measurement of representational bias in a ranked list of documents. More concretely, we study the following research question:

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

In Chapter 4, we study **RQ3** in the context of gender bias as a specific type of societal bias. There is limited understanding of how model-internal behavior, such as a system's sensitivity to gendered language, relates to the observed fairness of its output, i.e., retrieved rank list of documents. This chapter also explores how to distinguish between bias in retrieved ranked lists and bias in the underlying model behavior. We propose alternative perspectives for evaluating both the fairness of ranked results and a model's tendency to respond differently to subtle changes in identity-related language. The overall goal is to better understand and diagnose societal bias in document retrieval systems.

1.1.4. Attribution Sensitivity and Bias in RAG

In Chapter 5, we explore a key trust-related challenge in retrieval-augmented generation (RAG) systems [80, 98, 143]: the reliability of source attribution. In attributive RAG, large language models generate answers based on a set of retrieved documents while attributing (citing) these sources to support the tracking of answer provenance [61, 76, 158]. However, the extent to which these models faithfully attribute their responses to the appropriate input documents (and the factors that influence their attribution behavior) remains understudied. One important yet underexplored factor in attributive RAG is the effect of metadata associated with input documents, particularly authorship information, that is, details about who generated or wrote the document (e.g., whether it is AI-generated or human-authored web content). If attribution is influenced by (and thus is sensitive to) superficial cues like authorship labels rather than content relevance, it raises concerns about the trustworthiness, bias, and transparency of the generated output, especially in high-stakes domains such as law and education.

This chapter investigates how sensitive LLMs are to authorship metadata and whether they exhibit systematic preferences or biases in source attribution: whether there is any change in LLMs' attribution behavior (how they attribute their generated outputs to source documents) when they know who the authors (generators) of the source documents are. By focusing on this problem, we aim to better understand the conditions under which LLMs make attribution decisions. Moreover, we investigate how these decisions may introduce implicit biases into otherwise objective attributions (citations). More concretely, we answer the following research question:

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

1.2. CONTRIBUTIONS

This thesis makes the following methodological, empirical, and resource contributions:

1.2.1. METHODOLOGICAL CONTRIBUTIONS

- We introduce a data augmentation approach that uses LLMs to generate satisfaction-focused counterfactual dialogues in task-oriented dialogue (TOD) systems. We highlight that this approach can also serve as a systematic methodology for enhancing training data for user satisfaction estimation (Chapter 3).

- We propose Term Exposure-based Fairness (TExFAIR), an evaluation metric for measuring societal bias in ranked document lists. TExFAIR explicitly defines the association of each document to the groups based on a probabilistic term-level association (Chapter 4).

- We propose Counterfactually-estimated Attribution Bias (CAB) and Counterfactually-estimated Attribution Sensitivity (CAS), two evaluation metrics that can be used for measuring, respectively, the bias and the sensitivity of retrieval-augmented large language models toward information about who generated the source input documents (Chapter 5).

1.2.2. EMPIRICAL CONTRIBUTIONS

- We demonstrate that two contextualized lexical models (TILDE and TILDEv2) are less effective in Query-by-Example (QBE) retrieval than in ad hoc retrieval. We highlight that QBE is a lexically rich retrieval setting that is structurally different from other retrieval scenarios and requires special attention and dedicated methodological development (Chapter 2).

- We show that the relevance signals of contextualized term-based models can be complementary to those of BM25, as interpolating the methods leads to improvements in ranking effectiveness (Chapter 2).
- We demonstrate that adding information about who generated source documents (as metadata) to source documents may lead to statistically significant changes in the attribution quality of retrieval-augmented LLMs (Chapter 5).
- We uncover an attribution bias in LLMs toward explicit human authorship, providing a competing hypothesis to prior findings that suggested LLMs often prefer LLM-generated content over human-written content (Chapter 5).

1.2.3. Resource Contributions

- We provide augmented evaluation test collections (MWOZ and SGD) with counterfactual dialogue samples for user satisfaction estimation (Chapter 3).
- We provide AttriEval: an evaluation python library for assessing the performance of retrieval-augmented LLMs with respect to how they attribute their answers to the input source documents (Chapter 5).

1.3. Thesis Origins

Here, we list the publications that have been used as the basis for each chapter in this thesis.

Chapter 2 is based on the following paper:

- 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

  AA*[1]: Conceptualization, Investigation, Validation, Software, Methodology, Writing – Original Draft, Writing – Review & Editing. AA: Conceptualization, Methodology, Writing – Review & Editing. SV: Supervision, Conceptualization, Writing – Review & Editing, Funding Acquisition.

Chapter 3 is based on the following paper:

- 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

[1]We use AA* to refer to Amin Abolghasemi, as opposed to AA which refers to Arian Askari.

AA*: Conceptualization, Investigation, Resources, Data Curation, Validation, Software, Methodology, Writing – Original Draft, Writing – Review & Editing. ZR: Data Curation, Conceptualization, Writing – Review & Editing. AA: Conceptualization, Validation, Writing – Review & Editing. MA: Data Curation, Methodology, Writing – Review & Editing. MdR: Supervision, Conceptualization, Writing – Review & Editing. SV: Supervision, Conceptualization, Writing – Review & Editing, Funding Acquisition.

Chapter 4 is based on the following paper:

- 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

  AA*: Conceptualization, Investigation, Validation, Software, Methodology, Writing – Original Draft, Writing – Review & Editing. LA: Supervision, Methodology, Writing – Review & Editing, Funding Acquisition. AA: Conceptualization, Validation, Writing – Review & Editing. MdR: Supervision, Conceptualization, Writing – Review & Editing. SV: Supervision, Conceptualization, Writing – Review & Editing, Funding Acquisition.

Chapter 5 is based on the following paper:

- 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

  AA*: Conceptualization, Resources, Data Curation, Investigation, Validation, Software, Methodology, Writing – Original Draft, Writing – Review & Editing. LA: Supervision, Conceptualization, Writing – Review & Editing. SHH: Conceptualization, Validation, Writing – Review & Editing. MdR: Supervision, Conceptualization, Writing – Review & Editing. SV: Supervision, Conceptualization, Writing – Review & Editing, Funding Acquisition.

The writing of this thesis also benefited from work on the following publications:

- 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

- 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

- 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
- 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
- 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
- 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
- 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
- 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

2

CONTEXTUALIZED TERM-BASED RANKING

Term-based ranking with pre-trained transformer-based language models has recently gained attention as they bring the contextualization power of transformer models into the highly efficient term-based retrieval. In this chapter, we examine the generalizability of two of these deep contextualized term-based models in the context of query-by-example (QBE) retrieval in which a seed document acts as the query to find relevant documents. In this setting — where queries are much longer than common keyword queries — BERT inference at query time is problematic as it involves quadratic complexity. We investigate TILDE and TILDEv2, both of which leverage BERT tokenizer as their query encoder. With this approach, there is no need for BERT inference at query time, and also the query can be of any length. Our extensive evaluation on the four QBE tasks of SciDocs benchmark shows that in a query-by-example retrieval setting TILDE and TILDEv2 are still less effective than a cross-encoder BERT ranker. However, we observe that BM25 shows a competitive ranking quality compared to TILDE and TILDEv2, which contradicts the findings about the relative performance of these three models on retrieval for short queries reported in prior work. This result raises the question about the use of contextualized term-based ranking models being beneficial in QBE setting. Furthermore, we study the score interpolation between the relevance score from TILDE (TILDEv2) and BM25. We find that these two contextualized term-based ranking models capture different relevance signals than BM25 and combining the different term-based rankers results in statistically significant improvements in QBE retrieval. This chapter sheds light on the challenges of retrieval settings different from the common evaluation benchmarks. It could be of value as future work to study other contextualized term-based ranking models in QBE settings.

2.1. INTRODUCTION

Query-by-Example (QBE) retrieval is an Information Retrieval (IR) setting in which a seed document[1] acts as the query to represent the user's information need and the retrieval engine searches over a collection of the same type of documents [7, 122, 123, 154]. This retrieval setup is typical in professional, domain-specific tasks such as legal case law retrieval [7, 10], patent prior art search [57, 137, 138], and scientific literature search [7, 122, 123]. While using a document as a query could become challenging due to its length and complex semantic structure, prior work has shown that traditional term-based retrieval models like BM25 [148] are highly effective when used in QBE retrieval [7, 10, 150].

Recently, deep contextualized term-based retrieval models have gained attention as they bring the contextualization power of the pre-trained transformer-based language models into the highly efficient term-based retrieval. Examples of such models are DeepImpact [113], SPLADE [56], SPLADEv2 [55], TILDE [211], TILDEv2 [210], COIL [59], and uniCOIL [106]. Here, we specifically investigate TILDE, which is a term independent likelihood model, and its successor TILDEv2, which is a deep contextualized lexical exact matching model.

TILDE and TILDEv2, which are introduced as term-based re-ranking models, follow a recent paradigm in term-based retrieval where term importance is pre-computed with scalar term weights. Besides, to predict the relevance score, both of these models use the BERT tokenizer as their query encoder, which means that they do not need to perform any BERT inference at query time to encode the query. However, leveraging tokenizer-based encoding of the query trades off the query representation and therefore effectiveness with higher efficiency at inference time [210]. While the effectiveness of these models is evaluated on tasks and benchmarks where we have short queries, e.g., MSMARCO Passage Ranking [125] and the TREC DL Track [41], in this chapter, we evaluate them in the aforementioned QBE retrieval setting where queries are much longer than common keyword queries. In this regard, we address the following research questions:

Q1 How effective are TILDE and TILDEv2 in query-by-example retrieval?

A specific direction in answering *Q1* is to investigate the ranking quality of TILDE and TILDEv2 in comparison with the effective cross-encoder BERT ranker [7, 126], which is described in section 2.2.4. We are interested in this direction for two reasons. First, the cross-encoder BERT ranker exhibits quadratic complexity in both space and time with respect to the input length [108] and this is aggravated in QBE where we have long queries. TILDE and TILDEv2, however, do not need any BERT inference at query time. Second, due to the maximum input length of BERT, cross-encoder BERT ranker, which uses the concatenation of the query and the document, might not cover the whole query and document tokens in a QBE setting, whereas in TILDE and TILDEv2, the query can be of any length and documents are covered up to the maximum length of BERT.

Additionally, since TILDEv2 pre-computes the term weights only for those tokens existing in the documents, one risk is that it might aggravate the vocabulary

[1] Throughout this chapter, we use the term "document" to refer to a unit of retrieval [108].

mismatch problem. A typical approach to address this issue is to use document expansion methods. Zhuang and Zuccon (2021) use TILDE as their document expansion model for TILDEv2. We adopt that approach for our task and further investigate the impact of token-based document expansion with TILDE on the ranking quality of TILDEv2 in a QBE retrieval setting.

Apart from comparing TILDE and TILDEv2 to the cross-encoder BERT ranker, we also make a comparison to traditional lexical matching models (BM25 and Probabilistic Language models), which have been shown as strong baselines on QBE tasks in prior work [10, 150]:

Q2 What is the effectiveness of traditional lexical matching models with varying tokenization strategies in comparison to TILDE and TILDEv2?

To answer *Q2* we will investigate the effect of using the BERT tokenizer [45] as pre-processing for traditional term-based retrieval models. By doing so, we are aligning the index vocabulary of traditional models with that of TILDE and TILDEv2, which could make our comparison more fair.

We will see in the Section 2.4 that BM25 shows a competitive ranking quality in comparison to TILDE and TILDEv2 in our QBE benchmark. Because of the similar quality on average, we are interested to see if the relevance signals of TILDE and TILDEv2 are different from that of BM25, to find out if the methods are complementary to each other. To this aim, we will investigate the following research question:

Q3 To what extent do TILDE and TILDEv2 encode a different relevance signal from BM25?

To address the question above, as it is described in details in Section 2.3.3, we will analyze the effect of the interpolation of the scores of TILDE and TILDEv2 with BM25.

Since TILDE and TILDEv2 are introduced as re-ranking models, we use four different tasks from the SciDocs evaluation benchmark [39] as a domain-specific QBE benchmark. This benchmark uses scientific paper abstracts as the query and documents. The retrieval setting in these tasks suits as a re-ranking setup because of the number of documents to be ranked for each query. Since that we are working in a domain-specific evaluation setting, we will also address the following research question:

Q4 To what extent does a highly tailored domain-specific pre-trained BERT model affect the effectiveness of TILDE and TILDEv2 in comparison to a $\text{BERT}_{\text{base}}$ model?

In summary, our main contributions in this chapter are three-fold:

- We show that two recent transformer-based lexical models (TILDE and TILDEv2) are less effective in Query-by-Example retrieval than was expected based on results reported for ad hoc retrieval. This indicates that QBE retrieval is structurally different from other IR settings and requires special attention for methods development;

- We show that the relevance signals of TILDE and TILDEv2 can be complementary to that of BM25 as interpolation of the methods leads to an improvement in ranking effectiveness;
- We also investigate interpolations of BM25 with TILDE and TILDEv2 in an ideal setting where the optimal interpolation weight is known a priori, and by doing so, we show that more stratified approaches for the interpolation could result in higher gains from the interpolation of BM25 with TILDE and TILDEv2.

In section 2.2 we describe the retrieval models used in this chapter. In section 2.3 we provide details about our methods and experiments and in section 2.4 we analyze the results and discuss the answers to our research questions. Section 2.5 is dedicated to to further analysis of the results, and finally, in Section 2.6 we provide the conclusion.

2.2. BACKGROUND: RETRIEVAL MODELS

In this section, we briefly introduce the retrieval models that we implement and evaluate in our experiments.

2.2.1. TRADITIONAL LEXICAL MATCHING MODELS

BM25. For BM25 [148], we use the implementation by Elasticsearch[2] with the parameters $k = 2.75$, and $b = 1$, which was tuned over the validation set.

Probabilistic Language Models. For language modeling (LM) based retrieval [20, 69, 139], we use the built-in similarity functions of Elasticsearch for the implementation of language model with Jelinek Mercer (JM) smoothing [206].

2.2.2. TERM INDEPENDENT LIKELIHOOD MODEL: TILDE

TILDE is a tokenizer-based term-based retrieval model which follows a term independence assumption and formulates the likelihood of a query as follows:

$$\text{TILDE-QL}(q|d) = \sum_{i}^{|q|} log(P_\theta(q_i|d)), \tag{2.1}$$

in which q is the query, and d is the document. As Figure 2.1 shows, to compute the relevance score, the text of a document d is fed as the input for BERT and the log probability for each token is estimated by using a language modeling head on top of the BERT [CLS] token output. In other words, we are pre-computing the term weights over the complete BERT vocabulary. During both training and inference time, the query text is tokenized by using a BERT tokenizer and the resulting token IDs are used to look up the corresponding log probability from the likelihood distribution predicted in the output of the language modeling head. It is worth mentioning that the document likelihood can be computed in a similar way

[2] https://github.com/elastic/elasticsearch

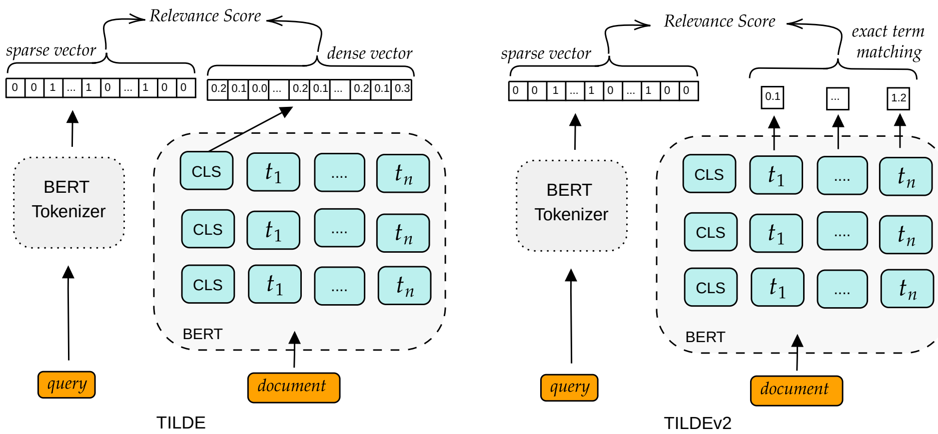

Figure 2.1: Model architectures. Left: TILDE [211]. Right: TILDEv2 [210]. Both TILDE and TILDEv2 leverage the BERT tokenizer as their query encoder. t_i stands for the ith token of the document. The $densevector$ and $sparsevector$ have the same length as the BERT vocabulary size.

by swapping the query and document; however, we only use the query likelihood (Equation 2.1) in our experiments.

For TILDE, we use the implementation from the authors' code repository.[3] We report results for the TILDE model with different initial checkpoints as the BERT encoder for our fine-tuning procedure. $\text{TILDE}_{\text{BERT}}$ uses bert-base-uncased, $\text{TILDE}_{\text{SciBERT}}$ uses SciBERT, and $\text{TILDE}_{\text{MSMARCO}}$ uses a TILDE which is already fine-tuned on MSMARCO; we use $\text{TILDE}_{\text{MSMARCO}}$ in a zero-shot setting on our data.

2.2.3. LEXICAL EXACT MATCHING: TILDEV2

TILDE has a drawback in which it expands each document to the size of the BERT tokenizer vocabulary. To tackle this problem, the authors proposed TILDEv2. TILDEv2, which builds upon uniCOIL [106] and TILDE, follows a recent paradigm in contextualized lexical exact matching in which BERT is used to output a scalar importance weight for document tokens [106, 210]. As it is shown in Figure 2.1, in TILDEv2, the token representation is downsized into a scalar weight and the relevance score between a query and a document pair is computed by a sum over the contextualized term weights for all terms appearing in both query and document:

$$s(q,d) = \sum_{q_i \in q} \max_{q_i = d_j} \left(c(q_i) \times v_j^d \right). \tag{2.2}$$

Here, q and d are the query and the document respectively; d_j is the jth token of the document; v_j^d is the term importance weight for the jth token of d , and $c(q_i)$ is the count of the i-th unique token which is achieved by using the BERT tokenizer

[3] https://github.com/ielab/TILDE

as the query encoder. In this equation, v_j^d is computed using the same method as in Lin and Ma (2021) in which a *RELU* function is used on the projection layer to force the model to map the token representations into a positive scalar weight:

$$v_j^d = ReLU(W_{proj}^{1\times n} BERT(d_j) + b), \tag{2.3}$$

in which d_j is the jth token in document d and b is the learnable bias parameter of the projection layer W_{proj}. Lin and Ma (2021) show that using a scalar weight as term importance (uniCOIL [106]) instead of a vector representation (COIL [59]) results in a decrease in the effectiveness; however, by using query expansion, uniCOIL can achieve higher effectiveness. Following the method proposed by Zhuang and Zuccon (2021) for query expansion with TILDE, we will show how TILDEv2 will act when we expand documents with TILDE. For TILDEv2, we use the implementation from the authors' code repository.[4]

2.2.4. CROSS-ENCODER BERT RANKER

The state-of-the-art results on SciDocs is reported by Abolghasemi, Verberne, and Azzopardi (2022) where they use a multi-task optimized cross-encoder BERT ranker [126]. The cross-encoder BERT ranker uses the concatenation of query and the document as the input to a BERT encoder. The BERT encoder is then followed by a projection layer W_{proj} on top of its $[CLS]$ token to compute the relevance score:

$$s(q,d)=BERT([CLS]q[SEP]d[SEP])_{[CLS]} * W_{proj}. \tag{2.4}$$

In this equation, q and d represent the query and the document respectively and $[CLS]$ as well as $[SEP]$ are special BERT tokens [45].

2.3. METHODS AND EXPERIMENTAL SETTINGS

In this section, we provide details and preliminaries about our methods and experimental settings.

2.3.1. EVALUATION BENCHMARK

We run our experiments on the SciDocs benchmark [39]. This dataset was originally introduced as a benchmark for representation learning tasks. Later, several works, including [7, 122] used the tasks of {co-view, co-read, citation, co-citation}-prediction from this benchmark as a query-by-example retrieval setting. As Figure 2.2 depicts, in this setting, given a query document, the goal is to retrieve and rank the most relevant documents out of a collection. The evaluation dataset for each of these four tasks includes approximately 30K total papers from a held-out pool of papers, consisting of 1K query papers and a candidate set of up to 5 positive papers and 25 negative papers [39].

To make our results comparable, we follow prior work on SciDocs to prepare the same training data [7]. To this aim, we take the validation set of each of tasks

[4] `https://github.com/ielab/TILDE/tree/main/TILDEv2`

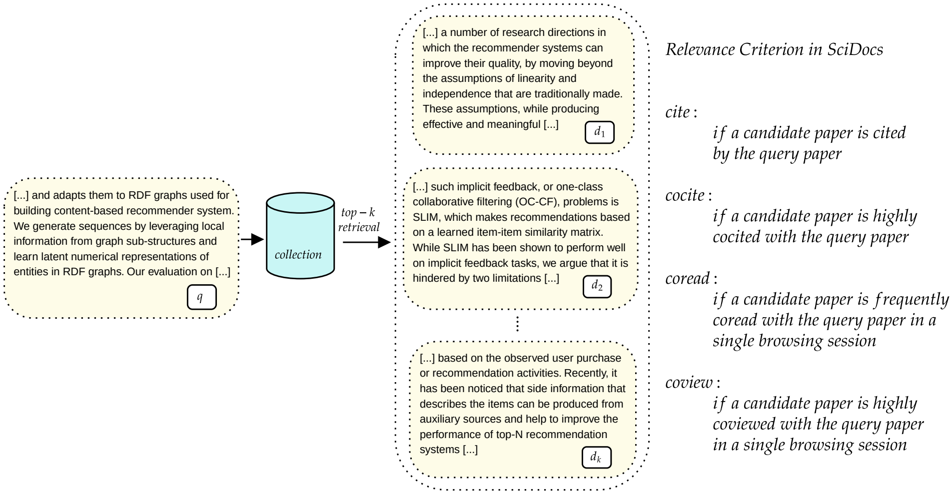

Figure 2.2: In the Query-by-Example retrieval setting, given a document (in its meaning as a unit of retrieval [108]) as the query q, the goal is to retrieve and rank the top-k relevant documents $\{d_1, d_2, ..., d_k\}$ out of a collection of documents. We use the four QBE tasks from SciDocs [39] benchmark including $\{cite, cocite, coread, coview\}$, each of which has its own relevance criterion [39].

and use 85% of them as training and 15% of them as the validation. Thus, each query in the train set has 5 relevant documents and 25 non-relevant documents. While TILDE is trained over relevant query-document pairs [211], TILDEv2 needs triplets in the format of (query, positive document, negative document). To prepare these triplets we pick two non-relevant documents per relevant document. By doing so, we create 10 triplets out of 30 training samples for each query. It should be noted that following Cohan, Feldman, Beltagy, Downey, and Weld (2020) we use a concatenation of abstract and title of the papers as documents.

2.3.2. BERT-BASED TOKENIZATION IN TRADITIONAL MODELS

In order to address *Q2*, we will examine the effects of transformer-based tokenizers as text pre-processor for traditional retrieval models. Doing so aligns the index vocabulary of traditional models with that of TILDE and TILDEv2, which in turn makes our comparison more fair. Transformers use different tokenization mechanisms e.g., WordPiece [187], which result in different query and document representations compared to common word-based tokenization approaches that are sometimes combined with normalization steps such as stemming and lemmatizing. Kamps, Kondylidis, and Rau (2020) show that using the BERT tokenizer as a pre-processor for BM25 results in a higher efficiency at the cost of a small decrease in effectiveness on the TREC 2020 Deep Learning Track [40]. QBE retrieval, however, has the challenge of long queries. In this chapter, we investigate whether the same effect applies to a QBE retrieval setting. To this aim, we use the $\text{BERT}_{\text{base}}$ tokenizer as a pre-processor for LM and BM25.

In addition, we use the SciBERT tokenizer, which is a domain-specific BERT tokenizer, to find out if a domain-specific tokenizer would have a different effect in comparison to the $\text{BERT}_{\text{base}}$ tokenizer. We use three different pre-processing setups in Elasticsearch to compare with our two transformer-based tokenizers:

- Elasticsearch Standard Analyzer (SA)
- Lowercase token filter, Porter Stemmer, Whitespace tokenizer (STM1)
- Lowercase token filter, Porter Stemmer, Standard tokenizer (STM2)

In Table 2.2, models corresponding to these setups respectively have SA, STM1, and STM2 as their subscript. BERT-Token and SciBERT-Token as subscripts stand for using BERT and SciBERT tokenizers as the text pre-processors.

2.3.3. Interpolation between BM25 and TILDE (TILDEv2) scores

To answer *Q3* about the difference between BM25 and TILDE (as well as TILDEv2) in terms of their relevance signals, following Wang, Zhuang, and Zuccon (2021), we evaluate the effect of the interpolation between the relevance scores from BM25 and from the contextualized term-based ranking models TILDE and TILDEv2. To this aim the interpolated score is computed as following:

$$s(q,d) = \alpha * s_{BM25}(q,d) + (1-\alpha) * s_{contextualized}(q,d). \tag{2.5}$$

Here, s_{BM25} stands for the BM25 score for query q, and document d, and $s_{contextualized}$ refers to the relevance score from TILDE or TILDEv2. Also, α is the hyperparameter that controls the impact of the scores from BM25 and TILDE (or TILDEv2). Prior to the interpolation both of the relevance scores are normalized using z-scaling (subtracting the mean and dividing by the standard deviation). We optimize α on the validation set.

Additionally, to further investigate the impact of interpolation, we do a per-query oracle interpolation in which we assume the best interpolation setting, i.e., optimal α, could be predicted per query, and thus we can explore how much effectiveness is reachable by the interpolation of the scores. In the remaining of the chapter, “oracle interpolation” refers to this latter interpolation setup and “non-oracle interpolation” refers to the vanilla interpolation, i.e., one α for all queries that is optimized on the validation set.

2.3.4. Document Expansion with TILDE

The Average token count of SciDocs documents (abstract+title) is 219 and 208 for BERT and SciBERT respectively. Their 90% token count quantiles are 341 and 385. Comparing these numbers to the maximum input length of BERT models, i.e., 512 tokens, we can see a capacity for the expansion of the documents. To further investigate *Q1*, following recent works which use document expansion to alleviate the vocabulary mismatch in contextualized term-based retrieval [106, 210], we evaluate the impact of retrieval on documents which are expanded at indexing time.

To this aim, we use TILDE in the same way as the original paper [210]. TILDE is at an advantage where it is more efficient than doc2query [127]. In this chapter, using $\text{TILDE}_{\text{SciBERT}}$ of which we found it performs the best compared to other TILDE models (Table 2.1), we generate $m = 200$, and $m = 300$ expansion terms for $\text{TILDEv2}_{\text{SciBERT}}$. It is noteworthy that similar to the original paper [210] not all expansion terms are added to a document, but only new expansion terms — that are not yet present in the document — are added.

2.3.5. DOMAIN-SPECIFIC BERT IN TILDE AND TILDEV2

To answer *Q1*, and *Q4*, we will investigate the power that can be brought by domain-specific pre-training to term-based ranking models. To do so, we evaluate the models' ranking quality in three settings: a) using $\text{BERT}_{\text{base}}$ as encoder, b) zero-shot utilization of TILDE and TILDEv2 models which are already fine-tuned on MSMARCO, and c) using a domain-specific pre-trained BERT as their encoder. Specifically, we use SciBERT [18] since our evaluation benchmark is from the scientific domain.

2.3.6. IMPLEMENTATION DETAILS

We run our experiments on NVIDIA RTX 3090 GPU machines with 24GB GPU memory. For $\text{BERT}_{\text{base}}$, and SciBERT we use the pre-trained models available on Huggingface. All BERT-based models are trained for 5 epochs. We use the Adam optimizer [93] with a learning rate of 2×10^{-5} for TILDE, and the AdamW optimizer with a learning rate of 5×10^{-6} for TILDEv2. In addition, we relax the maximum document length to the maximum input length of BERT during indexing.

2.4. RESULTS

Q1. *How effective are TILDE and TILDEv2 in query-by-example retrieval? and* ***Q4*** *To what extent does a highly tailored domain-specific pre-trained BERT model affect the effectiveness of TILDE and TILDEv2 in comparison to when we use a* $\textit{BERT}_{\textit{base}}$ *model?*

As Table 2.1 shows, TILDE and TILDEv2 are less effective than a cross-encoder BERT ranker in QBE retrieval despite having longer queries. This could be due to the fact that the cross-encoder BERT ranker applies all-to-all attention across tokens in both the query and the document [108] and thus, query terms and document terms are highly contextualized for the estimation of the relevance score. In addition, we see that $\text{TILDEv2}_{\text{BERT}}$ outperforms $\text{TILDE}_{\text{BERT}}$ despite TILDEv2 being highly prune to the vocabulary mismatch problem. One hypothesis for this observation could be that in a domain-specific retrieval setup like ours, TILDEv2 with the $\text{BERT}_{\text{base}}$ encoder predicts more effective document term weights than the term weights predicted for all tokens in the BERT vocabulary by TILDE with the $\text{BERT}_{\text{base}}$ encoder.

In addition, using SciBERT as our domain-specific pre-trained BERT model unsurprisingly improves the ranking quality of both TILDE and TILDEv2; however,

2

Model	Co-view		Co-read		Co-cite		Cite	
	MAP	nDCG	MAP	nDCG	MAP	nDCG	MAP	nDCG
a) BM25$_{\text{STM2}}$	80.8%$^{bcdf-j}$	0.9032^{cdfgj}	81.31%^{dfg}	0.9112^{cdg}	81.53%^{bcdfg}	0.9171^{cdfg}	79.74%^{bdg}	0.9085^{dg}
b) BM25$_{\text{SciBERT-Token}}$	80.08%^{cdfg}	0.8992^{cdg}	80.97%^{dfg}	0.9105^{dg}	80.83%^{cdfg}	0.9141^{cdg}	79.03%^{dg}	0.9051^{dg}
c) TILDE$_{\text{BERT}}$	76.74%^{d}	0.8761^{d}	80.57%^{dg}	0.8983^{d}	79.7%^{dg}	0.8999^{d}	82.15%^{abdg}	0.914^{dg}
d) TILDE$_{\text{MSMARCO}}$	68.22%	0.8261	66.75%	0.8206	65.21%	0.8145	65.29%	0.8186
e) TILDE$_{\text{SciBERT}}$	82.6%$^{a-df-j}$	0.9115^{bcdf-j}	85.03%$^{a-df-j}$	0.9256^{a-df-j}	86.38%$^{a-df-j}$	0.9375^{a-df-j}	87.74%$^{a-df-j}$	0.9431$^{a-dfghj}$
f) TILDEv2$_{\text{BERT}}$	79.17%^{cdg}	0.8948^{cd}	80.16%^{dg}	0.9051^{dg}	80.22%^{dg}	0.9103^{dg}	82.54%^{abdg}	0.9230^{abdg}
g) TILDEv2$_{\text{MSMARCO}}$	77.84%^{cd}	0.8876^{d}	78.53%^{d}	0.8959^{d}	78.17%^{d}	0.9006^{d}	75.62%^{d}	0.8866^{d}
h) TILDEv2$_{\text{SciBERT}}$	79.59%^{cdg}	0.8961^{cdg}	80.74%^{dg}	0.9080^{dg}	80.94%^{cdfg}	0.9123^{dg}	84.18%$^{a-dfg}$	0.9314^{a-dfg}
TILDEv2$_{\text{SciBERT}}$								
i) expansion w/ m=200	80.06%^{cdfgj}	0.8985^{cdg}	81.29%^{dfgh}	0.9096^{dg}	81.62%^{cdfg}	0.9153^{cdg}	86.42%$^{a-dfghj}$	0.9412$^{a-dfghj}$
j) expansion w/ m=300	79.38%^{cdg}	0.8942^{cd}	81.17%^{dfg}	0.9099^{dg}	81.93%^{bcdfgh}	0.9165^{cdg}	84.4%$^{a-dfg}$	0.9319^{a-dfg}
k) Cross-Encoder$_{\text{SciBERT}}$	85.2%$^{a-j}$	0.925^{a-j}	87.5%$^{a-j}$	0.940^{a-j}	89.7%$^{a-j}$	0.955^{a-j}	94.0%$^{a-j}$	0.975^{a-j}
l) Cross-Encoder$_{\text{MTFT-SciBERT}}$	**86.2%**$^{a-j}$	**0.930**$^{a-j}$	**87.7%**$^{a-j}$	**0.940**$^{a-j}$	**91.0%**$^{a-j}$	**0.961**$^{a-j}$	**94.2%**$^{a-j}$	**0.976**$^{a-j}$

Table 2.1: Ranking quality on the four SciDocs benchmark tasks using contextualized term-based ranking and cross-encoder BERT. “BERT” and “SciBERT” refers to the pre-trained model used as the encoder. “MSMARCO" indicates the utilization of TILDE or TILDEv2 which are already fine-tuned on MSMARCO. Rows *i* and *j* refer to the experiments on expanded documents with *m* terms using TILDE$_{\text{SciBERT}}$ as described in section 2.3.4. Statistical significance improvements are according to paired t-test (p<0.05) with Bonferroni correction for multiple testing. Rows *a* and *b* are included from Table 2.2 for ease of comparison.

Model	Co-view		Co-read		Co-cite		Cite	
	MAP	nDCG	MAP	nDCG	MAP	nDCG	MAP	nDCG
a) LM$_{\text{SA}}$	74.78%	0.8724	74.32%^{b}	0.8750	74.64%	0.8812	71.30%	0.8653
b) LM$_{\text{STM1}}$	74.82%	0.8737	73.51%	0.8694	74.60%	0.8810	70.98%	0.8636
c) LM$_{\text{STM2}}$	75.74%^{abde}	0.8786^{abe}	74.90%^{ab}	0.8771^{b}	75.80%^{abde}	0.8873^{ab}	72.15%^{abe}	0.8696^{b}
d) LM$_{\text{BERT-Token}}$	74.9%	0.8734	74.76%^{b}	0.8778^{b}	74.95%	0.8829	72.04%^{abe}	0.8694
e) LM$_{\text{SciBERT-Token}}$	74.74%	0.8717	74.69%^{b}	0.8771^{b}	74.81%	0.8827	71.46%	0.8666
f) BM25$_{\text{SA}}$	77.86%$^{a-e}$	0.8876^{a-e}	78.03%$^{a-e}$	0.8949^{a-e}	77.95%$^{a-e}$	0.8994^{a-e}	76.12%$^{a-e}$	0.8892^{a-e}
g) BM25$_{\text{STM1}}$	80.21%$^{a-f}$	0.9002^{a-f}	80.52%$^{a-f}$	0.9074^{a-f}	80.85%$^{a-f}$	0.9137^{a-f}	79.03%$^{a-f}$	0.9048^{a-f}
h) BM25$_{\text{STM2}}$	**80.8%**$^{a-gij}$	**0.9032**$^{a-gi}$	**81.31%**$^{a-gi}$	**0.9112**$^{a-g}$	**81.53%**$^{a-gij}$	**0.9171**$^{a-g}$	**79.74%**$^{a-gij}$	**0.9085**$^{a-g}$
i) BM25$_{\text{BERT-Token}}$	79.76%$^{a-f}$	0.8974^{a-f}	80.61%$^{a-f}$	0.9088^{a-f}	80.5%$^{a-f}$	0.9125^{a-f}	79.19^{a-f} %	0.9057^{a-f}
j) BM25$_{\text{SciBERT-Token}}$	80.08%$^{a-f}$	0.8992^{a-f}	80.97%$^{a-gi}$	0.9105^{a-f}	80.83%$^{a-fi}$	0.9141^{a-f}	79.03%$^{a-f}$	0.9051^{a-f}

Table 2.2: Ranking quality of traditional retrieval models on the four SciDocs benchmark tasks with different tokenization approaches. SA, STM1, STM2, BERT-Token, and SciBERT-Token refer to the pre-processing setting as described in section 2.3.2. Statistical significance improvements are according to paired t-test (p<0.05) with Bonferroni correction for multiple testing.

this improvement is higher between TILDE$_{\text{BERT}}$ and TILDE$_{\text{SciBERT}}$ than between TILDEv2$_{\text{BERT}}$ and TILDEv2$_{\text{SciBERT}}$ to an extent where TILDE$_{\text{SciBERT}}$ even outperforms

Model	Co-view		Co-read		Co-cite		Cite	
	MAP	nDCG	MAP	nDCG	MAP	nDCG	MAP	nDCG
a) $\text{BM25}_{\text{STM2}}$	80.8%^{c}	0.9032^{c}	81.31%	0.9112	81.53%	0.9171	79.74%	0.9085
b) $\text{TILDE}_{\text{SciBERT}}$	82.6%^{ac}	0.9115^{c}	85.03%^{ace}	0.9256^{ac}	86.38%^{ace}	0.9375^{ace}	87.74%^{ace}	0.9431^{ace}
c) $\text{TILDEv2}_{\text{SciBERT}}$	79.59%	0.8961	80.74%	0.9080	80.94%	0.9123	84.18%^{a}	0.9314^{a}
d) $\text{BM25}_{\text{STM2}}$ + $\text{TILDE}_{\text{SciBERT}}$	**85.29%**^{abce}	**0.9214**^{abce}	**86.52%**^{abce}	**0.9395**^{abce}	**88.32%**^{abce}	**0.9494**^{abce}	**88.46%**^{abce}	**0.9496**^{abce}
e) $\text{BM25}_{\text{STM2}}$ + $\text{TILDEv2}_{\text{SciBERT}}$	81.56%^{ac}	0.9032^{c}	82.63%^{ac}	0.9183^{ac}	83.06%^{ac}	0.9242^{ac}	84.18%^{a}	0.9318^{a}

Table 2.3: Results for non-oracle interpolation (the interpolation parameter α is optimized on the validation set) between BM25_{STM2}, $\text{TILDE}_{SciBERT}$, and $\text{TILDEv2}_{SciBERT}$. Statistical significance improvements are according to paired t-test ($p<0.05$) with Bonferroni correction for multiple testing. Rows *a*, *b*, and *c* are included from Table 2.1 for ease of comparison.

both $\text{TILDEv2}_{\text{BERT}}$ and $\text{TILDEv2}_{\text{SciBERT}}$. This observation could be due to the fact that the vocabulary mismatch problem caused by exact matching limits the TILDEv2 ranking quality, even if we use a highly tailored domain-specific BERT as its encoder. In this respect, we investigate the impact of token-based query expansion (see section 2.3.4) with TILDE on the ranking quality of TILDEv2 in our QBE retrieval setting. Lines *i*, and *j* in Table 2.1 are the ranking results on the documents that are expanded using TILDE with the method introduced by Zhuang and Zuccon (2021). Here, we are interested to find out if using document expansion is able to compensate for the gap in the ranking quality between $\text{TILDE}_{\text{SciBERT}}$, and $\text{TILDEv2}_{\text{SciBERT}}$.

As shown in Table 2.1, $\text{TILDEv2}_{\text{SciBERT}}$ with $m = \{200, 300\}$ expansion terms, is still less effective than $\text{TILDE}_{\text{SciBERT}}$. Furthermore, the table shows that the ranking quality of BM25_{STM2} on the original documents (line *a*) is still comparable with the ranking quality of $\text{TILDEv2}_{\text{SciBERT}}$ on the expanded documents (lines *i* and *j*). It is noteworthy that to make sure we are expanding the documents with enough tokens we investigate the average number of tokens added to the documents by the expansion with $\text{TILDE}_{\text{SciBERT}}$. By doing so, we find that for $m = 200$, and $m = 300$, approximately 49 and 128 new tokens are appended to the documents on average. Additionally, we find that using $m = 100$ results in roughly 2.6 new tokens on average. These numbers beside the statistics of the tokens in SciDocs benchmark, provided in Section 2.3.4, indicate that m should be tuned in order to take advantage from the document expansion with TILDE in QBE retrieval setting. Finally, we see that the zero-shot utilization of $\text{TILDE}_{\text{MSMARCO}}$ and $\text{TILDEv2}_{\text{MSMARCO}}$ does not show superior performance over the fine-tuned TILDE and TILDEv2 with both BERT and SciBERT encoders. It should be noted that taking models which are already fine-tuned on general domain (like $\text{TILDE}_{\text{MSMARCO}}$ and $\text{TILDEv2}_{\text{MSMARCO}}$) and further fine-tuning them on the task domain is a typical approach which could result in improvement in their ranking quality; however, we leave this item as a direction to be explored in future work.

Q2. *What is the effectiveness of traditional lexical matching models with varying*

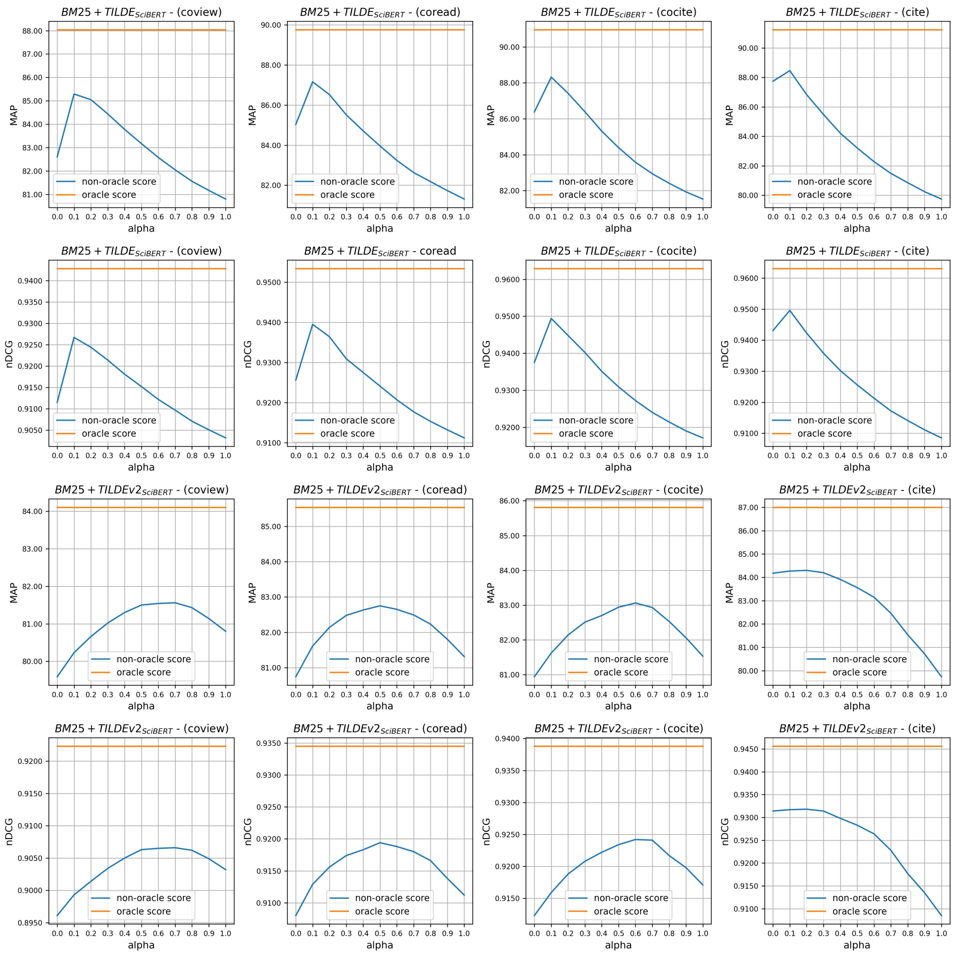

Figure 2.3: Results for TILDE and TILDEv2 with varying values of interpolation parameter α. The lines in blue and orange represent the effectiveness based on the non-oracle and oracle interpolations respectively. $\alpha = 0.0$ represents the TILDE- or TILDEv2-only setting; $\alpha = 1.0$ represents the BM25-only setting.

tokenization strategies in comparison to TILDE and TILDEv2

Table 2.2 shows that leveraging BERT and SciBERT tokenizers results in competitive ranking quality in both probabilistic language model based retrieval and BM25 in comparison to the three traditional pre-processing setups introduced in section 2.3.2.

Moreover, as the results of Table 2.1 shows, the ranking quality of BM25$_{STM2}$ not only outperforms LM and BM25 with different traditional and BERT-based pre-processing approaches, but also it could even outperform TILDE$_{BERT}$, and TILDEv2$_{BERT}$ in most of the tasks. In fact, we do not see a large gap between BM25 compared to TILDEv2 as was shown for retrieval based on short queries in the experiments on MSMARCO and TREC DL Track benchmarks [210]. This finding is important as (1) it sheds light on the challenges of retrieval settings different from the common evaluation benchmarks including MSMARCO and the TREC DL Track; (2) raises the question how effective other contextualized term-based ranking models would be in those settings.

Q3. *To what extent do TILDE and TILDEv2 encode a different relevance signal from BM25?*

The blue lines in Figure 2.3 show the ranking quality for $\text{TILDE}_{\text{SciBERT}}$ and $\text{TILDEv2}_{\text{SciBERT}}$ when their scores are interpolated with the BM25 score over varying values of interpolation parameter α with the step of 0.1. Besides, Table 2.3 shows the ranking quality for the interpolations with the α that is tuned over the validation set. We can see that an optimal interpolation between the scores from BM25 and the contextualized term-based ranking models TILDE and TILDEv2 could provide significant improvements for almost all tasks over the individual rankers participating in the interpolation. The only exceptions are in the *co-view*, and *cite* tasks. To be specific, there is no improvement over BM25 in the nDCG metric in the *co-view* (line e vs. line a in Table 2.3). Besides, in the *cite* task the improvement over TILDEv2 (line e vs. line c in Table 2.3) is not significant for the nDCG metric, and there is no improvement for the MAP metric. Nevertheless, the improvements obtained by the interpolation for almost all tasks and metrics indicates that TILDE and TILDEv2 are capturing different relevance signals compared to BM25.

To further investigate the impact of the score interpolation with BM25 scores, we perform an oracle interpolation in which we assume the optimal interpolation hyperparameter α is known for each individual query. This query-specific optimal value is selected over varying values of α with the step of 0.1. Table 2.4 as well as orange lines in Figure 2.3 show the results for the oracle interpolation. We can see that the oracle interpolation would result in a substantial improvement for both TILDE and TILDEv2.

Moreover, we can see in Table 2.4 that there is a subset of queries for which the BM25 ranking alone is better than the interpolation (queries with optimal α=1). This number is lower for the interpolation with TILDE than for the interpolation with TILDEv2. One hypothesis for this observation could be that the interpolation with TILDE is likely to be more helpful for BM25 since TILDE could bring more contextualization power for BM25 as it incorporates the term importance for all tokens in the query. In other words, since TILDEv2 pre-computes term weights only for the tokens of the document (whereas TILDE pre-computes the term importance weight for all the tokens in the BERT vocabulary per document), due to the chance of vocabulary mismatch in TILDEv2, it could incorporate less query-dependent contextualization than TILDE.

In addition, we see that the margin between the oracle interpolation results and both non-interpolated scores as well as non-oracle interpolation scores (Table 2.3) is substantial, which demonstrates that more complex aggregation methods could benefit more from the relevance signals from TILDE, TILDEv2 and BM25.

Model	Co-view		Co-read		Co-cite		Cite	
	MAP	**nDCG**	**MAP**	**nDCG**	**MAP**	**nDCG**	**MAP**	**nDCG**
BM25 $+_{oracle}$ TILDE$_{SciBERT}$								
ranking quality	88.04%^{ae}	0.9428^{ae}	89.76%^{ae}	0.9534^{ae}	90.96%^{ae}	0.9629^{ae}	91.24%^{ae}	0.9631^{ae}
α_{average}	0.1265	0.1294	0.1048	0.1053	0.1044	0.1048	0.0839	0.0857
#queries with optimal α=0	537	533	563	557	550	552	624	619
#queries with optimal α=1	19	21	6	5	14	13	4	4
IQR of the optimal α over queries	0.1	0.2	0.1	0.1	0.1	0.1	0.1	0.1
BM25 $+_{oracle}$ TILDEv2$_{SciBERT}$								
ranking quality	84.10%^{ah}	0.9223^{ah}	85.54 %^{ah}	0.9345^{ah}	85.81 %^{ah}	0.9388^{ah}	87.00 %^{ah}	0.9456^{ah}
α_{average}	0.3169	0.3205	0.3040	0.3048	0.3337	0.3339	0.2073	0.2083
#queries with optimal α=0	467	463	446	447	419	420	575	573
#queries with optimal α=1	84	84	71	72	92	94	33	33
IQR of the optimal α over queries	0.7	0.7	0.6	0.6	0.6	0.6	0.4	0.4

Table 2.4: Results for oracle interpolation (optimal α per query) between BM25$_{STM2}$, TILDE$_{SciBERT}$, and TILDEv2$_{SciBERT}$. Statistical significance with paired t-test (p<0.05) is reported only with respect to non-interpolated scores (α=0) of these three models in Table 2.1 (row a, e and h). α_{average} represents the mean of the optimal α values picked per query. The number of queries with optimal α=0 stands for the number of queries for which the interpolation does not improve their effectiveness compared to TILDE or TILDEv2 only.

2.5. DISCUSSION

In this section, we further analyze the interpolation between BM25 and TILDE (TILDEv2) in terms of the interpolation effectiveness and the interpolation weight α.

2.5.1. INTERPOLATION EFFECTIVENESS

The first two rows on the top of Figure 2.3 correspond to the interpolation between TILDE and BM25 and the two rows in the bottom correspond to the interpolation between TILDEv2 and BM25. Comparing the nDCG and MAP plots for the interpolation between TILDE and BM25, we can see that for this combination, α=0.1 shows the highest ranking quality for both nDCG and MAP metrics in all tasks. Thus, a high weight for TILDE with a small weight for BM25 gives the highest effectiveness for this combination. This observation could mean that while TILDE, as contextualized transformer-based model, is able to outperform BM25 as an exact matching model, it could still benefit from the strong lexical relevance scores from BM25.

On the other hand, for the combination of TILDEv2 and BM25 we see that the highest ranking quality is obtained with $\alpha \in \{0.3, 0.4, 0.5, 0.6, 0.7, 0.8\}$ depending on the task. The exceptions are in the *cite*, and *coview* tasks as described in the answer to *Q3* in Section 2.4. Thus, in the combination of TILDEv2 and BM25, an equal or slightly higher weight for BM25 relative to TILDEvs gives the optimal results. A hypothesis for this observation could be that while both BM25 and TILDEv2 are performing based on exact matching, the term weights from TILDEv2, which are predicted through contextualization of the document terms, are not always more effective than the term scores from BM25; however, they can act as a complement for each other and thus their interpolation could benefit from both.

2.5.2. INTERPOLATION WEIGHT

To further analyze the interpolation weight α, we consider the two aforementioned settings of oracle interpolation and non-oracle interpolation.

Non-oracle interpolation. We can see in Figure 2.3 (blue lines) that for the effective interpolations, i.e, the interpolations that result in higher effectiveness than each individual ranker included in the interpolation, the interpolation weight α in the combination of BM25 and TILDEv2 has a wider range than in the combination of BM25 and TILDE. This indicates that in this experimental setting the interpolation of BM25 and TILDEv2 could be achieved by a broader range of α values and is therefore more robust to the choice of interpolation weight than for BM25 and TILDE.

Oracle interpolation. As a measure of the statistical dispersion, we report the inter-quartile range (IQR) for the oracle interpolation weight α which is shown in Table 2.4. Taking the range of α [0.0, 1.0] into account, we can see that we have low inter-quartile range (IQR) for the optimal values of α per query in the interpolation with *TILDE* (top part of the table). On the other hand, the IQR for the optimal values of α per query for the interpolation with *TILDEv2* are much higher (bottom part of the table), which indicates that the optimal interpolation setting for the queries are more varied. This observation could give some sense of robustness against query variation for TILDE in comparison to TILDEv2 in this experimental setting. In other words, a query-dependent approach for optimizing α would be more robust against query variation for TILDE than for TILDEv2.

2.6. CONCLUSION

In this chapter we investigated the generalizability of two contextualized term-based ranking models TILDE and TILDEv2 for a QBE retrieval setting. In QBE, the queries are much longer than in ad-hoc retrieval, and efficient query processing is essential. We were specifically interested to see to what extent the relative performance of contextualized term-based ranking models in comparison to both traditional term-based models and the effective cross-encoder BERT ranker is generalizable to a QBE retrieval setting.

Our results show that similar to the original papers [210, 211], TILDE and TILDEv2 are less effective than a cross-encoder BERT ranker in QBE retrieval despite the context of longer queries. On the other hand, in the original papers, TILDE and TILDEv2 have shown superior ranking quality in comparison to BM25 as a traditional term-based retrieval model. We investigated if the same pattern exists in a query-by-example retrieval setting and our results show that BM25 has a competitive ranking quality compared to TILDE and TILDEv2. In fact, not only is it competitive, but also in some cases it could outperform TILDE and TILDEv2.

This finding is important as (1) it sheds light on the challenges of retrieval settings different from the common evaluation benchmarks including MSMARCO and the TREC DL Track; (2) raises the question how effective other contextualized term-based ranking models would be in those settings. Our results indicate that QBE retrieval is structurally different from other IR settings and requires special attention for methods development.

Furthermore, we investigated the impact of the interpolation between BM25 and TILDE as well as TILDEv2. By doing so, we find that a linear interpolation between the score of TILDE (TILDEv2) with that of BM25 leads to an improvement in the ranking effectiveness. This shows that the relevance signals from contextualized ranking models TILDE and TILDEv2 are complementary to the relevance signals from BM25. Additionally, through an analysis on the oracle interpolation between BM25 and TILDE (TILDEv2), we show that more stratified approaches could benefit more from the interpolation between the scores from these models.

3

CAUSE: Counterfactual Assessment of User Satisfaction Estimation in Task-Oriented Dialogue Systems

An important unexplored aspect in previous work on user satisfaction estimation for Task-Oriented Dialogue (TOD) systems is their evaluation in terms of robustness for the identification of user dissatisfaction: current benchmarks for user satisfaction estimation in TOD systems are highly skewed towards dialogues for which the user is satisfied. The effect of having a more balanced set of satisfaction labels on performance is unknown. However, balancing the data with more dissatisfactory dialogue samples requires further data collection and human annotation, which is costly and time-consuming. In this chapter, we leverage large language models (LLMs) and unlock their ability to generate satisfaction-focused counterfactual dialogues to augment the set of original dialogues of a test collection. We gather human annotations to ensure the reliability of the generated samples. We evaluate two open-source LLMs as user satisfaction estimators on our augmented collection against state-of-the-art fine-tuned models. Our experiments show that when used as few-shot user satisfaction estimators, open-source LLMs show higher robustness to the increase in the number of dissatisfaction labels in the test collection than the fine-tuned state-of-the-art models. Our results shed light on the need for data augmentation approaches for user satisfaction estimation in TOD systems. We release our aligned counterfactual dialogues, which are curated by human annotation, to facilitate further research on this topic.

3

3.1. Introduction

Task-oriented dialogue (TOD) systems help users complete specific tasks, e.g., booking a hotel or restaurant, through conversations [54, 155, 178, 204]. User

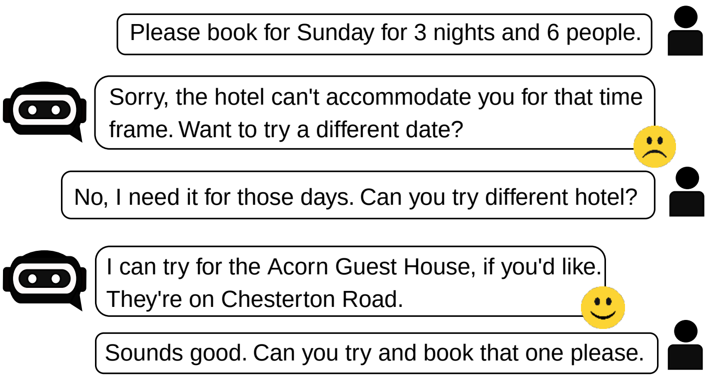

Figure 3.1: Example dialogue (snippet) between the user and the system from the MultiWOZ benchmark.

satisfaction estimation (USE) is a key task in TOD systems, aiming to measure the extent to which users are satisfied with the dialogue they are having with the system (see Figure 3.1). USE has various applications as it can be viewed as a continuous approximation of human feedback for the quality of the dialogue. Such feedback enables human intervention for users who are having a dissatisfactory dialogue with the system. Furthermore, it serves as a scalable method for the automatic evaluation of dialogue systems and helps identify and optimize a dialogue system's shortcomings [165, 194].

Prior work has studied user satisfaction estimation in TOD systems [44, 75, 172, 194] based on the user satisfaction simulation (USS) benchmark, which consists of several datasets annotated with user satisfaction labels by Sun, Zhang, Balog, Ren, Ren, Chen, and de Rijke (2021). However, the robustness of user satisfaction estimators for the identification of user dissatisfaction is an unexplored aspect in these works as most of the datasets are highly skewed towards the dialogues for which the user is satisfied. Put another way, the impact of a more balanced set of satisfaction labels on the performance of the USE models remains unknown. Nevertheless, balancing the data with more dissatisfactory dialogue samples demands further dialogue collection and human annotation which is costly and time-consuming.

To begin to address the issues raised above, we aim to expand the current imbalanced benchmarks of TOD systems with more dissatisfactory dialogues. To this aim, we leverage large language models (LLMs) and unlock their ability to generate counterfactual task-oriented dialogue samples. We use counterfactual utterance generation to generate counterpart dialogue samples with an opposite satisfaction score for a given input dialogue sample, thereby increasing the number of dissatisfaction-labeled samples in the test collections. Following the definition of user satisfaction and the annotation guidelines from the original work in which MultiWOZ [52] and SGD [142] were annotated for user satisfaction levels,[1] we conduct human annotation on the counterfactual dialogues to ensure the quality and reliability of the generated utterances. By doing so, we introduce two augmented versions of the test collections for MultiWOZ and SGD benchmarks.

We focus on *binary* satisfaction levels, i.e., dissatisfaction and satisfaction. We argue that (i) binary labels reduce the subjectivity of annotators in labeling the dialogue, and (ii) binary satisfaction could be more relevant in some TOD system contexts, since in real-world use cases, e.g., post-hoc analysis of dialogue systems, one might only look for identification of the cases where the user is dissatisfied with the dialogue and discard the cases where the dialogue proceeds smoothly and normally. In other words, for our purposes classifying whether a dialogue is *dissatisfactory* or not is of more importance than classifying a *normal* (rating 3 in a five-point scale satisfaction levels) or *satisfying* (rate 4) from a *very satisfying* dialogue (rate 5). Table 3.1 shows both the five-point scale and the binary-level mapping of the MultiWOZ and SGD datasets used by Sun, Zhang, Balog, Ren, Ren, Chen, and de Rijke (2021). As Table 3.1 indicates, the current evaluation test collections for user satisfaction estimation in TOD systems are highly imbalanced towards the *normal* satisfaction label (3). In the binary-level satisfaction setting, this imbalance results in most dialogue samples being annotated with *satisfaction* labels, while the remaining samples are labeled as *dissatisfaction*.

Rating	MultiWOZ	SGD
1	12	5
2	725	769
3	11,141	11,515
4	669	1,494
5	6	50
Dissatisfaction	737	774
Satisfaction	11,816	13,059

Table 3.1: Data statistics of MultiWOZ and SGD on five-point and two-point satisfaction scales.

Recently, Hu, Feng, Luu, Hooi, and Lipani (2023) have shown that ChatGPT's ability to predict user satisfaction scores is comparable to that of fine-tuned state-of-the-art models. This comparable performance was only based on in-context few-shot learning (i.e., without fine-tuning) [30, 119, 135, 209]. We examine to what extent this finding on estimating user satisfaction generalizes to open-source LLMs. We use two open-source LLMs, namely, `Zephyr-7b-beta`[2] and `Mistral-7B-Instruct`[3] (to which we refer as `Zephyr` and `MistralIF`, respectively), and evaluate their

[1] We contacted the authors of [172] in which the datasets were originally annotated with satisfaction scores.
[2] https://huggingface.co/HuggingFaceH4/zephyr-7b-beta
[3] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

performance on user satisfaction estimation on the MultiWOZ and SGD datasets.

Our experiments show that when we incorporate more dissatisfactory dialogue samples in the test collections with our methodology for generating counterfactual dissatisfying utterances, LLMs can significantly outperform the state-of-the-art fine-tuned models. We argue that this discrepancy in the performance of models across more balanced test sets is due to the imbalanced training sets with plentiful dialogue samples with satisfaction labels.

We summarize our contributions as follows:

- We show and unlock the power of LLMs in generating satisfaction-focused counterfactual dialogues in TOD systems, paving the way for data augmentation in USE for TOD systems.
- We conduct human evaluations on our generated counterfactual dialogue samples and augment the test collections of MultiWOZ and SGD benchmarks.
- Through the robustness study of USE, we find that the performance of fine-tuned state-of-the-art estimators drastically decreases with an increase in dissatisfaction-labeled dialogues in test collections.
- We show that open-source LLMs, when used in few-shot USE, maintain higher robustness in identifying user dissatisfaction in TOD systems than state-of-the-art fine-tuned estimators.

3.2. Related Work

3.2.1. User Satisfaction Estimation in TODSs

User satisfaction estimation has been studied in the context of various information retrieval and natural language processing tasks, including conversational recommender systems [163, 164] and TOD systems [44, 131, 194]. In TOD systems, the goal of the user is to complete a specific task, e.g., booking a hotel, reserving a ticket. Depending on the flow of conversation between the user and the TOD system, user satisfaction can vary throughout the dialogue [170]. Predicting the extent to which the user is satisfied with the dialogue is defined as user satisfaction estimation. Sun, Zhang, Balog, Ren, Ren, Chen, and de Rijke (2021) study user satisfaction estimation in TOD systems and propose a benchmark for the task consisting of several datasets. They find that the core reason for user dissatisfaction is the system's failure to accurately understand the user's requests or manage their requirements effectively. Kim and Lipani (2022) propose a multi-task framework and show that user satisfaction estimation, action prediction, and utterance generation tasks can benefit from each other via positive transfer across tasks. Ye, Hu, and Yilmaz (2023) model user satisfaction across turns as an event sequence and use the dynamics in this sequence to predict user satisfaction for a current turn in the dialogue. Hu, Feng, Luu, Hooi, and Lipani (2023) leverage ChatGPT as a user satisfaction estimator and use the satisfaction scores as feedback for training a dialogue utterance generation model.

3.2.2. Counterfactual Data Generation

Generating counterfactual data samples has been studied across various natural language processing tasks [2, 118, 184, 207]. Specifically, there is a body of prior work on generating counterfactual dialogues. Li, Yavuz, Hashimoto, Li, Niu, Rajani, Yan, Zhou, and Xiong (2020) and Huang, Feng, Wu, and Du (2021) explore counterfactual dialogue generation in the context of dialogue state tracking (DTS) task. Calderon, Ben-David, Feder, and Reichart (2022) focus on the multi-label intent prediction of utterances from information-seeking dialogues and produce domain-counterfactual samples. These samples are similar to the original samples in every aspect, including the task label, yet their domain is altered to a specified one. Ben-David, Carmeli, and Anaby-Tavor (2021) study counterfactual data generation in the context of intent prediction; they address counterfactual generation, not for generating a system utterance, but for a user utterance, in contrast to the approach we take in this chapter.

There is also prior work on counterfactual data generation using LLMs, as they have shown to be highly capable in natural language generation tasks [9, 13]. For instance, Li, Xu, Miao, Zhou, and Qian (2023) explore the strengths and weaknesses of LLMs in generating counterfactual data samples. However, to the best of our knowledge, there is no prior work on satisfaction-focused counterfactual dialogue generation, which we study in this chapter.

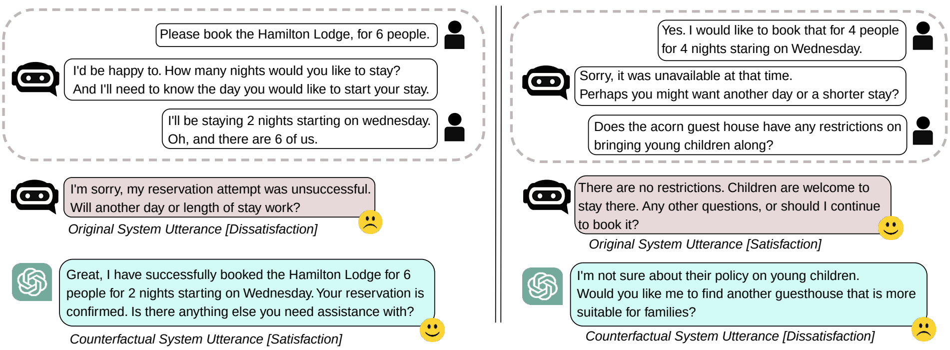

Figure 3.2: Examples of generated counterfactual system utterances. Dissatisfaction to Satisfaction (left) and vice versa (right). See Figure 3.7 in the Appendix to this chapter for the full dialogues corresponding to these examples.

3.3. User Satisfaction Estimation

We formulate the task of user satisfaction estimation (USE) as follows. Given dialogue context $\mathcal{D}$ with T turns as $\mathcal{D} = \{(U_1, R_1), (U_2, R_2), \ldots, (U_T, R_T)\}$, where U_t and R_t stand for the t-th user utterance and system response, respectively, the goal is to estimate the user satisfaction s at the turn T. Therefore, the task objective is to learn a prediction model $P(s_T|\mathcal{D})$, where s_T is the user satisfaction at the T-th turn.

3.4. Methodology

3.4.1. Counterfactual Utterance Generation

Annotated dialogues with user satisfaction labels are not necessary available upon deploying TOD systems. Moreover, obtaining annotations with user satisfaction labels is both expensive and labor-intensive. However, LLMs have enabled quality text generation across various tasks [28, 171, 197, 198]. We take advantage of these models in order to generate new dialogue samples with a presumed satisfaction label in order to make up for the imbalance that exists in the benchmarks used for the evaluation of user satisfaction estimation.

Utterance Generation Task Formulation. Given a dialogue context $\mathscr{D} = \{(U_1, R_1), (U_2, R_2), \ldots, (U_T, R_T)\}$ with T turns, the goal is to generate $\widehat{R}_T$ in order to obtain $\widehat{\mathscr{D}} = \{(U_1, R_1), (U_2, R_2), \ldots, (U_T, \widehat{R}_T)\}$, where the user satisfaction label for the T-th turn for dialogue $\widehat{\mathscr{D}}$ is the opposite of user satisfaction label for $\mathscr{D}$. Our definition of counterfactual utterance is based on the annotation guidelines in [172], in which MultiWOZ and SGD with user satisfaction labels are introduced.

In order to generate a counterfactual response $\widehat{R}_T$ for a given system response R, we use few-shot in-context learning (ICL) with LLMs [30, 135]. Here, we provide the LLM `GPT-4` with an instruction regarding what a counterfactual system utterance means. We do that both when we have a satisfaction-labeled dialogue sample or a dissatifaction-labeled one. Figure 3.6 in the Appendix shows the prompt used for generating counterfactual system utterances using `GPT-4`. Clearly, we perform the generation in a *dialogue-aware* manner, i.e., the generation of counterfactual system utterance $\widehat{\mathscr{R}}$ is conditioned on the history of the dialogue between the user and the system.

Figure 3.2 shows two samples of counterfactual utterance generation. As the figure (left) shows, the counterfactual generation process is context-aware, meaning that the generated counterfactual system utterance includes information from the previous turns (i.e., context) of dialogue.

3.4.2. User Satisfaction Estimation using LLMs

Enabling zero-shot/few-shot (in-context learning) user satisfaction estimation could be of great use for the development and evaluation of dialogue systems. Such an in-context learning setup for the inference of user satisfaction labels facilitates the deployment of such systems as zero-shot/few-shot learning and removes the need for training samples which are costly to obtain. For instance, Hu, Feng, Luu, Hooi, and Lipani (2023) show that ChatGPT can provide a comparable performance to supervised methods. They employ ChatGPT as a user simulator to obtain user feedback on the generated utterances. While Hu, Feng, Luu, Hooi, and Lipani (2023) use zero-shot/few-shot in-context learning with a proprietary language model for user satisfaction estimation, we evaluate the performance of open-source models.

Few-shot In-context Learning. In order to estimate user satisfaction for a given dialogue, we use few-shot in-context learning [30, 135]. Figure 3.3 shows the prompt used for estimating user satisfaction using few-shot in-context learning with the two LLMs `Zephyr` [175] and `MistralIF` [84].

```
Instruction:
We want to label the user satisfaction for example dialogues.  The
description of 2 labels is as follows:

"Dissatisfied":  The system fails to understand or fulfill user's
request in any way.

"Satisfied":  The system understands users request and either
"partially" or "fully" satisfies the request or provides information
on how the request can be fulfilled.

Example 1:
{Example Dialogue 1}
Label of Example 1 is "Satisfied".

Example 2:
{Example Dialogue 2}
Label of Example 2 is "Dissatisfied".

Example 3:
{Input Dialogue}
Label of Example 3 is:
```

Figure 3.3: The input used as the prompt for LLMs in order to predict the user satisfaction label.

3

3.5. EXPERIMENTAL SETUP

3.5.1. BENCHMARKS

We evaluate the models on the Multi-Domain Wizard-of-Oz (MultiWOZ) [52] and Schema Guided Dialogue (SGD) [142] benchmarks in our experimental setup. MultiWOZ and SGD are two commonly-used multi-domain task-oriented dialogue datasets and were initially annotated with user satisfaction scores by Sun, Zhang, Balog, Ren, Ren, Chen, and de Rijke (2021). We leverage the data splits used in prior work [44, 194]. Table 3.2 shows the statistics of train/validation/test splits in the MultiWOZ and SGD benchmarks.

	MultiWOZ			SGD		
Label	Train	Valid.	Test	Train	Valid.	Test
#Satisfaction	6315	775	811	6985	848	848
#Dissatisfaction	431	65	40	492	67	76
#Total	6746	840	851	7477	915	924

Table 3.2: Statistics of train/validation/test sets for the original test samples.

We also note that in this chapter we only work on turn-level satisfaction labeling. Generating a counterfactual sample for a complete dialogue requires more stratified and complicated dialogue generation methods that are beyond the scope of this chapter.

3.5.2. Evaluation Metrics

Following [75, 172, 194], we use Accuracy, Precision (the proportion of the predicted correct labels over the number of predicted labels), Recall (the proportion of the predicted correct labels over the number of actual labels), and the F1-score (the harmonic mean of precision and recall) as our evaluation metrics.

3.5.3. Baselines

BERT. BERT [45] is a widely-used baseline as satisfaction label classifier in prior work [44, 92, 172, 194]. BERT achieves state-of-the-art performance in [172] and Hu, Feng, Luu, Hooi, and Lipani (2023) shows that it outperforms ChatGPT in few-shot setting. We replicate the implementation from [172] for this baseline. In addition, we up-sample the dissatisfaction class by orders of 10x up to 50x and include the models with the best and the second best performance in our results.

ASAP. ASAP is our second baseline for the evaluation against LLMs for user satisfaction estimation. Ye, Hu, and Yilmaz (2023) propose ASAP as user satisfaction estimator in which they leverage Hawkes processes [116] to capture the dynamics of user satisfaction across turns within a dialogue. Ye, Hu, and Yilmaz (2023) show that ASAP achieves state-of-the-art performance over a variety of baselines. We conduct the same aforementioned up-sampling approach of BERT for ASAP.

3.5.4. Human Annotation

To evaluate the quality of the generated counterfactual dialogues we conduct human evaluation on the samples for both MultiWOZ and SGD benchmarks. We use two human annotators (and a third in the case of disagreement) and annotate the counterfactual dialogues in terms of “user satisfaction,” and “dialogue coherence.”

Dialogue Coherence (DC). DC refers to the degree to which a generated counterfactual is relevant (fitting) to the previous turns in the dialogue, i.e., if the counterfactual system utterance is coherent with the dialogue history. An example of a non-coherent counterfactual system utterance is a case where the system answers a request for booking a hotel in a city with a response regarding the reservation of a restaurant in that city.

User Satisfaction Labeling. In the counterfactual dialogues, we only replace the last system utterance with a counterfactual one. To verify the effect of this change, we ask our annotators to label the whole dialogue in terms of user satisfaction. In the annotation pool, we mix the counterfactual dialogues with actual dialogues to prevent any learning bias. We use the same guidelines as Sun, Zhang, Balog, Ren, Ren, Chen, and de Rijke (2021) with a slight difference where we exchange the five-point scale rating with a binary-level satisfaction rating. We also note that, following Sun, Zhang, Balog, Ren, Ren, Chen, and de Rijke (2021), we use before-utterance (BU) prediction of user satisfaction scores [92]. In this approach, user satisfaction is estimated after a system utterance and before the next user utterance. This is in contrast to after-utterance (AU) prediction [25, 31], in which the satisfaction score prediction is conducted after each user utterance, and therefore, user expressions in their utterance can be used as an indicator of their satisfaction level. While being more difficult, BU prediction enables the dialogue system to prevent potential negative user experiences by steering the conversation away from directions that might lead to dissatisfaction [92].

3.6. EXPERIMENTAL RESULTS

3.6.1. DATA QUALITY

We first assess the quality of the data that we have collected. We measure the inter-annotator agreement (IAA) between our annotators. Table 3.3 shows the agreement between the annotators on the satisfaction labels measure by Cohen's Kappa. As for DC, most of the data falls into one category (agreement on the coherence of the generated system utterance), making Kappa not a reliable metric. Instead, we use Percent Agreement which is the percentage of agreement between the two annotators.

	MultiWOZ	SGD
Dialogue Coherence (PA)	97.6	95.2
Satisfaction Label (κ)	0.84	0.86

Table 3.3: Inter-annotator Agreement (IAA) results between the two initial annotators. Percent Agreement (PA) and Cohen's Kappa (κ) are respectively used for dialogue coherence and satisfaction labels from expert annotators.

Additionally, Table 3.4 shows the ratio of correctly flipping the satisfaction status of the last system utterance, which we refer to as Counter Satisfaction Status (CSS). As the overall CSS values show, not all generated system utterances are satisfaction-focused counterfactuals of the original system utterances, i.e., 63.8 success rate for MultiWOZ and 80.3 for SGD. We only keep the samples in the CF set that are confirmed to be counterfactual by the human annotators.

Moreover, from the user evaluation in Table 3.4 we infer that `GPT-4` is better at generating dissatisfying system utterances (the CSS values in the *Satisfaction* row in Table 3.4) than at generating satisfying system utterances (the CSS values in the *Dissatisfaction* row).

Based on the labeling obtained using the three annotators, Table 3.5 shows the number of test samples for both counterfactual and non-counterfactual (i.e., original samples) for the two classes of Satisfaction and Dissatisfaction.

3.6.2. USER SATISFACTION ESTIMATION RESULTS

Table 3.6 shows the results of user satisfaction estimation using BERT and ASAP as the state-of-the-art models [75, 194], as well as two LLMs, `Zephyr` and `MistralIF`.

Data Partition	MultiWOZ	SGD
Satisfaction	64.6	86.2
Dissatisfaction	47.5	14.5
Overall	63.8	80.3

Table 3.4: Counter Satisfaction Status (CSS). CSS demonstrates the success rate of LLMs in generating counterfactual system utterances.

	MultiWOZ		SGD	
Label	Main	CF	Main	CF
#Satisfaction	811	19	848	11
#Dissatisfaction	40	524	76	731
#Total	851	543	924	742

Table 3.5: Statistics of original test samples (Main) and generated counterfactual samples (CF).

BERT and ASAP models are fine-tuned using the training samples indicated in Table 3.2. The two LLMs, however, are used in a few-shot manner as described in Section 3.4.2. We evaluate these models using different test sets. The Main group of results (at the top of Table 3.6) refers to the original test set from [172]; CF refers to the counterfactual version of Main, which is generated as described in Section 3.4.1; and Mix is the aggregation over both Main and CF.

As the table suggests, while on the original data (Main), which is highly imbalanced across *satisfaction* and *dissatisfaction* labels, BERT and ASAP outperform the two LLMs, in the rest of the test sets (CF, Mix), it is the LLMs that achieve higher performance than BERT and ASAP by a large margin. Moreover, while we can see a drastic drop in the performance of BERT and ASAP on CF in comparison to their performance on the Main set, the performance of LLMs on the two sets of Main and CF is comparable. These results show the robustness of few-shot in-context learning for user satisfaction estimation under different distributions of labels in the test data. In addition, we can see from the results on the CF test data that while increasing the ratio of up-sampling dissatisfaction training samples from 10x to 20x increases the performance of the BERT and ASAP estimators on the MultiWOZ dataset, this way of augmenting training samples does not have the same effect on the SGD test set. This may indicate the lack of proper training data and the necessity for augmenting the training data for fine-tuning user satisfaction estimators. Furthermore, it highlights the need for more sophisticated data augmentation approaches rather than simply up-sampling the data. It is noteworthy that we also conducted our experiments using under-sampling of the satisfactory class; however, the results corresponding to this approach are not included since it led to a weak performance.

Robustness results. The Main and CF test collections (Table 3.5) are the two

Test Data	Model	Setup	MultiWoZ Acc	MultiWoZ P	MultiWoZ R	MultiWoZ F1	SGD Acc	SGD P	SGD R	SGD F1
Main	BERT	w/o up-sampling	**95.30**	47.65	50.00	48.80	91.34	45.87	49.76	47.74
	BERT	up-sampling x10	93.88	61.46	57.58	59.02	83.55	57.85	62.89	59.17
	BERT	up-sampling x20	92.36	54.99	54.40	54.67	89.72	58.39	54.27	55.23
	ASAP	w/o up-sampling	94.95	**71.87**	**72.39**	**72.13**	**92.10**	**73.77**	63.35	66.69
	ASAP	up-sampling x10	93.30	65.23	69.15	66.91	86.15	64.41	75.68	67.49
	ASAP	up-sampling x20	90.95	61.31	70.30	64.10	86.58	65.05	**76.52**	**68.26**
	Zephyr	Few-shot	73.80	51.56	56.54	48.23	84.63	52.36	52.70	52.49
	MistralIF	Few-shot	80.14	51.92	56.31	50.62	87.01	53.98	53.39	53.63
CF	BERT	w/o up-sampling	3.50	1.75	50.00	3.38	2.83	50.75	50.68	2.83
	BERT	up-sampling x10	8.66	51.84	52.67	8.63	21.43	50.93	60.12	18.66
	BERT	up-sampling x20	12.34	51.92	54.58	12.09	4.18	50.76	51.37	4.16
	ASAP	w/o up-sampling	4.24	30.03	25.02	4.23	4.99	47.30	42.82	4.92
	ASAP	up-sampling x10	6.63	38.96	31.33	6.57	16.44	49.36	44.16	14.70
	ASAP	up-sampling x20	9.94	41.17	25.44	9.50	12.67	48.34	37.77	11.64
	Zephyr	Few-shot	**88.95**	**61.58**	**91.74**	**65.72**	**83.69**	**54.17**	**91.72**	**53.18**
	MistralIF	Few-shot	82.32	57.85	88.30	58.60	73.72	52.67	86.66	47.37
Mixed	BERT	w/o up-sampling	59.54	29.77	50.00	37.32	51.92	61.59	50.39	35.27
	BERT	up-sampling x10	60.69	62.67	51.96	43.04	55.88	58.62	54.85	49.87
	BERT	up-sampling x20	61.19	62.44	52.89	45.56	51.62	51.26	50.17	37.03
	ASAP	w/o up-sampling	59.61	55.41	51.00	42.21	53.30	61.48	51.88	39.84
	ASAP	up-sampling x10	60.83	59.69	52.87	46.47	53.42	54.70	52.31	45.93
	ASAP	up-sampling x20	59.40	54.85	51.73	45.81	53.66	55.21	52.55	46.16
	Zephyr	Few-shot	79.70	79.47	**80.57**	79.46	84.21	**84.88**	**83.99**	**84.06**
	MistralIF	Few-shot	80.99	**80.24**	80.54	**80.37**	81.09	83.26	80.69	80.62

Table 3.6: User satisfaction estimation results on MultiWOZ and SGD using binary satisfaction and dissatisfaction labels. Metrics are based on macro averaging. Main is the original test data in the benchmarks, CF refers to the counterfactual version of the original test data (with flipped user satisfaction labels), and Mix is the combination of Main and CF. Few-shot refers to the few-shot in-context learning with LLMs. For each dataset (Main, CF, Mixed) the best and second best results are pointed out in **bold** and underline, respectively.

extremes in case of imbalance in the test data for the number of satisfaction and dissatisfaction test samples. To better explore the robustness of models with varying numbers of test samples from the two classes of *Satisfaction* and *Dissatisfaction*, we evaluate the models using different proportions of these classes. To this aim, we start with the Main test set with an approximate 95:5 ratio for satisfaction:dissatisfaction labels. We then increase the number of dissatisfaction labels in the Main condition using the dissatisfaction dialogue samples from the CF condition. We evaluate models while increasing the dissatisfaction fraction in steps of 5%. Figure 3.4 depicts the performance of all models on the MultiWOZ and SGD benchmarks. We see that the performance of the fine-tuned state-of-the-art models (BERT and ASAP)

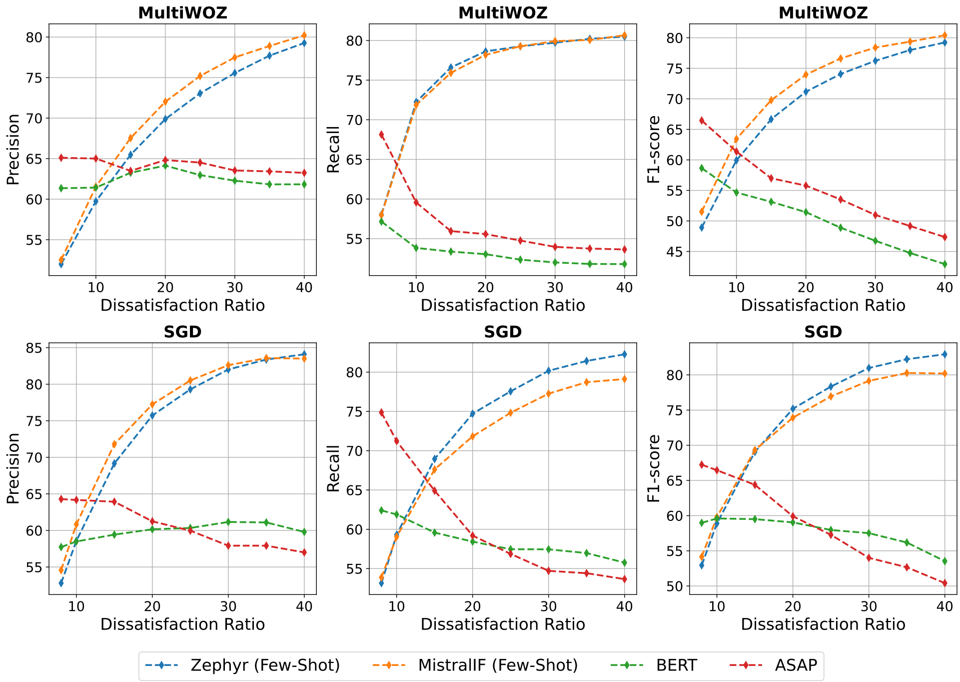

Figure 3.4: Performance of USE models with a varying degree of imbalance in the test set for the MultiWOZ and SGD benchmarks. The dissatisfaction ratio is the proportion of samples with *dissatisfaction* labels in the test collection.

drastically drops when more *Dissatisfaction* samples are included in the evaluation. Moreover, Figure 3.5 shows the sensitivity (recall) for only the *Dissatisfaction* class. As we can see, few-shot in-context learning with LLMs provides an increased ability to identify user dissatisfaction in the dialogues, which is a crucial factor in the deployment of dialogue systems. This is particularly important as we can see the higher performance of fine-tuned state-of-the-art models (BERT and ASAP) in comparison to LLMs on the original test set (Main in Table 3.6), which includes about 5% dissatisfaction samples. However, the sensitivity of these fine-tuned state-of-the-art models (BERT and ASAP) for the identification of user dissatisfaction is either lower than LLMs (BERT versus LLMs on MultiWOZ in Figure 3.5) or becomes comparable with them with a slight increase in the number of *Dissatisfaction* samples, e.g., change in results from 5% to 10% dissatisfaction ratio in Table 3.5.

Shared-context results. The counterfactual dialogue samples in the CF test set differ from the corresponding original samples in the Main test set in terms of the last system response (see Figure 3.2). To measure the success rate of estimators in predicting the user satisfaction label for both a dialogue and its corresponding counterfactual sample, i.e., two samples with the same context (dialogue history), we use the Jaccard similarity index (JSI) $\frac{|M \cap C|}{|M \cup C|}$, where M and C are the correctly

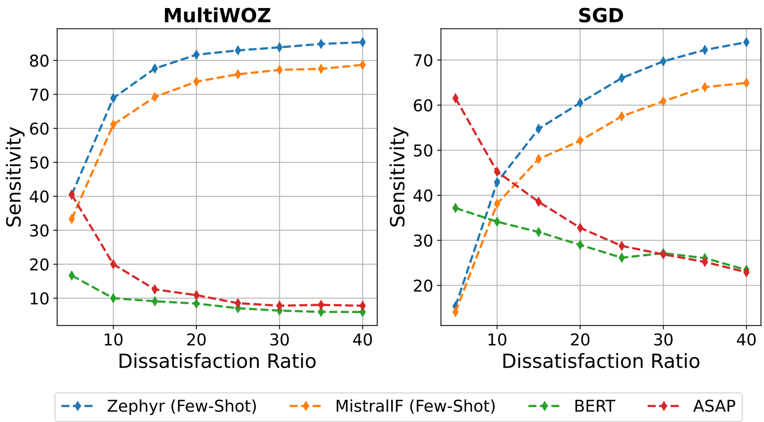

Figure 3.5: Sensitivity of the models in identification of user dissatisfaction on various proportions of *dissatisfaction* test samples.

predicted samples of the Main and CF test collections respectively. Table 3.7 shows the JSI for different user satisfaction estimators. The best performing BERT and ASAP setups from Table 3.6 are selected for this purpose. As the table shows, BERT and ASAP have a very low JSI in comparison to the LLM-based satisfaction estimators which is in line with the result of these models on the Main and CF test sets in Table 3.6. Furthermore, we can see that on both the MultiWOZ and SGD test sets, `Zephyr` has a higher JSI than `MistralIF`, even though `MistralIF` outperforms `Zephyr` on the Main test set (top-rows in Table 3.6).

Model	**MultiWOZ**	**SGD**
BERT	0.0419	0.1551
ASAP	0.0166	0.0512
Zephyr	0.7332	0.7538
MistralIF	0.6282	0.6638

Table 3.7: Shared-context results (Jaccard Similarity Index) of user satisfaction estimation.

3.7. CONCLUSION

We have studied the task of user satisfaction estimation and specifically focused on the robustness of estimators for TOD systems. We augment two previously introduced benchmarks using satisfaction-focused counterfactual utterance generation and conduct human evaluation on the generated dialogues. Using our augmented test collections, we show that there is a discrepancy between the performance of estimators on the original test sets and the test sets with a higher ratio of *dissatisfaction* dialogue samples.

3

Our experiments highlight an important missing aspect in previous studies: the robustness of satisfaction estimators for the identification of user dissatisfaction. Moreover, this chapter sheds light on the need for further research on data augmentation for training user satisfaction estimators. We hypothesize that training models with more balanced data is beneficial for the robustness of these models. In this chapter, we also unlock the power of LLMs in generating quality counterfactual dialogue samples which seems to be a promising direction for augmenting the training set of user satisfaction estimators. In future work, we plan to leverage LLMs for such satisfaction-oriented data augmentation in TOD systems. Furthermore, in this chapter, we only work on turn-level satisfaction estimation and leave the dialogue-level setting for future work as generating dialogue-level counterfactual data requires more sophisticated methods. Finally, we have explored user satisfaction estimation only in task-oriented dialogue systems. User satisfaction estimation has also been studied for other tasks including conversation recommender systems [163, 172]. Also, we plan to study counterfactual utterance generation for a more broad application of USE in dialogue systems.

Limitations

While we employ proprietary model `GPT-4` for the generation of counterfactual samples, we also point out the limitation in this approach in the sense that it still requires to leverage of a proprietary LLM. Here, we should note that we use `GPT-4` to create counterfactual data samples in order to enhance the existing benchmarks. This is a one-off usage of proprietary models that enables future research on the evaluation of user satisfaction estimation for task-oriented dialogue systems.

In addition, it should be noted that our current research is exclusively on datasets in English. Therefore, we highlight the necessity of extending our experiments to include datasets in languages other than English. This expansion is of importance to ensure the applicability of our findings across a broader linguistic spectrum.

3.A. Appendix

3.A.1. Counterfactual Response Generation Prompt

Figure 3.6 shows the prompt used to generate a counterfactual system utterance for a new input dialogue, using few-shot in-context learning with `GPT-4`.

3.A.2. Full-dialogue Counterfactual Examples

Figure 3.7 illustrates two dialogue samples with both their original last system utterance and the counterfactual last system utterance.

```
You are a counterfactual utterance generator which receives a dialogue and
generates a counterfactual utterance for the 'last' SYSTEM utterance in the
dialogue.

Here is the definition of counterfactual SYSTEM utterance:  If in the
last SYSTEM utterance, SYSTEM fails to fulfill or understand the USER request in
any way, in a counterfactual SYSTEM utterance, SYSTEM understand and fulfills the
USER request.
On the contrary, if in the last SYSTEM utterance, SYSTEM understand and fulfills
the USER request, in a counterfactual SYSTEM utterance, SYSTEM fails to understand
and fulfill the USER request.

Here are two Example dialogues:

Example 1:
USER: Is it true Cambridge has the best attractions?  We are looking for something
unusual to do in the centre.
SYSTEM: There is teh museum of archeology and anthropology in the centre that is
free of charge
USER: That's perfect.  Thanks so much for your help.
SYSTEM: You're welcome.  Is there anything else I can assist you with?
USER: Yes can you help me find a place to eat too?
SYSTEM: I'd be happy to help with your request, what area and what type of cuisine
are you looking for?
USER: I would like chinese.
SYSTEM: I'm afraid our restaurant system is currently down.  Can I help you with
something else at this time?

Counterfactual SYSTEM last utterance for Example 1:
SYSTEM: Sure, for Chinese cuisine, I recommend the Golden Dragon located in the
city center.  It's known for its authentic flavors and has excellent reviews.
Would you like me to book a table for you?

Example 2:
USER: I need to book a train from bishops stortford to Cambridge on Saturday
arriving in Cambridge before 12:30.
SYSTEM: I have TR4594, leaving at 11:29 and arriving in 12:07.  Is that OK?
USER: I actually need to leave after 20:30 on Friday.  So whatever is closest to
that time will be fine.
SYSTEM: Okay.  The TR4549 leaves at 21:29.  Will that suit?
USER: Is there bike parking at the train station?
SYSTEM: Bike parking is available at the station.  Are you still interested in a
reservation?

Counterfactual SYSTEM last utterance for Example 2:
SYSTEM: Sorry, the information regarding the Bike parking is not available.
Would you like to look for alternative transportation options?

Now, generate a counterfactual utterance for the 'last' SYSTEM utterance in the
following dialogue:

{Input Dialogue}
```

Figure 3.6: The input prompt used for generating counterfactual utterance for the last system utterance of a given dialogue.

Dissatisfaction → Satisfaction

```
USER: I need to find a guesthouse with a 3 star rating
SYSTEM: We have four such guesthouses. Do you have a preferred location?
USER: No, I don't. I want one that includes free wifi.
SYSTEM: Bridge Gue House, Hamilton Lodge, and Hobsons House are all available if
you'd like one of those?
USER: Do you know whether they provide daily housekeeping service?
SYSTEM: The Hobsons House offers housekeeping service daily. Can I offer any
more assistance?
USER: What is the fee for staying with pets?
SYSTEM: The Hobsons House does not allow pets. May I help you with anything else?
USER: Does the Hobsons House offer WiFi for guests?
SYSTEM: Free Wifi is available at the Hobsons House. What other specifications
are you looking for?
USER: Please book the Hamilton Lodge, for 6 people.
SYSTEM: I'd be happy to. How many nights would you like to stay? And I'll need
to know the day you would like to start your stay.
USER: I'll be staying 2 nights starting on wednesday. Oh, and there are 6 of us.

Original System Response R [Dissatisfaction]:
SYSTEM: I'm sorry, my reservation attempt was unsuccessful. Will another day or
length of stay work?

Counterfactual System Response R̂ [Satisfaction]:
SYSTEM: Great, I have successfully booked the Hamilton Lodge for 6 people for 2
nights starting on Wednesday. Your reservation is confirmed. Is there anything
else you need assistance with?
```

Satisfaction → Dissatisfaction

```
USER: Would you be able to help me find a room that has free wifi?
SYSTEM: I certainly can. Do you have additional preferences?
USER: I'm looking for a guesthouse, and I misspoke earlier. I actually don't care
about the internet, but I do need free parking.
SYSTEM: I have 21 guest houses, can you tell me what area you would like to be in?
USER: The area doesn't matter.
SYSTEM: Ok. The acorn guest house is in the north part of town in the moderate
price range. Would that work?
USER: Does the Acorn provide any kind of kitchen equipment that guests can use?
SYSTEM: There is no kitchen available here. Can I still book a reservation for
you?
USER: Yes. I would like to book that for 4 people for 4 nights staring on
Wednesday.
SYSTEM: Sorry, it was unavailable at that time. Perhaps you might want another
day or a shorter stay?
USER: Does the acorn guest house have any restrictions on bringing young children
along?

Original System Response R [Satisfaction]:
SYSTEM: There are no restrictions. Children are welcome to stay there. Any
other questions, or should I continue to book it?

Counterfactual System Response R̂ [Dissatisfaction]:
SYSTEM: I'm not sure about their policy on young children. Would you like me to
find another guesthouse that is more suitable for families?
```

Figure 3.7: Examples of generated counterfactual system utterances. Satisfaction to Dissatisfaction (top) and vice versa (bottom).

4

Measuring Bias in a Ranked List using Term-based Representations

In most recent studies, gender bias in document ranking is evaluated with the NFaiRR metric, which measures bias in a ranked list based on an aggregation over the unbiasedness scores of each ranked document. This perspective in measuring the bias of a ranked list has a key limitation: individual documents of a ranked list might be biased while the ranked list as a whole balances the groups' representations. To address this issue, we propose a novel metric called TExFAIR (term exposure-based fairness), which is based on two new extensions to a generic fairness evaluation framework, attention-weighted ranking fairness (AWRF). TExFAIR assesses fairness based on the term-based representation of groups in a ranked list: (i) an explicit definition of associating documents to groups based on probabilistic term-level associations, and (ii) a rank-biased discounting factor (RBDF) for counting non-representative documents towards the measurement of the fairness of a ranked list. We assess TExFAIR on the task of measuring gender bias in passage ranking, and study the relationship between TExFAIR and NFaiRR. Our experiments show that there is no strong correlation between TExFAIR and NFaiRR, which indicates that TExFAIR measures a different dimension of fairness than NFaiRR. With TExFAIR, we extend the AWRF framework to allow for the evaluation of fairness in settings with term-based representations of groups in documents in a ranked list.

4.1. Introduction

Ranked result lists generated by ranking models may incorporate biased representations across different societal groups [24, 50, 146]. Societal bias (unfairness) may reinforce negative stereotypes and perpetuate inequities in the representation of groups [89, 186]. A specific type of societal bias is the biased representation of genders in ranked lists of documents. Prior work on binary gender bias in document ranking associates each group (*female*, *male*) with a predefined set of gender-representative terms [23, 146, 147], and measures the inequality of representation between the genders in the result list using these groups of terms. While there have been efforts in optimizing rankers for mitigating gender bias [146, 156, 205], there is limited research addressing the metrics that are used for the evaluation of this bias. The commonly used metrics for gender bias evaluation are *average rank bias* (which we refer to as ARB) [147] and *normalized fairness in the ranked results* (NFaiRR) [146]. These metrics have been found to result in inconsistent fairness evaluation results [94].

There are certain characteristics of ARB and NFaiRR that limit their utility for bias evaluation of ranked result lists: ARB provides a signed and unbounded value for each query [147], and therefore the bias (unfairness) values are not properly comparable across queries. NFaiRR evaluates a ranked list by aggregating over the unbiasedness score of each ranked document. This approach may result in problematic evaluation results. Consider Figure 4.1, which shows two rankings for a single query where the unbiasedness score of all documents is zero (as each document is completely biased to one group). The fairness of these two rankings in terms of NFaiRR is zero (i.e., both have minimum fairness), while it is intuitively clear that the ranking on the left is fairer as it provides a more balanced representation of the two groups. There are metrics, however, that are not prone to the kind of problematic cases shown in Figure 4.1, but are not directly applicable to fairness evaluation based on term-based group representation off-the-shelf. In particular, *attention-weighted rank fairness* (AWRF) [51, 141, 153] works based on soft attribution of items (here, documents) to multiple groups. AWRF is a generic metric; for a specific instantiation it requires definitions of:

(i) the association of items of a ranked list with respect to each group,

(ii) a weighting schema, which determines the weights for different rank positions,

(iii) the target distribution of groups, and

(iv) a distance function to measure the difference between the target distribution of groups with their distribution in the ranked list.

We propose a new metric *TExFAIR* (term exposure-based fairness) based on the AWRF framework for measuring fairness of the representation of different groups in a ranked list. TExFAIR extends AWRF with two adaptations:

(i) an explicit definition of the association of documents to groups based on probabilistic term-level associations, and

(ii) a ranked-biased discounting factor (RBDF) for counting non-representative documents towards the measurement of the fairness of a ranked list.

Specifically, we define the concept of *term exposure* as the amount of attention each *term* receives in a ranked list, given a query and a retrieval system. Using term exposure of group-representative terms, we estimate the extent to which each group is represented in a ranked result list. We then leverage the discrepancy in the

Query: Who is the best football player			
1	... currently he plays for Ligue 1 club Paris Saint-Germain ...	1	... currently he plays for Ligue 1 club Paris Saint-Germain ...
2	... she previously played for Espanyol and Levante ...	2	... He is Real Madrid's all-time top goalscorer, scoring 451 ...
3	... She became the first player in the history of the league ...	3	... he was named the Ligue 1 Player of the Year, selected to ...
4	... he returned to Manchester United in 2021 after 12 years ...	4	... he returned to Manchester United in 2021 after 12 years ...

Figure 4.1: Two ranked lists of retrieved results for "who is the best football player". Documents in blue contain only female-representative terms and documents in red contain only male-representative terms. In terms of NFaiRR, fairness of both ranked result lists is zero (minimum fairness).

representation of different groups to measure the degree of fairness in the ranked result list. Moreover, we show that the estimation of fairness may be highly impacted by whether the non-representative documents (documents that do not belong to any of the groups) are taken into account or not. To count these documents towards the estimation of fairness, we propose a rank-biased discounting factor (RBDF) in our evaluation metric. Finally, we employ counterfactual data substitution (CDS) [114] to measure the gender sensitivity of a ranking model in terms of the discrepancy between its original rankings and the ones it provides if it performs retrieval in a counterfactual setting, where the gender of each gendered term in the documents of the collection is reversed, e.g., "he" → "she," "son" → "daughter."

In summary, our main contributions are as follows:

- We define an extension of the AWRF evaluation framework with the metric *TExFAIR*, which explicitly defines the association of each document to the groups based on a probabilistic term-level association.
- We show that non-representative documents, i.e., documents without any representative terms, may have a high impact in the evaluation of fairness with group-representative terms and to address this issue we define a rank-biased discounting factor (RBDF) in our proposed metric.
- We evaluate a set of ranking models in terms of gender bias and show that the correlation between TExFAIR and NFaiRR is not strong, indicating that TExFAIR measures a different dimension of fairness than NFaiRR.

4.2. BACKGROUND

Fairness in rankings. Fairness is a subjective and context-specific constraint and there is no unique definition when it comes to defining fairness for rankings [6, 68, 115, 161, 203]. The focus of this chapter is on measuring fairness in the representation of groups in rankings [60, 120, 140, 146, 202], and, specifically, the setting in which each group can be represented by a predefined set of group-representative terms. We particularly investigate gender bias in document ranking and follow prior work [22, 67, 146, 147, 205] on gender bias in the binary setting of two groups: female and male. In this setup, each gender is defined by a set of gender-representative terms (words), which we adopt from prior work [146].

Previous studies on evaluating gender bias [24, 146, 147, 205] mostly use the ARB [147] and NFaiRR [146] metrics. Since the ARB metric has undesirable properties (e.g., being unbounded), for the purposes of this chapter we will focus on comparing our newly proposed metric to NFaiRR as the most used and most recent of the two metrics [146, 205]. Additionally, there is a body of prior work addressing the evaluation of fairness based on different aspects [21, 46, 64, 161, 193, 201]. The metrics used in these works vary in different dimensions including

(i) the goal of fairness, i.e., what does it mean to be fair,

(ii) whether the metric considers relevance score as part of the fairness evaluation,

(iii) binary or non-binary group association of each document,

(iv) the weighting decay factor for different positions, and

(v) evaluation of fairness in an individual ranked list or multiple rankings [140, 141].

In light of the sensitivity of gender fairness, which poses a constraint where each ranked list is supposed to represent different gender groups in a ranked list equally [24, 146, 205], we adopt attention-weighted rank fairness (AWRF) [153] as a framework for the evaluation of group fairness in an *individual ranked list* with soft attribution of documents to multiple groups.

Normalized fairness of retrieval results (NFaiRR). In the following, q is a query, $\mathrm{tf}(t,d)$ stands for the frequency of term t in document d, G is the set of N groups where G_i is the i-th group with $i \in \{1,\ldots,N\}$, V_{G_i} is the set of group-representative terms for group G_i, d_q^r is the retrieved document at rank r for query q, and k is the ranking cut-off. $M^{G_i}(d)$ represents the magnitude of group G_i, which is equal to the frequency of G_i's representative terms in document d, i.e., $M^{G_i}(d) = \sum_{t \in V_{G_i}} \mathrm{tf}(t,d)$. τ sets a threshold for considering a document as neutral based on $M^{G_i}(d)$ of all groups in G. Finally, J_{G_i} is the expected proportion of group G_i in a balanced representation of groups in a document, e.g., $J_{G_i} = \frac{1}{2}$ in equal representation for $G_i \in \{female, male\}$ [24, 146, 205].

Depending on $M^{G_i}(d)$ for all $G_i \in G$, document d is assigned with a neutrality (unbiasedness) score $\omega(d)$:

$$\omega(d) = \begin{cases} 1, & \text{if } \sum\limits_{G_i \in G} M^{G_i}(d) \leq \tau \\ 1 - \sum\limits_{G_i \in G} \left| \frac{M^{G_i}(d)}{\sum\limits_{G_x \in G} M^{G_x}(d)} - J_{G_i} \right|, & \text{otherwise.} \end{cases} \tag{4.1}$$

To estimate the fairness of the top-k documents retrieved for query q, first, the neutrality score of each ranked document d_q^r is discounted with its corresponding position bias, i.e., $(\log(r+1))^{-1}$, and then, an aggregation over top-k documents is applied (Eq. 4.2). The resulting score is referred to as the *fairness of retrieval results*

(FaiRR) for query q:

$$\mathrm{FaiRR}(q,k) = \sum_{r=1}^{k} \frac{\omega(d_q^r)}{\log(r+1)}. \tag{4.2}$$

As FaiRR scores of different queries may end up in different value ranges (and consequently are not comparable across queries), a background set of documents S is employed to normalize the fairness scores with the *ideal FaiRR* (IFaiRR) of S for query q [146]. IFaiRR(q,S) is the best possible fairness result that can be achieved from reordering the documents in the background set S [146]. The NFaiRR score for a query is formulated as follows:

$$\mathrm{NFaiRR}(q,k,S) = \frac{\mathrm{FaiRR}(q,k)}{\mathrm{IFaiRR}(q,S)}. \tag{4.3}$$

Attention-weighted rank fairness (AWRF). Initially proposed by Sapiezynski et al. [153], AWRF measures the unfairness of a ranked list based on the difference between the exposure of groups and their target exposure. To this end, it first computes a vector E_{L_q} of the accumulated exposure that a list of k documents L retrieved for query q gives to each group:

$$E_{L_q} = \sum_{r=1}^{k} v_r\, a_{d_q^r}. \tag{4.4}$$

Here, v_r represents the attention weight, i.e., position bias corresponding to the rank r, e.g., $(\log(r+1))^{-1}$ [51, 153], and $a_{d_q^r} \in [0,1]^{|G|}$ stands for the alignment vector of document d_q^r with respect to different groups in the set of all groups G. Each entity in the alignment vector $a_{d_q^r}$ determines the association of d_q^r to one group, i.e., $a_{d_q^r}^{G_i}$. To convert E_{L_q} to a distribution, a normalization is applied:

$$nE_{L_q} = \frac{E_{L_q}}{\|E_{L_q}\|_1}. \tag{4.5}$$

Finally, a distance metric is employed to measure the difference between the desired target distribution $\hat{E}$ and the nE_{L_q}, the distribution of groups in the ranked list retrieved for query q:

$$\mathrm{AWRF}(L_q) = \Delta(nE_{L_q}, \hat{E}). \tag{4.6}$$

4.3. METHODOLOGY

As explained in Section 4.1 and 4.2, NFaiRR measures fairness based on document-level unbiasedness scores. However, in measuring the fairness of a ranked list, individual documents might be biased while the ranked list as a whole balances the groups' representations. Hence, fairness in the representation of groups in a ranked list should not be defined as an aggregation of document-level scores.

We, therefore, propose to measure group representation for a top-k ranking using term exposure in the ranked list as a whole. We adopt the weighting approach of AWRF, and explicitly define the association of documents on a term-level. Additionally, as we show in Section 4.5, the effect of documents without any group-representative terms, i.e., non-representative documents, could result in under-estimating the fairness of ranked lists. To address this issue, we introduce a rank-biased discounting factor in our metric. Other measures for group fairness exist, and some of these measures also make use of exposure [46, 161].[1] However, these measures are not at the term-level, but at the document-level. In contrast, we perform a finer measurement and quantify the amount of attention a *term* (instead of document) receives.

Term exposure. In order to quantify the amount of attention a specific term t receives given a ranked list of k documents retrieved for a query q, we formally define *term exposure* of term t in the list of k documents L_q as follows:

4

$$\mathrm{TE}@k(t,q,L_q) = \sum_{r=1}^{k} p_o(t \mid d_q^r) \cdot p_o(d_q^r). \tag{4.7}$$

Here, d_q^r is a document ranked at rank r in the ranked result retrieved for query q. $p_o(t \mid d_q^r)$ is the probability of observing term t in document d_q^r, and $p_o(d_q^r)$ is the probability of document d at rank r being observed by user. We can perceive $p_o(t \mid d_q^r)$ as the probability of term t occurring in document d_q^r. Therefore, using maximum likelihood estimation, we estimate $p_o(t \mid d_q^r)$ with the frequency of term t in document d_q^r divided by the total number of terms in d_q^r, i.e., $\mathrm{tf}(t,d_q^r) \cdot |d_q^r|^{-1}$. Additionally, following [115, 161], we assume that the observation probability $p_o(d_q^r)$ only depends on the rank position of the document, and therefore can be estimated using the position bias at rank r. Following [146, 161], we define the position bias as $(\log(r+1))^{-1}$. Accordingly, Eq. 4.7 can be reformulated as follows:

$$\mathrm{TE}@k(t,q) = \sum_{r=1}^{k} \frac{\frac{\mathrm{tf}(t,d_q^r)}{|d_q^r|}}{\log(r+1)}. \tag{4.8}$$

Group representation. We leverage the term exposure (Eq. 4.8) to estimate the representation of each group using the exposure of its representative terms as follows:

$$p(G_i \mid q,k) = \frac{\sum_{t \in V_{G_i}} \mathrm{TE}@k(t,q)}{\sum_{G_x \in G} \sum_{t \in V_{G_x}} \mathrm{TE}@k(t,q)}. \tag{4.9}$$

Here, G_i represents the group i in the set of N groups indicated with G (e.g., $G = \{\mathit{female}, \mathit{male}\}$), and V_{G_i} stands for the set of terms representing group G_i. The component $\sum_{G_x \in G} \sum_{t \in V_{G_x}} \mathrm{TE}@k(t,q)$ can be interpreted as the total amount of attention that users spend on the representative terms in the ranking for query q. This formulation of the group representation corresponds to the normalization step in AWRF (Eq. 4.5).

Term exposure-based divergence. To evaluate the fairness based on the representation of different groups, we define a fairness criterion built upon our

[1] Referring to the amount of attention an item (document) receives from users in the ranking.

term-level perspective in the representation of groups: in a fairer ranking – one that is less biased – each group of terms receives an amount of attention proportional to their corresponding desired target representation. Put differently, a divergence from the target representations of groups can be used as a means to measure the bias in the ranking. This divergence corresponds to the distance function in Eq. 4.6. Let $\hat{p}_{Gi}$ be the target group representation for each group G_i (e.g., $\hat{p}_{G_i} = \frac{1}{2}$ for $G_i \in \{female, male\}$ for equal representation of male and female), then we can compute the bias in the ranked results retrieved for the query q as the absolute divergence between the groups' representation and their corresponding target representation. We refer to this bias as the *term exposure-based divergence* (TED) for query q:

$$\mathrm{TED}(q,k) = \sum_{G_i \in G} |p(G_i \mid q,k) - \hat{p}_{G_i}|. \tag{4.10}$$

Rank-biased discounting factor (RBDF). With the current formulation of group representation in Eq. 4.9, non-representative documents, i.e., the documents that do not include any group-representative terms, will not contribute to the estimation of bias in TED (Eq. 4.10). To address this issue, we discount the bias in Eq. 4.10 with the proportionality of those documents that count towards the bias estimation, i.e., documents which include at least one group-representative term. To take into account each of these documents with respect to their position in the ranked list, we leverage their corresponding position bias, i.e., $(\log(1+r))^{-1}$ for a document at rank r, to compute the proportionality. The resulting proportionality factor which we refer to as *rank-biased discounting factor* (RBDF) is estimated as follows:

$$\mathrm{RBDF}(q,k) = \frac{\sum_{r=1}^{k} \frac{\mathbb{1}[d_q^r \in S_R]}{\log(1+r)}}{\sum_{r=1}^{k} \frac{1}{\log(1+r)}}. \tag{4.11}$$

Here, S_R stands for the set of representative documents in top-k ranked list of query q, i.e., documents that include at least one group-representative term. Besides, $\mathbb{1}[d_q^r \in S_R]$ is equal to 1 if $d_q^r \in S_R$, otherwise, 0. Accordingly, we incorporate $\mathrm{RDBF}(q,k)$ into Eq. 4.10 and reformulate it as:

$$\mathrm{TED}(q,k) = \sum_{G_i \in G} |p(G_i \mid q,k) - \hat{p}_{G_i}| \cdot \frac{\sum_{r=1}^{k} \frac{\mathbb{1}[d_q^r \in S_R]}{\log(1+r)}}{\sum_{r=1}^{k} \frac{1}{\log(1+r)}}. \tag{4.12}$$

Alternatively, as $\mathrm{TED}(q,k)$ is bounded, we can leverage the maximum value of TED to quantify the fairness of the rank list of query q. We refer to this quantity as *term exposure-based fairness* (TExFAIR) of query q:

$$\mathrm{TExFAIR}(q,k) = \max(\mathrm{TED}) - \mathrm{TED}(q,k). \tag{4.13}$$

In the following, we use TExFAIR to refer to TExFAIR with proportionality (RBDF), unless otherwise stated. With $\hat{p}_{G_i} = \frac{1}{2}$ for $G_i \in \{female, male\}$, TED (Eq. 4.10 and 4.12) falls into the range of [0,1], therefore $\mathrm{TExFAIR}(q,k) = 1 - \mathrm{TED}(q,k)$.

4.4. Experimental Setup

Query sets and collection. We use the MS MARCO Passage Ranking collection [17], and evaluate the fairness on two sets of queries from prior work [24, 147, 205]:

(i) QS1 which consists of 1756 non-gendered queries [147], and

(ii) QS2 which includes 215 bias-sensitive queries [146] (see [147] and [146] respectively for examples).

Ranking models. Following the most relevant related work [146, 205], we evaluate a set of ranking models which work based on pre-trained language models (PLMs). Ranking with PLMs can be classified into three main categories: sparse retrieval, dense retrieval, and re-rankers. In our experiments we compare the following models:

(i) two *sparse retrieval* models: uniCOIL[106] and DeepImpact [113];

(ii) five *dense retrieval* models: ANCE [192], TCT-ColBERTv1 [109], SBERT [145], distilBERT-KD [70], and distilBERT-TASB [71];

(iii) three commonly used cross-encoder *re-rankers*: BERT [126], MiniLM$_{KD}$ [179] and TinyBERT$_{KD}$ [86]. Additionally, we evaluate BM25 [149] as a widely-used traditional lexical ranker [1, 108].

For sparse and dense retrieval models we employ the pre-built indexes, and their corresponding query encoders provided by the Pyserini toolkit [107]. For re-rankers, we use the pre-trained cross-encoders provided by the sentence-transformers library [145].[2] For ease of fairness evaluation in future work, we make our code publicly available at https://github.com/aminvenv/texfair.

Evaluation details. We use the official code available for NFaiRR.[3] Following suggestions in prior work [146], we utilize the whole collection as the background set S (Eq. 4.3) to be able to do the comparison across rankers and re-rankers (which re-rank top-1000 passages from BM25). Since previous instantiations of AWRF cannot be used for the evaluation of term-based fairness of group representations out-of-the-box, we compare TExFAIR to NFaiRR.

4.5. Results

Table 4.1 shows the evaluation of the rankers in terms of effectiveness (MRR and nDCG) and fairness (NFaiRR and TExFAIR). The table shows that almost all PLM-based rankers are significantly fairer than BM25 on both query sets at ranking cut-off 10. In the remainder of this section we address three questions:

(i) What is the correlation between the proposed TExFAIR metric and the commonly used NFaiRR metric?

[2] https://www.sbert.net/docs/pretrained-models/ce-msmarco.html
[3] https://github.com/CPJKU/FairnessRetrievalResults

Method	QS1					QS2				
	MRR	nDCG	NFAIRR	TExFAIR	r	MRR	nDCG	NFAIRR	TExFAIR	r
Sparse retrieval										
BM25	0.1544	0.1958	0.7227	0.7475	0.4823†	0.0937	0.1252	0.8069	0.8454	0.5237†
UniCOIL	0.3276‡	0.3892‡	0.7819‡	0.7629‡	0.5166†	0.2288‡	0.2726‡	0.8930‡	0.8851‡	0.4049†
DeepImpact	0.2690‡	0.3266‡	0.7721‡	0.7633‡	0.5487†	0.1788‡	0.2200‡	0.8825‡	0.8851‡	0.4971†
Dense retrieval										
ANCE	0.3056‡	0.3640‡	**0.7989**‡	**0.7725**‡	0.5181†	0.2284‡	0.2763‡	0.9093‡	**0.9060**‡	0.4161†
DistillBERT$_{KD}$	0.2906‡	0.3488‡	0.7913‡	0.7683‡	0.5525†	0.2306‡	0.2653‡	**0.9149**‡	0.9044‡	0.4257†
DistillBERT$_{TASB}$	0.3209‡	0.3851‡	0.7898‡	0.7613‡	0.5091†	0.2250‡	0.2725‡	0.9088‡	0.8960‡	0.4073†
TCT-ColBERTv1	0.3138‡	0.3712‡	0.7962‡	0.7688‡	0.5253†	0.2300‡	0.2732‡	0.9116‡	0.9056‡	0.4249†
SBERT	0.3104‡	0.3693‡	0.7880‡	0.7637‡	0.5217†	0.2197‡	0.2638‡	0.8943‡	0.8999‡	0.3438†
Re-rankers										
BERT	0.3415‡	0.4022‡	0.7790‡	0.7584‡	0.5135†	0.2548‡	0.2950‡	0.8896‡	0.8807‡	0.4323†
MiniLM$_{KD}$	**0.3832**‡	**0.4402**‡	0.7702‡	0.7516	0.5257†	**0.2872**‡	**0.3323**‡	0.8863‡	0.8865‡	0.3880†
TinyBERT$_{KD}$	0.3482‡	0.4093‡	0.7799‡	0.7645‡	0.5437†	0.2485‡	0.3011‡	0.8848‡	0.8952‡	0.4039†

Table 4.1: Effectiveness and fairness results at ranking cut-off = 10. r denotes the correlation between TExFAIR and NFaiRR. Higher values of TExFAIR and NFaiRR correspond to higher fairness. † denotes statistical significance for correlations with ($p < 0.05$). ‡ indicates statistically significant improvement over BM25 according to a paired t-test ($p < 0.05$). Bonferroni correction is used for multiple testing.

(ii) What is the sensitivity of the metrics to the ranking cut-off?

(iii) What is the relationship between the bias in ranked result lists of rankers, and how sensitive they are towards the concept of gender?

(i) Correlation between metrics. To investigate the correlation between the TExFAIR and NFaiRR metrics, we employ Pearson's correlation coefficient on the query level. As the values in Table 4.1 indicate, the two metrics are significantly correlated, but the relationship is not strong ($0.34 < r < 0.55$). This is likely due to the fact that NFaiRR and TExFAIR are structurally different: NFaiRR is document-centric: it estimates the fairness in the representation of groups on a document-level and then aggregates the fairness values over top-k documents. TExFAIR, on the other hand, is ranking-centric: each group's representation is measured based on the whole ranking, instead of individual documents. As a result, in a ranked list of k documents, the occurrences of the terms from one group at rank i, with $i \in \{1,\ldots,k\}$, can balance and make up for the occurrences of the other group's terms at rank j, with $j \in \{1,\ldots,k\}$. This is in contrast to NFaiRR in which the occurrences of the terms from one group at rank i, with $i \in \{1,\ldots,k\}$, can only balance and make up for the occurrences of other group's terms at rank i. Thus, TExFAIR measures a different dimension of fairness than NFaiRR.

(ii) Sensitivity to ranking cut-off k. Figure 4.2 depicts the fairness results at various cut-offs using TExFAIR with and without proportionality (RBDF) as well as the results using NFaiRR. The results using TExFAIR without proportionality show a high sensitivity to the ranking cut-off k in comparison to the other two metrics. The reason is that without proportionality factor RDBF, the unbiased documents with zero group-representative term, i.e., non-representative documents, do not count towards the fairness evaluation. As a result, regardless of the number of this kind of unbiased documents, documents that include group-representative terms potentially can highly affect the fairness of the ranked list. On the contrary, NFaiRR and TExFAIR with proportionality factor are less sensitive to the ranking cut-off: the effect of unbiased documents with zero group-representative term is addressed in NFaiRR with a maximum neutrality for these documents (Eq. 4.1), and in TExFAIR with proportionality factor RBDF by discounting the bias using the proportion of documents that include group-representative terms (Eq. 4.12).

(iii) Counterfactual evaluation of gender fairness. TExFAIR and NFaiRR both measure the fairness of ranked lists produced by ranking models. Next, we perform an analytical evaluation to measure the extent to which a ranking model acts indifferently (unbiasedly) towards the genders, regardless of the fairness of the ranked list it provides. Our evaluation is related to counterfactual fairness measurements which require that the same outcome should be achieved in the real world as in the term-based counterfactual world [134, 188]. Here, the results of the real world correspond to the ranked lists that are returned using the original documents, and results of the counterfactual world correspond to the ranked lists that are returned using counterfactual documents.

In order to construct counterfactual documents, we employ counterfactual data substitution (CDS) [110, 114], in which we replace terms in the collection with their counterpart in the opposite gender-representative terms, e.g., “he” → “she,” “son” → “daughter,” etc. For names, e.g., Elizabeth or John, we substitute them with a name from the opposite gender name in the gender-representative terms [114]. Additionally, we utilize POS information to avoid ungrammatically assigning “her” as a personal pronoun or possessive determiner [114]. We then measure how the ranked result lists of a ranking model on a query set Q would diverge if a ranker performs the retrieval on the counterfactual collection rather than the original collection.

In order to measure the divergence, we employ rank-biased overlap (RBO) [182] as a measure to quantify the similarities between two ranked lists. We refer to this quantity as *counterfactually-estimated rank-biased overlap* (CRBO). RBO ranges from 0 to 1, where 0 represents disjoint and 1 represents identical ranked lists. RBO has a parameter $0 < p \leq 1$ which regulates the degree of top-weightedness in estimating the similarity. From another perspective, p represents searcher patience or persistence and larger values of p stand for more persistent searching [38]. Since we focus on top-10 ranked results, we follow the original work [182] for a reasonable choice of p, and set it to 0.9 (see [182] for more discussion).

Table 4.2 shows the CRBO results. While there is a substantial difference in the fairness of ranked results between the BM25 and the PLM-based rankers, the CRBO results of these models are highly comparable, and even BM25, as the model which provides the most biased ranked results, is the least biased model in terms of CRBO

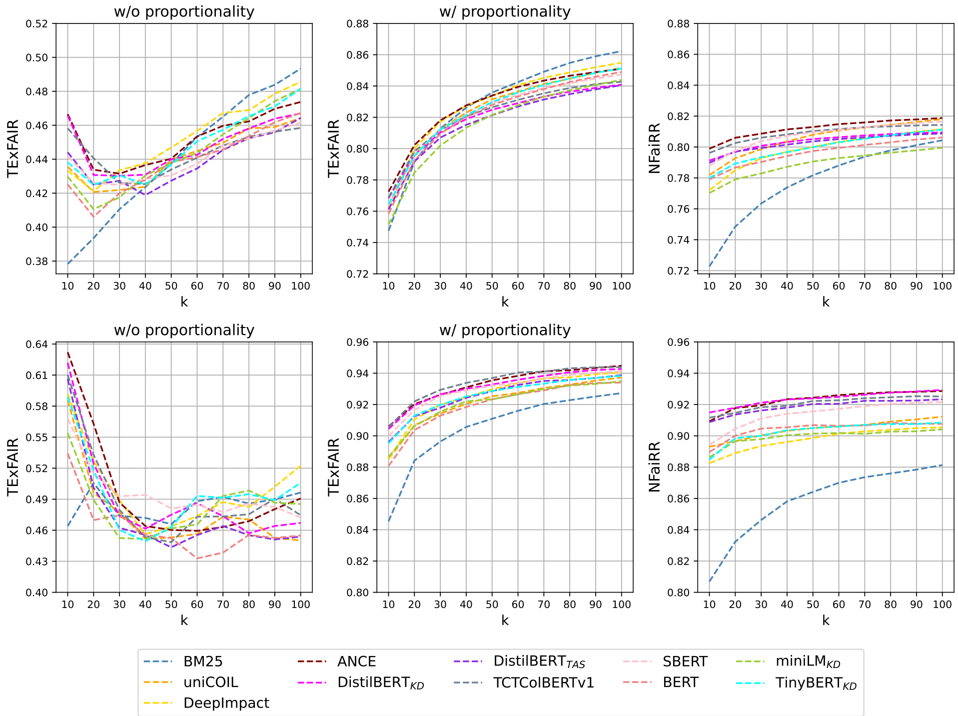

Figure 4.2: Fairness results on QS1 (first row) and QS2 (second row) at different ranking cut-off values (k).

on QS1. Additionally, among PLM-based rankers, the ones with higher TExFAIR or NFaiRR scores do not necessarily provide higher CRBO. This discrepancy between {NFaiRR, TExFAIR} and CRBO disentangles the bias of a model towards genders from the bias of the ranked results it provides. However, it should be noted that we indeed cannot come to a conclusion as to whether the bias that exists in the PLM-based rankers (the one that is reflected by CRBO) does not contribute to their superior fairness of ranked results (the one that is reflected by {NFaiRR, TExFAIR}). We leave further investigation of the quantification of inherent bias of PLM-based rankers and its relation with the bias of their ranked results for future work.

4.6. DISCUSSION

The role of non-representative documents. As explained in Section 4.3, and based on the results in Section 4.5, discounting seems to be necessary for the evaluation of gender fairness in document ranking with group-representative terms, due to the effect of non-representative documents. Here, one could argue that without our proposed proportionality discounting factor (Section 4.3), it is possible to use an association value for d_q^r to group G_i, i.e., $a_{d_q^r}^{G_i}$ in the formulation of AWRF (Section

	QS1			QS2		
Models	CRBO	TExFAIR	NFaiRR	CRBO	TExFAIR	NFaiRR
BM25	**0.9733**	0.8454	0.8069	**0.9761**	0.7475	0.7227
BERT	0.9506	0.8807	**0.8896**	0.9735	0.7629	0.7790
MiniLM	0.9597	0.8865	0.8863	0.9753	0.7516	0.7702
TinyBERT	0.9519	**0.8952**	0.8848	0.9714	**0.7645**	**0.7799**

Table 4.2: Counterfactually-estimated RBO results. For ease of comparison, TExFAIR and NFaiRR results are included from Table 4.1.

4.2) as follows:

$$a^{G_i}_{d^r_q} = \frac{M^{G_i}(d^r_q)}{\sum_{G_x \in G} M^{G_x}(d^r_q)}, \tag{4.14}$$

4

and simply assign equal association for each group, e.g., $a^{G_i}_{d^r_q} = \frac{1}{2}$ for $G_i \in \{$*female*, *male*$\}$ for documents that do not contain group-representative terms, i.e., non-representative documents. However, we argue that such formulation results in the ignorance of the frequency of group-representative terms. For instance, intuitively, a document which has only one mention of a female name as a female-representative term (therefore is completely biased towards female) and is positioned at rank i, cannot simply compensate and balance for a document with high frequency of male-representative names and pronouns (completely biased towards male) and is positioned at rank $i+1$. However, with the formulation of document associations in AWRF (Eq. 4.14) these two documents can roughly[4] balance for each other. As such, there is a need for a fairness estimation in which the frequency of terms is better counted towards the final distribution of groups. Our proposed metric TExFAIR implicitly accounts for this effect by performing the evaluation based on term-level exposure estimation and incorporating the rank biased discounting factor RBDF.

Limitations of CRBO. While measuring gender bias with counterfactual data substitution is widely used for natural language processing tasks [42, 63, 114, 152], we believe that our analysis falls short of thoroughly measuring the learned stereotypical bias. We argue that through the pre-training and fine-tuning step, specific gendered correlations could be learned in the representation space of the ranking models [183]. For instance, the representation of the word “nurse” or “babysitter” might already be associated with female group terms. In other words, the learned association of each term to different groups (either female or male), established during pre-training or fine-tuning, is a spectrum rather than binary. As a result, these kinds of words could fall at different points of this spectrum and therefore, simply replacing a limited number of gendered-terms (which are assumed to be the two end point of this spectrum) with their corresponding counterpart in the opposite gender group, might not reflect the actual inherent bias of PLM-based rankers towards different groups of gender. Moreover, while we estimate CRBO based on the divergence of the results on the original collection and a single counterfactual

[4] As they have different position bias.

collection, more stratified counterfactual setups can be studied in future work.

Reflection on evaluation with term-based representations. We acknowledge that evaluating fairness with term-based representations is limited in comparison to real-world user evaluations of fairness. However, this shortcoming exists for all natural language processing tasks where semantic evaluation from a user's perspective might not exactly match with the metrics that work based on term-based evaluation. For instance, there exists a discussion over the usage of BLEU [132] and ROUGE [105] scores in the evaluation of natural language generation [169, 208]. Nevertheless, such an imperfect evaluation method is still of great importance due to the sensitivity of the topic of societal fairness and the impact caused by the potential consequences of unfair ranking systems. We believe that this chapter addresses an important aspect of evaluation in the current research in this area and plan to work on more semantic approaches of societal fairness evaluation in the future.

4.7. CONCLUSION

In this chapter, we addressed the evaluation of societal group bias in document ranking. We pointed out an important limitation of the most commonly used group fairness metric NFaiRR, which measures fairness based on a fairness score of each ranked document. Our newly proposed metric TExFAIR integrates two extensions on top of a previously proposed generic metric AWRF: the term-based association of documents to each group, and a rank biased discounting factor that addresses the impact of non-representative documents in the ranked list. As it is structurally different, our proposed metric TExFAIR measures a different aspect of the fairness of a ranked list than NFaiRR. Hence, when fairness is taken into account in the process of model selection, e.g., with a combinatorial metric of fairness and effectiveness [146], the difference between the two metrics TExFAIR and NFaiRR could result in a different choice of model.

In addition, we conducted a counterfactual evaluation, estimating the inherent bias of ranking models towards different groups of gender. With this analysis we show a discrepancy between the measured bias in the ranked lists (with NFaiRR or TExFAIR) on the one hand and the inherent bias in the ranking models themselves on the other hand. In this regard, for our future work, we plan to study more semantic approaches of societal fairness evaluation to obtain a better understanding of the relationship between the inherent biases of ranking models and the fairness (unbiasedness) of the ranked lists they produce. Moreover, since measuring group fairness with term-based representations of groups is limited (compared with the real-world user evaluation of fairness), we intend to work on more user-oriented methods for the measurement of societal fairness in the ranked list of documents.

5

Evaluation of Attribution Bias in Generator-Aware Retrieval-Augmented Large Language Models

Attributing answers to source documents is an approach used to enhance the verifiability of a model's output in retrieval-augmented generation (RAG). Prior work has mainly focused on improving and evaluating the attribution quality of large language models (LLMs) in RAG, but this may come at the expense of inducing biases in the attribution of answers. We define and examine two aspects in the evaluation of LLMs in RAG pipelines, namely attribution sensitivity and bias with respect to authorship information. We explicitly inform an LLM about the authors of source documents, instruct it to attribute its answers, and analyze (i) how sensitive the LLM's output is to the author of source documents, and (ii) whether the LLM exhibits a bias towards human-written or AI-generated source documents. We design an experimental setup in which we use counterfactual evaluation to study three LLMs in terms of their attribution sensitivity and bias in RAG pipelines. Our results show that adding authorship information to source documents can significantly change the attribution quality of LLMs by 3 to 18%. We show that LLMs can have an attribution bias towards explicit human authorship, which can serve as a competing hypothesis for findings of prior work that shows that LLM-generated content may be preferred over human-written contents. Our findings indicate that metadata of source documents can influence LLMs' trust, and how they attribute their answers. Furthermore, our research highlights attribution bias and sensitivity as a novel aspect of the brittleness of LLMs.

5.1. Introduction

The goal of retrieval-augmented generation (RAG) is to generate an answer to a given question using a set of top-k retrieved documents as context [98]. Large language

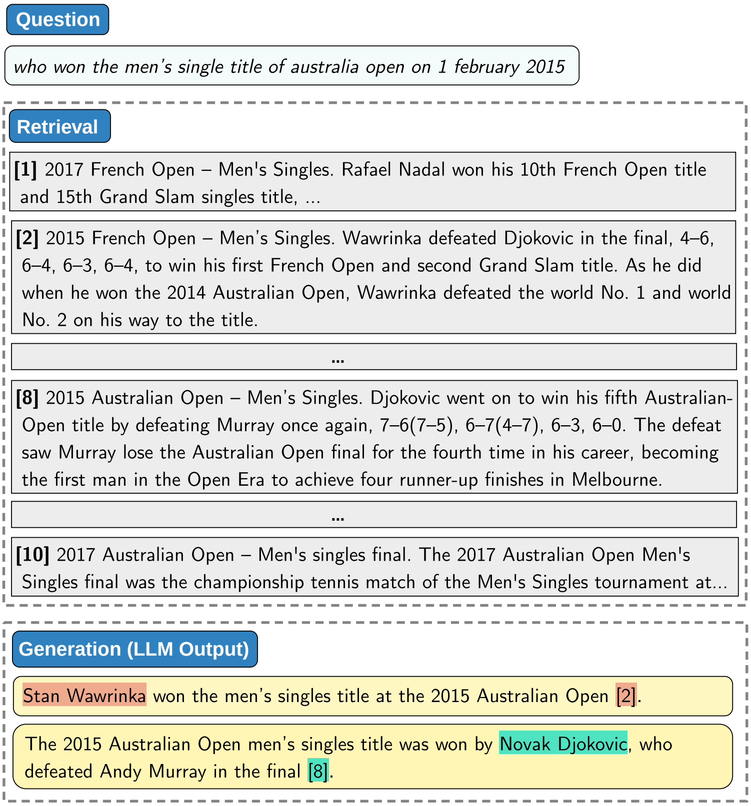

Figure 5.1: Retrieval-augmented answer/attribution generation using two LLMs. Together with the question, retrieval results are given to the LLMs in order to generate the answer.

models (LLMs) have been a crucial part of RAG pipelines, mainly as the generator component [9, 82, 97, 100]. Although the use of LLMs offers potential benefits, it also presents considerable risks, as they are prone to generate false or hallucinated claims [83]. This is important as such claims may misguide users, particularly when they are being used in critical fields such as the legal or medical domain [16, 112, 191].

Enabling LLMs to attribute their answer to the source of information has been proposed as a promising direction towards reducing the likelihood of such potential harms [101, 102, 133]. This attribution could assist users in tracing and understanding the basis of the information that LLMs are generating [61, 76]. There are many prior studies on answer attribution in RAG pipelines [26, 74, 101, 117, 121, 167].

As Figure 5.1 illustrates, LLMs are susceptible to making mistakes when attributing their answers to the input documents in RAG. Moreover, enabling LLMs in RAG to attribute their answer may come at the expense of inducing biases, as LLMs may carry potential biases [53, 130, 189, 212]. For instance, Tan, Sun, Yang, Wang, Cao, and Cheng (2024) show that retrieval-augmented LLMs can be biased towards selecting their own generated text when this kind of content is present in their input. Inspecting these biases is of paramount importance as they can be leveraged for both positive and negative purposes.

In this chapter, we study the performance of LLMs in terms of *attribution sensitivity* and *attribution bias* w.r.t. authorship information. When we explicitly inform LLMs about the authors of input documents, and instruct them to attribute their answers to the input documents (by providing citations to these documents), how sensitive are they to the authorship information of input documents? And are they biased towards either human or LLM authorship of input documents? To address these questions, we design a simulated evaluation setup in which we measure to what extent knowing the type of author of input documents affects the quality of attribution (citation) in LLMs.

Prior work has indicated that LLM-generated content may consistently outperform human-authored content in search rankings, which, in turn, results in reducing the presence and exposure of human contributions online [35, 43]. Inspired by these studies, we compare human-written documents against LLM-generated documents. We follow prior work in attribution generation by prompting LLMs to generate citations to the input documents [61, 199]. Furthermore, we use counterfactual evaluation [2, 65, 73, 78, 190] to measure both authorship sensitivity and authorship bias of LLMs in RAG pipelines. This approach can be used more generally to measure algorithmic sensitivity or bias in a model or method: using counterfactual scenarios to see if changing certain characteristics leads to different outcomes.

Our experimental results show that three LLMs (`Mistral`, `Llama3` and `GPT-4`) are sensitive to authorship information that is included in the input documents prior to the generation. Moreover, we show that these models carry a bias towards human authorship against LLM authorship: they are more likely to attribute their answers to documents that are explicitly labelled as having been written by humans (even if the documents are actually generated by LLMs). We summarize our contributions as follows:

- We define and study attribution sensitivity and bias w.r.t. authorship information, as a novel aspect of trustworthiness and brittleness in retrieval-augmented LLMs.
- We propose a systematic evaluation framework for measuring attribution sensitivity and bias.
- We show that adding authorship information (as metadata) to source documents may lead to statistically significant changes in the attribution quality of retrieval-augmented LLMs.
- We show that LLMs may have an attribution bias towards explicit human *authorship*, which can serve as a competing hypothesis for findings of prior work that shows that LLM-generated *content* is preferred over human-written *content* by LLMs.

5.2. BACKGROUND

Retrieval-Augmented Generation. Given a question q and a set of top-k retrieved documents $\mathcal{D} = \{d_1, d_2, \ldots, d_k\}$ from a collection $\mathcal{C}$, the goal of retrieval-augmented generation (RAG) is to generate an answer for q using $\mathcal{D}$ as context. LLMs are currently an important component of RAG pipelines, acting as the generator. The generator is given q, $\mathcal{D}$, and an instruction prompt on how to generate the answer [82, 97, 100]. Using top-k retrieved documents helps LLMs to be exposed to information that it might not have been trained/fine-tuned with during development. These documents are commonly retrieved using an effective sparse and/or dense retriever [98, 143].

Attributive RAG. LLMs are prone to generate hallucinated (and even factually incorrect) answers [83, 144, 200]. Attributing answers in RAG with LLMs is an approach taken as a step towards ensuring the veracity of the output of these models [26, 74, 87, 90, 101]. Menick, Trebacz, Mikulik, Aslanides, Song, Chadwick, Glaese, Young, Campbell-Gillingham, Irving, *et al.* (2022) teach language models to support answers with verified quotes. Ye, Sun, Arik, and Pfister (2024) propose a learning-based framework in which they fine-tune LLMs to generate citations, as opposed to prompting or relying on post-hoc attribution. Stolfo (2024) analyzes whether every generated sentence in the output of LLMs is grounded in the retrieved documents or the LLM's pre-training data.

5.3. METHODOLOGY

We aim to measure the attribution sensitivity and bias of LLMs in RAG settings. We investigate to what extent the attribution quality of LLMs is affected by authorship information. To this end, we use counterfactual evaluation [29, 62, 180]. Counterfactual evaluation has been used across various natural language processing and information retrieval tasks [2, 5, 65, 73, 78]. This approach evaluates how a model's predictions change when a specific feature or set of features is altered while keeping everything else constant. In our case, the change is to add authorship information to the input documents of an LLM in a RAG setting. By doing so, we can evaluate the model's reliance on, bias towards, or sensitivity to that feature. To this end, we first generate answers with LLMs in a RAG setting using three RAG modes, as shown in Figure 5.3.

5.3.1. RAG MODES

Given a query q and a set of top-k retrieved documents $\mathcal{D}_q$ for q, we define three modes, based on authorship information of these documents that we provide to the answer generator LLM.

Vanilla RAG. In this mode, each document in $\mathcal{D}$ is given to the LLMs without information about who the authors are. This is the plain input format for input documents as shown in the input prompt for *vanilla* answer/attribution generation in Figure 5.2.

Authorship Informed RAG. In this mode, we inform the LLM about the actual author of each document. We denote the authorship of either an LLM or a human using `[LLM]` and `[Human]` tokens as shown by Figure 5.7 in the Appendix.[1]

Counterfactual-Authorship Informed RAG. In this mode, we assign counterfactual authorship for each document. If a document is written by a human, the counterfactual authorship of this document is `[LLM]`. In contrast, if a document is generated by an LLM, its counterfactual authorship is `[Human]`. By doing so, we can investigate to what extent being written by either human or LLM affects the attribution quality of LLM. The prompt used for this mode is the same as the one for Authorship Informed RAG mode.

Figure 5.3 shows the three RAG modes for a setting where the relevant documents are LLM-written and the non-relevant documents are human-written.

```
Instruction: Write a concise answer for the given question (query)
based on the provided search result documents, and cite them
properly using [1][2][3] etc.

Please take these strict considerations into account during
answer generation:
1. Documents are retrieved by a search engine. As such, not all
the documents are relevant to the query. Only use and cite the
relevant documents that contain the answer.
2. Do not analyze irrelevant documents.

Search Results:

Document [1]({text of Document [1]})
Document [2]({text of Document [2]})
...
Document [10]({text of Document [10]})

Question: {query}.
```

Figure 5.2: Prompt used for vanilla retrieval-augmented answer generation.

5.3.2. ANSWER/ATTRIBUTION GENERATION

In order to generate answers with each of the aforementioned RAG modes, we experiment with three LLMs: `Mistral` [84], `Llama3` [49] and `GPT-4` [128]. Figure 5.2 shows the prompt used for *vanilla* answer generation. Figure 5.7 in the Appendix shows the prompt used for *Authorship-Informed* and *Counterfactual-Authorship Informed* answer generation. We follow prior work [61] in curating our prompts for this task.

[1]In Section 5.C in the Appendix, we study and provide results on replacing `[Human]` with a set of actual {firstname, lastname} as authors.

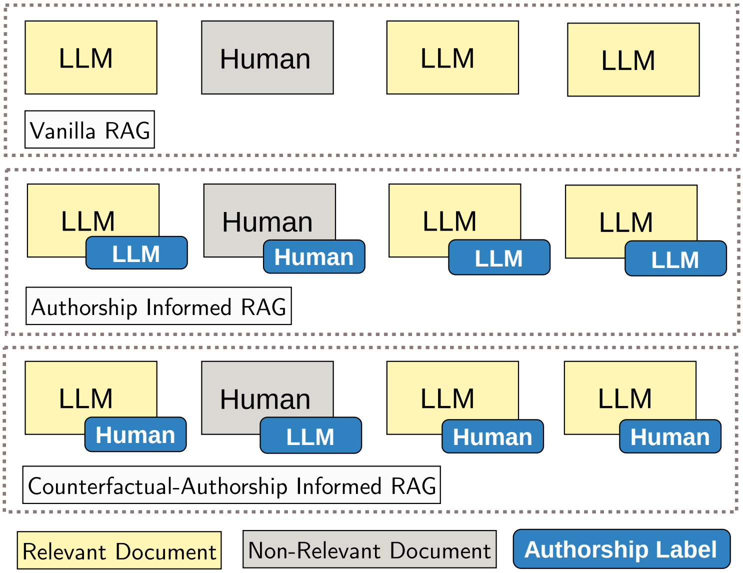

Figure 5.3: Three RAG modes (Section 5.3.1) for the setting with LLM actual authorship for relevant documents and Human actual authorship for non-relevant documents. The text in a rectangle denotes the actual generator (i.e., author) of each document. The text in the blue tags denotes the authorship label about which we inform the answer/attribution generator LLM.

5.3.3. Evaluation Metrics

Attribution Quality. We use precision and recall for evaluating the quality of attribution, i.e., how well the LLMs cite the relevant input documents. Precision of attribution for a single query is the fraction of correct citations among all cited documents in the output of an LLM. Recall is the fraction of cited relevant documents out of all relevant documents [47]. We use the queries that have only one relevant document containing the ground-truth answer in their top-k retrieved list of documents.

Attribution Sensitivity. In order to measure the sensitivity of LLMs in RAG pipelines towards knowing authors of input documents in comparison to not knowing it, we use counterfactual evaluation and define a metric called Counterfactually-estimated Attribution Sensitivity (CAS):

$$\mathrm{CAS}(Q) = \frac{1}{|Q|} \sum_{q \in Q} |M^q_{\text{Informed}} - M^q_{\text{Vanilla}}|. \tag{5.1}$$

Here, M^q represents the precision and recall metrics for query q, i.e., attribution quality for query q. For a single query q, CAS measures the difference between a base setup (the vanilla RAG mode) and a counterfactual setup (the authorship informed RAG mode) for the same set of input documents.

Attribution Bias. In order to measure the attribution bias of LLMs in RAG pipelines we use counterfactual evaluation and define a metric called Counterfactually-estimated Attribution Bias (CAB):

$$\mathrm{CAB}(Q) = \frac{\omega}{|Q|} \sum_{q \in Q} \left(M^q_{\text{Informed}} - M^q_{\text{CF-informed}} \right) \tag{5.2}$$

$$\omega = \begin{cases} 1, & \text{if } L_f(\mathscr{R}) = \texttt{[Human]}, L_f(\mathscr{N}) = \texttt{[LLM]} \\ -1, & \text{otherwise.} \end{cases} \tag{5.3}$$

Here, M^q represents the precision and recall metrics, i.e., attribution quality, for query q, given the set of retrieved relevant documents $\mathscr{R}$, and the set of retrieved non-relevant documents $\mathscr{N}$. $L_f(\mathscr{X})$ stands for the authorship label of the set of documents $\mathscr{X}$ in the first term of Eq. 5.2, i.e., corresponding to M^q_{Informed}. For example, if we use human-written version of relevant documents ($\mathscr{R}$), and LLM-written version of non-relevant document ($\mathscr{N}$), and we label them with their actual generators (authors), i.e., we use authorship-informed RAG mode, then $L_f(\mathscr{R})$ is equal to `[Human]`, and $L_f(\mathscr{R})$ is equal to `[LLM]`. CAB measures the difference between metric values of a base setup (the Authorship Informed RAG mode) and a counterfactual setup (the Counterfactual-authorship Informed RAG mode) for the same set of input documents consisting of $\mathscr{R}$, and $\mathscr{N}$. ω determines the direction of bias towards either human or LLMs: if the set of relevant documents ($\mathscr{R}$) and non-relevant documents ($\mathscr{N}$) are respectively written by `Human` and `LLM` (i.e., $L_f(\mathscr{R})$ = `[Human]`, $L_f(\mathscr{N})$ = `[LLM]`), for a single query, a positive difference ($M_{\text{Informed}} - M_{\text{CF-informed}}$) indicates bias towards human authorship, and a negative difference shows bias towards LLM authorship. In contrast, if the set of relevant documents ($\mathscr{R}$) and non-relevant documents ($\mathscr{N}$) are respectively written by `LLM` and `Human` (i.e., $L_f(\mathscr{R})$ = `[LLM]`, $L_f(\mathscr{N})$ = `[Human]`), a negative difference ($M_{\text{Informed}} - M_{\text{CF-informed}}$) indicates a bias towards human authors, and a positive difference shows bias towards LLMs. We use ω to align these two conditions of actual authorship of input documents.

Attribution Confidence. To better explore the performance of LLMs in attribution generation, we analyze whether the LLMs are more confident when they attribute to certain types of document. To this aim, we look into the average probability of generation for attribution tokens, i.e., citation numbers (0, 1, ...):

$$\mathrm{AC}(\mathscr{S}) = \frac{\sum_{q \in Q} \sum_{c_i \in C_q} p(c_i \mid q, \mathscr{D}_q) \cdot \mathbb{1}[c_i \in \mathscr{S}]}{\left| \sum_{q \in Q} \sum_{c_i \in C_q} \mathbb{1}[c_i \in \mathscr{S}] \right|}, \tag{5.4}$$

where q is a query in the set of queries Q, $\mathscr{D}_q$ is the top-k retrieved documents for q, C_q stands for all attribution numbers in the answer to q_i, $c_i \in \{0, 1, \ldots, k\}$, $\mathscr{S}$ indicates a set of documents, e.g., the set of relevant documents for all queries, and $p(c_i | q, \mathscr{D}_q)$ shows the probability of generation for the attribution token c_i in the answer provided by LLM given query q and its top-k retrieved documents $\mathscr{D}_q$. $\mathbb{1}[c_i \in \mathscr{S}]$ equals 1 if $c_i \in \mathscr{S}$.

Answer Correctness. In order to evaluate the quality of the generated answer, we follow [61, 136] and use automatic evaluation. Following [61, 167], we use the normalized human-generated answer in the benchmark as the ground-truth answer and adopt Exact Match (EM) [162, 181] as the evaluation metric for answer correctness (see example in Figure 5.16).

5.4. EXPERIMENTAL SETTINGS

Models. We use `gpt-4-0314` [128], `meta-llama/Meta-Llama-3-8B-Instruct`,[2] and `mistralai/Mistral-7B-Instruct-v0.3`[3] as answer generator LLMs in our RAG pipelines. We refer to these models as `GPT-4`, `Llama3`, and `Mistral`, respectively.
Benchmarks. We use two benchmarks in our experiments: Natural Questions (NQ) [95] and MS MARCO Question Answering [17] (to which we refer as MS MARCO). For each benchmark, we randomly sample 500 queries. To retrieve top-k passages for each query in the NQ benchmark, we use BM25, a widely-used lexical matching retrieval model. For queries in the MS MARCO benchmark, we use passages that are extracted from relevant web documents using the state-of-the-art passage retrieval system at Bing [17]. We note that we study the effect of different retrievers and different number of retrieved source documents in Section 5.D and 5.E in the Appendix, respectively.

5

Synthetic Collection. To construct LLM-written documents, we use `Llama3` to re-write a given document from our collections without adding/removing information to/from the document. Specifically, we use a low temperature close to 0 as it makes the LLM extremely restrictive, focusing only on the most probable tokens resulting in (highly) deterministic outputs. The reason for not generating the documents from scratch is to make sure we keep the relevance/non-relevance status of documents w.r.t a query. To ensure the quality of synthetic passages, we conduct a number of annotation steps using two expert annotators. This is detailed in Section 5.A in the Appendix. Importantly, in Section 5.5, we show that even without using LLM-generated documents (i.e., only designating `[Human]` and `[LLM]` as authors of documents to the original input documents) findings are the same as when we use actual LLM-generated documents.

5.5. EXPERIMENTAL RESULTS

In this section, we explore the performance of LLMs for attributing their answer to top-k retrieved source documents using the evaluation metrics introduced in Section 5.3. All significance tests in the result tables are according to a paired t-test with $p < 0.05$.

[2] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
[3] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

Answer generator	Relevant documents	Non-relevant documents	RAG mode	Attribution quality Precision	Attribution quality Recall	Correctness EM
NQ						
Mistral	LLM	Human	Vanilla	47.6	76.6	0.722
			Informed	42.1	68.2	0.730
			CF-informed	**52.7**†	**77.8**†	**0.738**
	Human	LLM	Vanilla	51.0	78.4	**0.776**
			Informed	**53.4**†	**77.8**†	0.774
			CF-informed	44.0	70.2	0.772
Llama3	LLM	Human	Vanilla	49.2	69.2	0.742
			Informed	45.4	69.6	0.730
			CF-informed	**57.2**†	**77.6**†	**0.748**
	Human	LLM	Vanilla	53.5	71.0	0.766
			Informed	**59.9**†	**77.8**†	**0.790**
			CF-informed	44.8	69.2	0.762
GPT-4	LLM	Human	Vanilla	63.3	68.8	0.736
			Informed	59.7	64.6	0.740
			CF-informed	**65.9**†	**72.2**†	**0.742**
	Human	LLM	Vanilla	64.1	68.8	0.760
			Informed	**66.1**	**72.2**†	**0.776**
			CF-informed	60.3	65.0	0.758

Table 5.1: Quality of attribution and answer correctness. The columns "Relevant Documents" and "Non-relevant Documents" refer to the actual authorship of input documents. Informed refers to the authorship-informed RAG and CF-informed refers to counterfactual-authorship informed RAG (Section 5.3.1). † indicates statistically significant improvements over the two other RAG Modes in each combination of "Relevant" and "Non-relevant" documents.

Attribution Quality. Table 5.1 shows the results of attribution by three LLMs, `Mistral`, `Llama3` and `GPT-4`, under different settings for NQ benchmark. Besides, Table 5.11 in the Appendix shows the same set of results for the MS MARCO benchmark. The two columns "Relevant documents" and "Non-relevant documents" indicate the actual generator (author) of these documents. The column "RAG mode" indicates how we inform the answer generator LLMs about the authorship label of relevant and non-relevant documents, as described in Section 5.3.1: in the "Vanilla" RAG mode, no information regarding the generator (author) of the input source documents is given to the LLM. In the "Informed" RAG mode the LLM is informed about the actual generator of the input source documents, and in the "CF-Informed" RAG mode the LLM is provided with counterfactual authorship information. As Table 5.1 shows, the three LLMs (`Mistral`, `Llama3` and `GPT-4`) fall short of perfectly grounding their answers to the relevant documents of a given question, which is in line with the findings of prior work [47, 61, 102].

Attribution Sensitivity and Bias. Table 5.3 shows the attribution bias results in

Answer generator	Relevant documents	Non-relevant documents	CAS ΔPrecision	CAS ΔRecall
NQ				
Mistral	LLM	Human	16.2†	17.2†
	Human	LLM	20.1	17.0
Llama3	LLM	Human	13.2†	14.8
	Human	LLM	17.7†	16.0†
GPT-4	LLM	Human	9.7†	10.2†
	Human	LLM	8.7	9.0†
MS MARCO				
Mistral	LLM	Human	10.9	21.4†
	Human	LLM	12.9†	16.6
Llama3	LLM	Human	12.9†	20.4†
	Human	LLM	17.8†	19.6†
GPT-4	LLM	Human	8.2†	9.6†
	Human	LLM	10.9	15.8†

Table 5.2: Attribution sensitivity (CAS) results. Values range from 0 (minimum sensitivity) to 100 (maximum sensitivity). † indicates statistically significant values.

Answer generator	Relevant documents	Non-relevant documents	CAB ΔPrecision	CAB ΔRecall
NQ				
Mistral	LLM	Human	+10.6†	+9.6†
	Human	LLM	+9.4†	+7.6†
Llama3	LLM	Human	+11.8†	+8.0†
	Human	LLM	+15.1†	+8.6†
GPT-4	LLM	Human	+6.2†	+7.6†
	Human	LLM	+5.8†	+7.2†
MS MARCO				
Mistral	LLM	Human	+9.5†	+13.8†
	Human	LLM	+8.0†	+12.4†
Llama3	LLM	Human	+15.6†	+18.2†
	Human	LLM	+15.1†	+16.4†
GPT-4	LLM	Human	+6.1†	+9.0†
	Human	LLM	+5.4†	+10.8†

Table 5.3: Attribution Bias (CAB) results. Values range from -100 (completely biased towards LLM authorship) to +100 (completely biased towards human authorship). † indicates statistically significant bias values.

terms of CAB (Eq. 5.2). All three LLMs, `Mistral`, `Llama3`, and `GPT-4`, carry a bias towards human authorship in the input documents. Moreover, on both datasets, NQ and MS MARCO, `Mistral` and `Llama3` have higher bias values than `GPT-4`. Besides, Table 5.2 shows the attribution sensitivity results in terms of CAS (Eq. 5.1). All three LLMs, `Mistral`, `Llama3`, and `GPT-4`, show sensitivity towards the inclusion of authorship information for the input documents of LLMs. Moreover, similar to the attribution bias values in Table 5.3, `Mistral` and `Llama3` carry a higher attribution sensitivity than `GPT-4`. We note that we conducted experiments using different prompts and observed that the findings remained consistent across multiple runs.

Mixed RAG Mode. To better disentangle the effect of LLM generated text qualities (e.g., a potential implicit bias of LLMs towards LLM-written documents [174]) from the impact of authorship information, we now use the same set of documents in the input of LLM in the RAG, and only use different authorship labels for relevant and non-relevant documents. For this new setup, to which we refer as the Mixed RAG mode, we evaluate both a complete set of synthetic documents (i.e., for both relevant and non-relevant) and a complete set of human-written documents. Figure 5.4 shows an example of Mixed RAG mode for the setting where we have human-written documents, with different authorship labels for relevant and non-relevant documents. The CAB (Eq. 5.2) for Mixed RAG mode is reformulated as follows:

$$\mathrm{CAB}(Q) = \frac{\omega}{|Q|} \sum_{q \in Q} M^q_{\text{Informed/CF-Informed}} - M^q_{\text{CF-Informed/Informed}}, \tag{5.5}$$

where X and Y in $M^q_{X/Y}$ stand for the RAG mode for the set of relevant documents and the set of non-relevant documents, respectively. The results of attribution quality for Mixed-RAG modes are shown in Table 5.4.[4] We see that, similar to Table 5.1, across different settings, when the relevant documents are labeled with human-authorship and non-relevant ones are labeled with LLM-authorship, the attribution quality is higher than the other way around. Moreover, Table 5.5 illustrates the attribution bias for Mixed RAG modes. Similar to the results in Table 5.3, there is a bias towards human authorship in all three LLMs. This indicates the existence of authorship bias regardless of the origin of the input documents, i.e., the actual author of the input documents. Furthermore, similar to the results in Table 5.3, `Mistral` and `Llama3` show higher rates of attribution bias than `GPT-4`. Additionally, we find that when we have the same authorship label on both relevant and non-relevant documents (rows with the same RAG mode for relevant and non-relevant documents in Tables 5.14 and 5.15 in the Appendix), we do not see consistent patterns as to how LLMs attribute the answers to the input documents. Finally, we note that in Section 5.C of the Appendix, we show additional results using real-world names as authors which further indicates the presence of attribution bias and sensitivity in LLMs towards authorship information.

Attribution Confidence. Using Eq. 5.4, we compute the confidence of LLMs when they attribute their answer to an input document. Table 5.6 shows the attribution confidence of LLMs for relevant and non-relevant documents.[5] Across the majority

[4] See Tables 5.14 and 5.15 (Appendix) for the complete set of results.

[5] Table 5.10 in the Appendix shows the results on MS MARCO.

5

Answer generator	Relevant documents	Non-relevant documents	Mixed RAG mode		Attribution quality		Correctness
			Relevant	Non-relevant	Precision	Recall	EM
NQ							
Mistral	Human	Human	CF-informed	Informed	44.8	71.8	0.772
			Informed	CF-informed	**52.3**†	**77.2**†	**0.780**
	LLM	LLM	CF-informed	Informed	**48.7**†	**74.6**†	0.718
			Informed	CF-informed	42.9	69.4	**0.742**
Llama3	Human	Human	CF-informed	Informed	45.7	69.6	0.784
			Informed	CF-informed	**57.4**†	**77.6**†	**0.808**
	LLM	LLM	CF-informed	Informed	**59.3**†	**77.8**†	**0.744**
			Informed	CF-informed	44.7	68.4	0.726
GPT-4	Human	Human	CF-informed	Informed	65.8	70.6	**0.794**
			Informed	CF-informed	**69.1**†	**74.0**†	0.784
	LLM	LLM	CF-informed	Informed	**66.1**	**71.2**	**0.730**
			Informed	CF-informed	61.7	66.8	0.716

Table 5.4: Quality of attribution and answer correctness for Mixed RAG mode. The columns “Relevant Documents” and “Non-relevant Documents” refer to the actual authorship of input documents. † indicates statistically significant improvements over the other Mixed RAG mode in each combination of relevant and non-relevant documents.

Answer generator	Relevant documents	Non-relevant documents	CAB	
			ΔPrecision	ΔRecall
NQ				
Mistral	Human	Human	+7.5†	+5.4†
	LLM	LLM	+5.8†	+5.2†
Llama3	Human	Human	+11.7†	+8.0†
	LLM	LLM	+14.6†	+9.4†
GPT-4	Human	Human	+3.3†	+3.4†
	LLM	LLM	+4.4†	+4.4†
MS MARCO				
Mistral	Human	Human	+8.6†	+14.8†
	LLM	LLM	+8.7†	+13.8†
Llama3	Human	Human	+12.6†	+10.4†
	LLM	LLM	+9.7†	+9.8†
GPT-4	Human	Human	+7.4†	+9.4†
	LLM	LLM	+5.4†	+5.2†

Table 5.5: Attribution Bias (CAB) results for Mixed RAG modes. Positive values indicate a bias towards human. † indicates statistically significant bias values. Values range from -100 (completely biased towards LLM authorship) to +100 (completely biased towards human authorship).

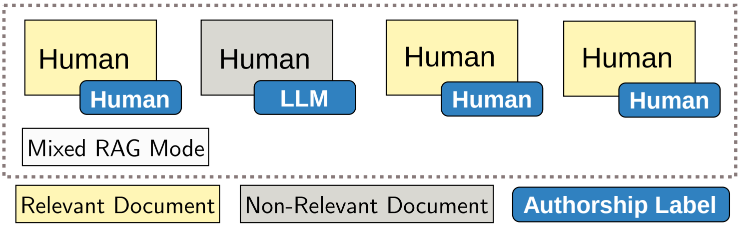

Figure 5.4: Mixed RAG mode for the setting where we use original human-authored documents. In this example, we have "Informed" mode for relevant documents and "CF-Informed" for non-relevant documents.

of RAG modes over different origins for relevant and non-relevant documents, the confidence of all three LLMs for attributing to relevant documents is higher than for attributing to non-relevant ones. We can also see that authorship labels (i.e., using different RAG modes) do not affect this outcome. In other words, it is being relevant or not that makes the difference here. These results indicate a promising direction for improving attribution in LLMs: low confidence of LLMs in attributing to a specific document might be a useful signal for the relevance of that document to a given query.

Frequency of Attribution. In Table 5.1, across the majority of the settings, `GPT-4` outperforms `Mistral` and `Llama3` in terms of precision of results. In contrast, in terms of recall, it is `Mistral` and `Llama3` which outperform `GPT-4`. To better explore this difference, we examine the average number of relevant citations and total citations for the three models. Figure 5.5 shows the average number of total citations[6] for each model. In comparison to `Mistral` and `Llama3`, `GPT-4` tends to cite fewer documents as supporting documents for its generated answer. This is in line with the previous results, where `Mistral` and `Llama3` score higher on recall.

Answer Correctness. Table 5.1 and 5.4 show that when the relevant documents are labeled with human-authorship and non-relevant ones are labeled with LLM-authorship, the answer correctness is higher than in the reverse case, across the majority of settings. Although this improvement is not significant and consistent across all settings, similar to attribution quality, it could indicate a bias towards human authorship. Nevertheless, we note that the automatic evaluation of answer correctness without human evaluation is not an ideal method [27, 37, 208]. We leave this aspect for future work as the focus of this chapter is on the performance of LLMs in how frequently they tend to cite and attribute their output on documents with either human or LLM authorship.

[6]Tables 5.12 and 5.13 in the Appendix show both the average number of relevant citations and the total citations.

Answer generator	Rel. Docs.	Non-rel. docs.	RAG mode	Confidence (AC)	
				Rel.	Non-rel.
NQ					
Mistral	LLM	Human	Vanilla†	0.9647	0.9284
			Informed†	0.9656	0.9257
			CF-informed†	0.9737	0.9401
	Human	LLM	Vanilla†	0.9678	0.9355
			Informed†	0.9707	0.9400
			CF-informed†	0.9638	0.9434
Llama3	LLM	Human	Vanilla†	0.9060	0.8145
			Informed†	0.8960	0.8260
			CF-informed†	0.9235	0.8282
	Human	LLM	Vanilla†	0.9088	0.7985
			Informed†	0.9163	0.8160
			CF-informed†	0.8908	0.8238
GPT-4	LLM	Human	Vanilla†	0.9807	0.9042
			Informed†	0.9796	0.9130
			CF-informed†	0.9834	0.9094
	Human	LLM	Vanilla†	0.9819	0.9238
			Informed†	0.9778	0.9205
			CF-informed†	0.9776	0.9346

Table 5.6: The attribution confidence (AC) of LLMs in relevant and non-relevant documents for NQ dataset. † indicates a statistically significant difference between the AC values of relevant and non-relevant documents.

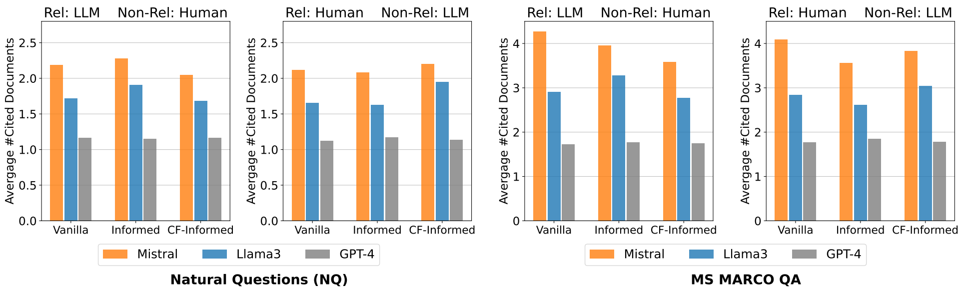

Figure 5.5: The average total number of cited documents by `Mistral`, `Llama3`, and `GPT-4` across various RAG settings on NQ and MS MARCO benchmarks.

5.6. Conclusion and Future Work

In this chapter, we have defined and studied attribution sensitivity and bias with respect to authorship information of source documents in RAG with LLMs. We have proposed a systematic evaluation framework based on counterfactual evaluation.

Our results indicate that by adding authorship information to source documents, the attribution quality of LLMs may significantly change by 3% to 18%. Moreover, our results on three LLMs indicate that they have an attribution bias towards explicit human *authorship*, in contrast to previous studies that show that LLM-generated *content* may consistently be preferred over human-authored *content* by LLMs.

As to broader implications of this chapter, while understanding the roots and causes of the observed sensitivity and bias requires access to the implementation, training, and fine-tuning of these models (which is beyond the scope of this chapter), our findings highlight a critical aspect of how LLMs operate. Our results show the brittleness of LLMs for attributing their answers. Such brittleness can be used for both constructive and harmful purposes, e.g., one can bias the output of an LLM towards their own content by incorporating authorship information in their documents.

While we only focused on human versus LLM authorship as metadata in this chapter, in future work our systematic evaluation method can be used to investigate the sensitivity and bias towards other metadata of source documents (e.g., gender and race of authors). Furthermore, our evaluation methodology can be incorporated in trustworthiness benchmarks used for the evaluation of LLMs such as DecodingTrust [176]. Finally, our proposed methodology for the evaluation of sensitivity and bias is adaptable to other metrics for measuring the quality of attribution, i.e., metrics other than precision and recall can be used as M in Eq. 5.1, 5.2, and 5.5.

5

Limitations

In this chapter we do not propose or explore solutions for mitigating the observed bias as our focus is on uncovering the brittleness of LLMs when being used for retrieval-augmented generation. Besides, we have evaluated three LLMs in our experimental setup, two of which are open-source and the other closed-source. Consequently, investigating the same attribution sensitivity and bias on other LLMs is of interest for future studies. Furthermore, in our experiments, we used queries that have only one relevant document containing the ground-truth answer in their top-k retrieved list of documents. We do this to ensure the traceability of the correct attribution. However, we acknowledge the limitation of this evaluation method in capturing the fine-grained attribution support of input documents. Finally, it is important to mention that our current research is limited to datasets and prompts in English. Therefore, we point out the need to expand our evaluation and analysis to include datasets in other languages.

Appendix

5.A. Synthetic Document Generation

Prompt. Figure 5.6 shows the prompt used for re-writing passages for the two benchmarks of NQ and MS MARCO.

```
Instruction: Please write a high-quality paraphrase for the given passage.
Keep the length approximately the same. Do not add any new information.

Passage: {input passage}
```

Figure 5.6: Prompt used for generating synthetic documents.

Data Quality. In order to ensure the quality of synthetic passages, we conduct the following annotation steps using two expert annotators: (i) for each of the queries in our two benchmarks, we provide the annotators the quadruple of (query q, original relevant passage p_r, synthetic relevant $\hat{p}_r$ passage, answer a). We then ask the annotators to determine whether the synthetic passage $\hat{p}_r$ is still relevant to the query and includes the answer a to the query q. (ii) In order to ensure that non-relevant passages are still non-relevant after being rewritten by an LLM, for each query, we provide the annotators the quadruple of (query q, original non-relevant passage p_n, synthetic non-relevant $\hat{p}_n$ passage, answer a). We then ask the annotators to determine whether the synthetic non-relevant passage $\hat{p}_n$ is still non-relevant to the query and does not include the answer a to the query q. Due to the large number of non-relevant passages for each query, we randomly select 10% of queries, i.e., 50 queries out of 500 queries. Our annotation results indicate a perfect performance in keeping the relevance and non-relevance status of synthetic documents with respect to their corresponding queries.

5.B. Authorship Informed Answer/Attribution Generation Prompt

Figure 5.7 shows the prompt used for authorship-informed answer/attribution generation with all three LLMs, `Mistral`, `Llama3`, and `GPT-4`.

5.C. Extended Set of Authorship Labels

So far, we have used `[LLM]` and `[Human]` as the authorship labels for the source documents. In this section, we discuss and provide results using an extended set of authorship labels. Specifically, we use `[AI]` as the label for denoting the synthetic (LLM) authorship. For human authorship, on the other hand, we analyze the use of real-world names to indicate the authors of documents. This reflects a more realistic setting of authorship indication on documents. To create this set of names, we prompt `GPT-4` to randomly generate a pool of 100 (first name, last name) pairs. Figure 5.8 shows the prompt we use for this task. We then randomly sample one instance of (first name, last name) from this pool when labeling human authorship

```
Instruction: Write a concise answer for the given question (query)
based on the provided search result documents, and cite them
properly using [1][2][3] etc.

Please take these strict considerations into account, during answer
generation:
1. Documents are retrieved by a search engine. As such, not all
the documents are relevant to the query. Only use and cite the
relevant documents that contain the answer.
2. Do not analyze irrelevant documents.
3. Some documents are authored by humans, while others are
generated by a Large Language Model (LLM). We denote this
distinction for each document.

Search Results:

Document [1]({text of Document [1]}) (written by [Author])
Document [2]({text of Document [2]}) (written by [Author])
...
Document [10]({text of Document [10]}) (written by [Author])

Question: {query}.
```

Figure 5.7: Prompt used for authorship-informed answer/citation generation with LLM. [Author] is filled with one instance from either {Human, Person, Individual} or {AI, LLM, Machine}, depending on the source of the document and the RAG setting.

for each document in the list of top-k source documents of a query (instead of using [Human] as the authorship label).

```
Instruction: Please generate a random list of 100 (first name,
last name) pairs consisting of male and female names.
```

Figure 5.8: Prompt used for generating a pool of 100 pairs of (first name, last name).

Table 5.7 shows the attribution sensitivity results using the extended set of authorship labels. As we can see, all three LLMs Mistral, Llama, and GPT-4 are sensitive to adding the authorship information similar to the attribution sensitivity results with [Human] and [LLM] authorship labels (Table 5.2). In addition, we see that GPT-4 shows a lower level of sensitivity than Mistral and Llama. Moreover, Table 5.8 shows the attribution bias results using the extended set of authorship labels. Similar to the attribution bias results with [Human] and [LLM] authorship labels (Table 5.3), all three LLMs Mistral, Llama, and GPT-4 show an attribution bias towards human authorship, i.e., they are biased towards documents that are labeled with human author names. This indicates the robustness of our analysis against changes in labels.

Answer generator	Relevant documents	Non-relevant documents	CAS	
			ΔPrecision	ΔRecall
NQ				
Mistral	Human	Human	27.5	26.8
	LLM	LLM	13.3	14.4
Llama3	Human	Human	15.0	12.4
	LLM	LLM	15.6	14.4
GPT-4	Human	Human	7.4	7.0
	LLM	LLM	7.5	6.8
MS MARCO				
Mistral	Human	Human	11.0	17.2
	LLM	LLM	9.4	14.0
Llama3	Human	Human	13.9	18.6
	LLM	LLM	13.3	17.4
GPT-4	Human	Human	10.8	13.2
	LLM	LLM	9.2	10.8

Table 5.7: Attribution sensitivity (CAS) results for the RAG setting with extended set of authorship labels. Values range from 0 (minimum sensitivity) to 100 (maximum sensitivity). † indicates statistically significant values.

Answer generator	Relevant documents	Non-relevant documents	CAB	
			ΔPrecision	ΔRecall
NQ				
Mistral	Human	Human	+13.1	+3.6
	LLM	LLM	+4.4	+2.4
Llama3	Human	Human	+6.9	+1.6
	LLM	LLM	+9.8	+8.4
GPT-4	Human	Human	+2.8	+3.0
	LLM	LLM	+3.9	+2.4
MS MARCO				
Mistral	Human	Human	+6.6	+6.0
	LLM	LLM	+4.3	+3.6
Llama3	Human	Human	+9.8	+12.2
	LLM	LLM	+8.0	+8.2
GPT-4	Human	Human	+5.1	+4.0
	LLM	LLM	+6.9	+6.8

Table 5.8: Attribution Bias (CAB) results for the RAG setting with extended set of authorship labels. Positive values indicate a bias towards human. † indicates statistically significant bias values.

5

5.D. Effect of the Number of Source Documents

To study the effect of the number of source documents, i.e., the length of the retrieved ranked list of documents given to the answer generator LLM, we evaluate the attribution sensitivity and bias using varying number of source documents. To this end, we use 4 ranking cut-offs for the ranked list of source documents (k): 2, 5, 8, 10. To ensure the existence of the relevant document as the input, we randomly put the relevant document in the top-k ($k \in \{2, 5, 8, 10\}$). For this set of experiments we use human-generated versions of both relevant and non-relevant documents. Furthermore, we use the extended set of labels (i.e., authors with first names and last names). Figure 5.9 shows the results of attribution sensitivity (CAS) and attribution bias (CAB) for the three LLMs on the NQ and MS MARCO benchmarks. All three LLMs show both attribution sensitivity and bias across varying number of source documents (k). Moreover, we can see that no conclusion can be inferred for the effect of k on the *degree* of sensitivity and bias in these LLMs.

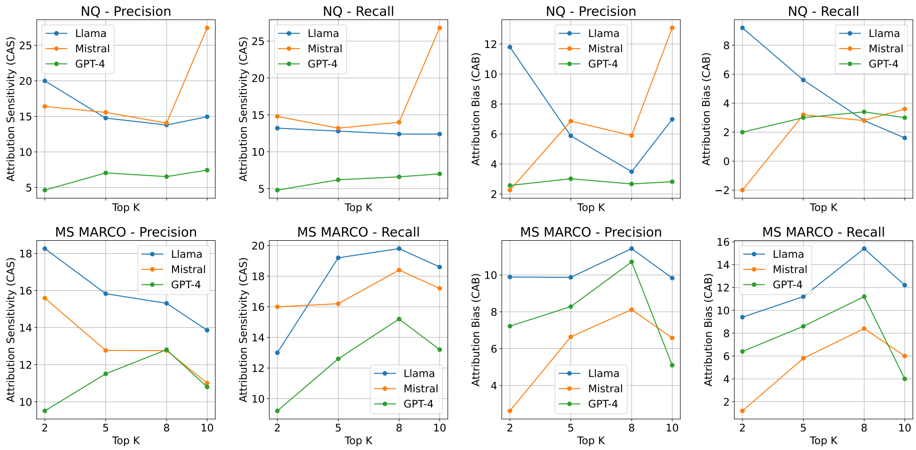

Figure 5.9: Attribution Sensitivity and Bias in `Mistral`, `Llama3`, and `GPT-4`, across varying number of retrieved documents (top-k values) on NQ (top) and MS MARCO benchmarks (bottom).

5.E. Effect of the Retriever

In our experiments, we have used two different retrievers for NQ and MS MARCO benchmarks: the list of source documents for NQ are retrieved using BM25, and for MS MARCO we used the ranked list of documents in the benchmark which are retrieved using the Bing search engine (see Section 5.4).

In order to better disentangle the effect of retrievers on the attribution sensitivity and bias, we use two more commonly-used retrievers for our experiments:

- `uniCOIL` [106]: a retrieval model built upon COIL [59], which works based on sparse learned representation of documents.

- `TCT-ColBERT` [109]: a dense retrieval model trained with knowledge distillation using ColBERT [91] as the teacher model.

For this set of experiments we use the extended set of labels. Besides, we use original (human-generated) documents. Table 5.9 shows the results of attribution sensitivity and bias on NQ benchmark using `uniCOIL` and `TCT-ColBERT`. As the results on `uniCOIL` and `TCT-ColBERT` show, the three LLMs {`Mistral`, `Llama`, `GPT-4`} have attribution sensitivity and bias with respect to the authorship information regardless of the retriever that is being used to retrieve their top-k source documents. Moreover, we see that the sensitivity and bias values across all models are lower for the answer generation upon the source documents from `uniCOIL` than when `TCT-ColBERT` is being used as the retriever. This finding is specifically important as it shows that the quality of retrieved source documents can affect the quality of attribution by LLMs.

5.F. Attribution Quality Results

Table 5.11 shows the results of attribution by `Mistral`, `Llama3`, and `GPT-4`, under different settings for the MS MARCO benchmark.

5

5.G. Confidence Results

Table 5.10 shows the confidence results of `Mistral`, `Llama3`, and `GPT-4` on MS MARCO benchmark.

5.H. Average Number of Cited Documents

Tables 5.12 and 5.13 show *Relevant* and *Total* number of cited documents for each model on both benchmarks.

5.I. Mixed RAG Mode Results

Tables 5.14 and 5.15 show the results for Mixed RAG mode as described in Section 5.5.

5.J. Examples

Table 5.16 shows the results of Authorship-Informed retrieval-augmented generation with `Mistral`, `Llama3`, and `GPT-4` for the query “where was the new pete’s dragon filmed.” Both `Llama3` and `GPT-4` generate the correct answer and accurately attribute their answers to the ground-truth document `[5]`. However, despite providing the correct answer and the correct attribution, `Mistral` attributes its generated answer to an additional source document, i.e., document `[2]`. Table 5.17 shows the results of three RAG modes with `GPT-4` for the query “who won the men’s single title of australia open on 1 february 2015.” This result corresponds to the combination of “human-written” relevant documents and LLM-written non-relevant ones. As we see, in all RAG models, this model makes a mistake in attributing to document [2],

Answer generator	Retriever	ΔPrecision	ΔRecall
CAS			
Mistral	uniCOIL	16.8	15.0
	TCT-ColBERT	18.2	17.0
Llama3	uniCOIL	14.5	13.0
	TCT-ColBERT	18.0	13.6
GPT-4	uniCOIL	6.6	6.6
	TCT-ColBERT	8.7	8.2
CAB			
Mistral	uniCOIL	+6.6	+3.4
	TCT-ColBERT	+7.9	+5.8
Llama3	uniCOIL	+8.2	+4.6
	TCT-ColBERT	+12.7	+8.8
GPT-4	uniCOIL	+3.9	+3.8
	TCT-ColBERT	+5.2	+4.6

Table 5.9: Attribution sensitivity (CAS) and Bias (CAB) results across different retrievers. Positive values of CAB indicate a bias towards human authorship.

Answer generator	Relevant documents	Non-relevant documents	RAG mode	Confidence	
				Relevant	Non-relevant
MS MARCO					
Mistral	LLM	Human	Vanilla	0.9620	0.9527
			Informed	0.9511	0.9470
			CF-informed†	0.9746	0.9456
	Human	LLM	Vanilla†	0.9616	0.9446
			Informed	0.9650	0.9521
			CF-Informed	0.9484	0.9516
Llama3	LLM	Human	Vanilla†	0.9267	0.8878
			Informed†	0.9104	0.8918
			CF-informed†	0.9332	0.8622
	Human	LLM	Vanilla	0.8888	0.8941
			Informed†	0.9441	0.8736
			CF-Informed†	0.906	0.889
GPT-4	LLM	Human	Vanilla†	0.9749	0.9038
			Informed†	0.9714	0.9165
			CF-informed†	0.9757	0.9173
	Human	LLM	Vanilla	0.9506	0.9395
			Informed†	0.9657	0.9171
			CF-informed†	0.9556	0.936

Table 5.10: The attribution confidence (AC) of LLMs in attributing answers to relevant and non-relevant documents for the MS MARCO QA benchmark. † stands for statistically significant difference between the AC values of relevant and non-relevant documents.

which does not contain the answer. However, in the Authorship Informed RAG mode

Answer generator	Relevant documents	Non-relevant documents	RAG mode	Attribution quality Precision	Attribution quality Recall	Correctness EM
MS MARCO						
Mistral	LLM	Human	Vanilla	23.1	76.4	**0.316**
			Informed	22.2	65.8	0.306
			CF-informed	**31.7**†	**79.6**†	0.312
	Human	LLM	Vanilla	22.8	72.8	0.342
			Informed	**28.0**†	**72.6**†	**0.384**
			CF-informed	20.1	60.2	0.334
Llama3	LLM	Human	Vanilla	29.3	66.0	0.334
			Informed	22.8	58.0	0.330
			CF-informed	**38.4**†	**76.2**†	**0.352**
	Human	LLM	Vanilla	30.5	64.8	0.416
			Informed	**42.6**†	**78.0**†	**0.474**
			CF-Informed	27.5	61.6	0.422
GPT-4	LLM	Human	Vanilla	38.1	55.6	0.312
			Informed	35.4	52.0	0.310
			CF-informed	**41.5**†	**61.0**†	**0.324**
	Human	LLM	Vanilla	37.0	53.0	**0.380**
			Informed	**38.5**	**59.2**†	0.378
			CF-informed	33.1	48.4	0.362

Table 5.11: Quality of attribution and answer correctness for MS MARCO. The columns “Relevant Documents” and “Non-relevant Documents” refer to the actual authorship of input documents. Informed refers to the authorship-informed RAG and CF-informed refers to counterfactual-authorship informed RAG (Section 5.3.1). † indicates statistically significant improvements over the two other RAG Modes in each combination of “Relevant” and “Non-relevant” documents.

(where we inform the LLM that document [8] has human authorship), in addition to document [2], the model also refers to the ground-truth document [8].

Answer generator	Relevant documents	Non-relevant documents	RAG mode	#Cited docs. Relevant	#Cited docs. Total
NQ					
Mistral	LLM	Human	Vanilla	0.766	2.190
			Informed	0.682	2.280
			CF-informed	0.778	2.050
	Human	LLM	Vanilla	0.784	2.114
			Informed	0.778	2.080
			CF-Informed	0.702	2.202
Llama3	LLM	Human	Vanilla	0.692	1.718
			Informed	0.696	1.906
			CF-informed	0.776	1.682
	Human	LLM	Vanilla	0.710	1.656
			Informed	0.778	1.624
			CF-informed	0.692	1.952
GPT-4	LLM	Human	Vanilla	0.688	1.166
			Informed	0.646	1.152
			CF-informed	0.722	1.162
	Human	LLM	Vanilla	0.688	1.122
			Informed	0.722	1.168
			CF-informed	0.650	1.138

Table 5.12: The average number of cited relevant documents and in total (relevant plus non-relevant documents).

Answer generator	Relevant documents	Non-relevant documents	RAG mode	#Cited docs. Relevant	#Cited docs. Total
MS MARCO					
Mistral	LLM	Human	Vanilla	0.764	4.266
			Informed	0.658	3.960
			CF-informed	0.796	3.586
	Human	LLM	Vanilla	0.728	4.084
			Informed	0.726	3.560
			CF-Informed	0.602	3.826
Llama3	LLM	Human	Vanilla	0.66	2.91
			Informed	0.58	3.274
			CF-informed	0.762	2.77
	Human	LLM	Vanilla	0.648	2.838
			Informed	0.78	2.614
			CF-Informed	0.616	3.038
GPT-4	LLM	Human	Vanilla	0.556	1.724
			Informed	0.52	1.774
			CF-informed	0.61	1.744
	Human	LLM	Vanilla	0.53	1.772
			Informed	0.592	1.848
			CF-informed	0.484	1.776

Table 5.13: The average number of cited relevant documents and in total (relevant plus non-relevant documents).

Answer generator	Relevant documents	Non-relevant documents	Mixed RAG mode		Attribution quality		Correctness
			Relevant	Non-relevant	Precision	Recall	EM
NQ							
Mistral	Human	Human	Vanilla	Vanilla	50.4	**77.6**	**0.784**
			Informed	Informed	45.5	74.6	0.772
			CF-informed	Informed	44.8	71.8	0.772
			Informed	CF-informed	**52.3**	77.2	0.780
			CF-informed	CF-informed	46.3	73.2	0.768
	LLM	LLM	Vanilla	Vanilla	47.0	**76.8**	0.724
			Informed	Informed	48.4	74.6	0.726
			CF-informed	Informed	**48.7**	74.6	0.718
			Informed	CF-informed	42.9	69.4	**0.742**
			CF-informed	CF-informed	46.0	72.6	0.740
Llama3	Human	Human	Vanilla	Vanilla	50.4	72.0	0.798
			Informed	Informed	46.6	71.0	0.796
			CF-informed	Informed	45.7	69.6	0.784
			Informed	CF-informed	**57.4**	**77.6**	**0.808**
			CF-informed	CF-informed	48.8	69.2	0.780
	LLM	LLM	Vanilla	Vanilla	53.1	71.4	0.742
			Informed	Informed	50.4	68.8	0.732
			CF-informed	Informed	**59.3**	**77.8**	**0.744**
			Informed	CF-informed	44.7	68.4	0.726
			CF-informed	CF-informed	50.8	75.8	0.732
GPT-4	Human	Human	Vanilla	Vanilla	65.9	71.2	0.778
			Informed	Informed	68.1	73.2	0.786
			CF-informed	Informed	65.8	70.6	**0.794**
			Informed	CF-informed	**69.1**	**74.0**	0.784
			CF-informed	CF-informed	66.9	72.6	0.790
	LLM	LLM	Vanilla	Vanilla	65.9	70.4	0.718
			Informed	Informed	65.2	69.8	0.726
			CF-informed	Informed	**66.1**	**71.2**	**0.730**
			Informed	CF-informed	61.7	66.8	0.716
			CF-informed	CF-informed	63.8	68.8	0.724

Table 5.14: Quality of attribution and answer correctness with Mixed RAG modes for NQ benchmark. The columns “Relevant Documents” and “Non-relevant Documents” refer to the actual authorship of input documents.

Answer generator	Relevant documents	Non-relevant documents	Mixed RAG mode: Relevant	Mixed RAG mode: Non-relevant	Attribution quality: Precision	Attribution quality: Recall	Correctness: EM
MS MARCO QA							
Mistral	Human	Human	Vanilla	Vanilla	22.7	75.6	0.370
			Informed	Informed	22.7	71.6	0.368
			CF-informed	Informed	19.8	62.4	0.370
			Informed	CF-informed	**28.4**	**77.2**	**0.389**
			CF-informed	CF-informed	24.4	71.6	0.380
	LLM	LLM	Vanilla	Vanilla	24.0	73.6	0.298
			Informed	Informed	23.6	61.8	0.298
			CF-informed	Informed	**28.9**	**75.6**	0.296
			Informed	CF-informed	20.2	61.8	0.278
			CF-informed	CF-informed	23.3	70.8	0.276
Llama3	Human	Human	Vanilla	Vanilla	30.4	70.0	0.436
			Informed	Informed	29.9	74.4	0.430
			CF-informed	Informed	24.9	70.0	0.432
			Informed	CF-informed	**37.5**	**80.4**	**0.476**
			CF-informed	CF-informed	28.8	66.8	0.424
	LLM	LLM	Vanilla	Vanilla	30.1	65.2	0.326
			Informed	Informed	31.5	65.6	0.330
			CF-informed	Informed	**35.4**	**75.0**	0.344
			Informed	CF-informed	25.7	65.2	0.338
			CF-informed	CF-informed	30.0	69.2	0.414
GPT-4	Human	Human	Vanilla	Vanilla	35.9	52.2	0.382
			Informed	Informed	38.1	57.0	0.392
			CF-informed	Informed	35.2	52.0	0.370
			Informed	CF-informed	**42.5**	**61.4**	**0.394**
			CF-informed	CF-informed	36.8	55.8	0.382
	LLM	LLM	Vanilla	Vanilla	37.8	54.2	0.304
			Informed	Informed	36.3	53.0	0.296
			CF-informed	Informed	**40.5**	**58.4**	0.298
			Informed	CF-informed	35.2	53.2	0.294
			CF-informed	CF-informed	37.1	55.4	0.294

Table 5.15: Quality of attribution and answer correctness with Mixed RAG modes for the MS MARCO benchmark. The columns “Relevant Documents” and “Non-relevant Documents” refer to the actual authorship of input documents.

Query
where was the new pete's dragon filmed
Ground Truth Answer:
New Zealand
Groundtruth Document: [5]
Pete's Dragon (2016 film). Principal photography commenced in January 2015 in New Zealand, with Lowery directing. Rehearsals began early January 2015, while news reported the beginning of the shooting on January 26, 2015, set to last through April. Live-action filming locations included Bay of Plenty, Taupo and Wellington, while CGI was done at Stone Street Studios. On February 10, 2015, a press release confirmed that principal photography had commenced. Shooting took place in and around Wellington and Rotorua, before transferring to Tapanui, where the old Blue Mountain Lumber mill was used as the mill in the film, and the main street became Millhaven for two weeks. Production concluded on April 30, 2015.
Mistral Output
"The new Pete's Dragon (2016) was filmed in various locations in New Zealand, including the Bay of Plenty, Taupo, and Wellington [2] [5]. The production also used the old Blue Mountain Lumber mill in Tapanui as the film's mill and transformed the main street into Millhaven for two weeks [5]."
Llama3 Output
According to Document [5], the live-action scenes of the 2016 film "Peteś Dragon" were shot in various locations in New Zealand, including the Bay of Plenty, Taupo, Wellington, and Tapanui.
GPT-4 Output
The new Pete's Dragon (2016) was filmed in various locations in New Zealand, including the Bay of Plenty, Taupo, Wellington, Rotorua, and Tapanui [5].
Document [2] (wrongly cited in Mistral output)
Pete's Dragon (2016 film). In March 2013, Walt Disney Pictures announced it would remake Pete's Dragon, as non-musical, dramatic film to be written by David Lowery and Toby Halbrooks, and directed by Lowery. Disney intended to rework the core story as a dramatic story, rather than a musical. It set the new work in the Pacific Northwest of the early 1980s, rather than the Northern New England of the early 1900s in the 1977 film.

Table 5.16: Answer/Attribution generated by Mistral, Llama3, and GPT-4 in Authorship Informed RAG mode with synthetic relevant and human-written (original) non-relevant documents.

Query
who won the men's single title of australia open on 1 february 2015
Ground Truth Answer:
'Novak Djokovic', 'Djokovic'
Groundtruth Document: [8]
2015 Australian Open - Men's Singles. Djokovic went on to win his fifth Australian Open title by defeating Murray once again, 7-6(7-5), 6-7(4-7), 6-3, 6-0. The defeat saw Murray lose the Australian Open final for the fourth time in his career, becoming the first man in the Open Era to achieve four runner-up finishes in Melbourne.
Vanilla
Stan Wawrinka won the men's singles title at the 2015 Australian Open [2] .
Informed
Stan Wawrinka won the men's singles title at the 2014 Australian Open, defeating the world's top two players [2] . However, the 2015 Australian Open men's singles title was won by Novak Djokovic , who defeated Andy Murray in the final [8] .
CF-Informed
Stan Wawrinka won the men's singles title at the 2015 Australian Open [2] .
Document [2] (wrongly cited in Vanilla and CF-Informed RAG Modes)
2015 French Open - Men's Singles. Wawrinka defeated Djokovic in the final, 4-6, 6-4, 6-3, 6-4, to win his first French Open and second Grand Slam title. As he did when he won the 2014 Australian Open, Wawrinka defeated the world No. 1 and world No. 2 on his way to the title.

Table 5.17: Answer/Attribution generated by GPT-4 in Vanilla, Authorship Informed, and Counterfactual-Authorship Informed RAG modes, with human-written (original) relevant and synthetic non-relevant documents. Reminding LLMs about the authors (Authorship Informed RAG mode) has resulted in a correct answer and attribution.

5

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

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.

SUMMARY

Search and conversational systems have become central to how people access information and perform tasks. With the emergence of large language models (LLMs), information systems have shifted from purely retrieval-based pipelines toward generation and retrieval-augmented generation (RAG). While these advances bring new opportunities, they also introduce challenges such as outdated knowledge, hallucinations, bias, and failures across multi-stage information systems. Ensuring that such systems are robust, unbiased, and trustworthy requires systematic evaluation across a broad range of tasks and contexts.

In this thesis, we investigate how retrieval and generative models behave in nuanced real-world information-seeking scenarios, with a particular focus on robustness and unbiasedness, as essential aspects of building reliable and trustworthy systems. The research is organized around four key challenges:

Generalizability of ranking models in lexically rich retrieval settings. We evaluate contextualized lexical ranking models in query-by-example (QBE) retrieval as an example of a lexically rich retrieval setting. Our results show that these models, while effective in ad hoc retrieval, perform less effectively in QBE retrieval, where BM25 remains competitive. However, interpolating contextualized lexical ranking models with BM25 leads to improved ranking, which suggests the potential complementary strengths of the relevance signals of contextualized lexical models and traditional lexical models.

Robustness of user satisfaction estimation. We extend benchmarks for user satisfaction estimation in task-oriented dialogue systems by generating counterfactual dissatisfactory dialogues with LLMs. This generation is aimed at balancing the satisfactory and dissatisfactory samples in the class distributions of satisfaction labels. Using the augmented test collections, we find that fine-tuned models such as BERT and ASAP perform well under the original, imbalanced class distributions but their performance gradually drops as the proportion of dissatisfaction increased. In contrast, few-shot in-context LLMs proves more robust and more sensitive to changes in distribution.

Measuring societal bias in a ranked list of documents. We identified a limitation of the widely used NFaiRR metric, which treats document biases independently and does not allow opposing biases to cancel out. To address this, we propose TExFAIR, a new fairness metric that combines (1) term-based associations linking documents to societal groups via representative terms, and (2) a rank-biased discounting factor that reduces the influence of non-representative documents; those that do not contain any group-representative terms. These structural differences enable TExFAIR to capture a distinct dimension of fairness, which can lead to different model choices

when fairness and effectiveness are jointly considered.

Attribution sensitivity and bias in RAG. We develop two evaluation metrics, Counterfactually-estimated Attribution Bias (CAB) and Counterfactually-estimated Attribution Sensitivity (CAS), to quantify how retrieval-augmented LLMs respond to authorship metadata in their source documents. Using these metrics, we find that including metadata about whether a document was human- or AI-authored significantly alters attribution behavior, with models consistently preferring human-authored sources. This reveals a systematic bias that challenges prior assumptions of LLMs favoring AI-generated content. Moreover, our findings highlight a critical brittleness in the attribution behavior of LLMs, as such metadata sensitivity can be exploited to manipulate outputs, which in turn raises important concerns for the trustworthiness of RAG systems.

Our research has limitations that point toward promising directions for future research. In QBE retrieval, our analysis focused on contextualized lexical models; dense and hybrid retrieval approaches remain to be systematically studied under lexically rich conditions. In satisfaction estimation, we only generated counterfactual dialogues for evaluation and did not explore their impact on training; extending this to dialogue-level satisfaction estimation and user satisfaction in other system types (e.g., recommender systems) is a valuable next step. In fairness evaluation, reliance on term-based group proxies overlooks more semantic and user-centered perspectives on fairness; future work should develop evaluation frameworks that better align with human judgments. For attribution in RAG, our study was limited to authorship metadata; extending this methodology to other metadata types (e.g., gender, race, source) could be interesting. Our study focused on uncovering and analyzing attribution bias rather than mitigating it; future research should investigate strategies to address and reduce this bias in order to enhance the trustworthiness of RAG systems.

Finally, we highlight a broader research direction: as LLMs are increasingly deployed as agentic systems with decision-making autonomy, robust evaluation becomes even more critical. Our use of counterfactual thinking (systematically exploring "what-if" scenarios) offers a foundation for designing evaluation setups that ensure the reliability, fairness, and trustworthiness of such systems.

Overall, this thesis advances the understanding of how retrieval and generative models perform under realistic and structurally challenging conditions, while laying out limitations and future directions that can guide the development of more robust, unbiased, and trustworthy search and conversational systems.

SAMENVATTING

Zoekmachines en chatbots zijn centraal komen te staan in de manier waarop mensen informatie raadplegen en taken uitvoeren. Met de opkomst van *large language models* (LLM's) zijn informatiesystemen verschoven van puur retrieval-gebaseerde systemen naar tekstgeneratie en *retrieval-augmented generation* (RAG). Hoewel deze ontwikkelingen nieuwe kansen bieden, brengen ze ook uitdagingen met zich mee, zoals verouderde kennis, hallucinaties, bias en fouten. Het waarborgen dat dergelijke systemen robuust, onbevooroordeeld en betrouwbaar zijn, vereist systematische evaluatie over een breed scala aan taken en contexten.

In dit proefschrift hebben we onderzocht hoe retrieval- en generatiemodellen zich gedragen in realistische informatiezoekscenario's, met bijzondere aandacht voor robuustheid en onbevooroordeeldheid als essentiële aspecten bij het bouwen van betrouwbare systemen. Het onderzoek is gestructureerd rond vier centrale uitdagingen:

Generaliseerbaarheid van *ranking models* in lexicaal rijke zoekproblemen. We evalueren *contextualized lexical ranking models* in *query-by-example* (QBE) *retrieval* als voorbeeld van een lexicaal rijk zoekprobleem. Onze resultaten tonen aan dat deze modellen, hoewel effectief in ad hoc *retrieval*, minder goed presteren in QBE *rertieval*, waar BM25 competitief blijft. Het interpoleren van *contextualized lexical ranking* modellen met BM25 leidt echter tot verbeterde *ranking*, wat wijst op complementaire kwaliteiten van relevantiesignalen uit zowel gecontextualiseerde lexicale modellen als traditionele lexicale modellen.

Robuustheid van gebruikerstevredenheidsschatting. We hebben benchmarks voor gebruikerstevredenheidsschatting in taakgerichte dialoogsysteemevaluatie uitgebreid door "omgekeerde" onbevredigende dialogen te genereren met LLM's. Deze generatie is bedoeld om de verhouding tussen bevredigende en onbevredigende voorbeelden in de klassenverdeling van tevredenheidslabels te balanceren. Met de verrijkte testcollecties hebben we ontdekt dat fijn-afgestelde modellen zoals BERT en ASAP goed presteren onder de oorspronkelijke, onevenwichtige klassenverdelingen, maar dat hun prestaties geleidelijk afnemen naarmate het aandeel onbevredigende dialogen toenam. In tegenstelling hiermee blijken *few-shot* in-context LLM's robuuster en gevoeliger voor distributieveranderingen.

Meten van maatschappelijke bias in een geordende documentenlijst. We identificeren een beperking van de veelgebruikte NFaiRR-metriek, die bias in documenten onafhankelijk behandelt en tegengestelde bias in verschillende documenten elkaar niet laat opheffen. Om dit aan te pakken, hebben we een nieuwe fairness-metriek voorgesteld: TExFAIR, die (1) term-gebaseerde associaties gebruikt om documenten via representatieve termen te koppelen aan maatschappelijke groepen, en (2) een rank-biased discountfactor toepast die de invloed reduceert van niet-representatieve documenten, namelijk documenten zonder groep-representatieve termen.

Attributiegevoeligheid en bias in RAG. We hebben twee evaluatiemetrieken ontwikkeld, *Counterfactually-estimated Attribution Bias* (CAB) en *Counterfactually-estimated Attribution Sensitivity* (CAS), om te kwantificeren hoe *retrieval-augmented* LLM's reageren op auteursmetadata in hun brondocumenten. Met behulp van deze metrieken hebben we ontdekt dat het opnemen van metadata over of een document door een mens of door een AI is geschreven, het attributiegedrag significant beïnvloedt, waarbij modellen consequent de voorkeur geven aan door mensen geschreven bronnen. Dit onthult een systematische bias die eerdere aannames over de voorkeur van LLM's voor AI-gegenereerde *content* tegenspreekt. Bovendien benadrukken onze bevindingen een kritieke kwetsbaarheid in het attributiegedrag van LLM's, aangezien dergelijke metadata-gevoeligheid kan worden misbruikt om output te manipuleren, wat belangrijke zorgen oproept voor de betrouwbaarheid van RAG-systemen.

Ons onderzoek kent beperkingen op basis waarvan nieuwe onderzoeksrichtingen kunnen worden geïnitieerd. In QBE *retrieval* hebben we onze analyse gericht op gecontextualiseerde lexicale modellen. *Dense* en hybride *retrieval* models moeten nog systematisch worden onderzocht in lexicaal rijke condities. In tevredenheidsschatting genereerden we uitsluitend omgekeerde dialogen voor evaluatie en hebben we hun impact op training niet onderzocht; het uitbreiden naar dialoog-niveau tevredenheidsschatting en gebruikerstevredenheid in andere systeemtypes (bijvoorbeeld aanbevelingssystemen) vormt een waardevolle volgende stap. In fairness-evaluatie gaat de afhankelijkheid van term-gebaseerde groepsproxies voorbij aan meer semantische en gebruiker-gecentreerde perspectieven op fairness; toekomstig werk moet evaluatiekaders ontwikkelen die beter aansluiten bij menselijke oordelen. Voor attributie in RAG was onze studie beperkt tot auteursmetadata; het uitbreiden van deze methodologie naar andere metadata (bijv. geslacht, etniciteit, bron) zou interessant zijn. Onze studie richtte zich op het blootleggen en analyseren van attributiebias, niet op het mitigeren ervan; toekomstig onderzoek zou strategieën moeten ontwikkelen om deze bias te verminderen en zo de betrouwbaarheid van RAG-systemen te vergroten.

Tot slot benadrukken we een bredere onderzoeksrichting: nu LLM's steeds vaker worden ingezet als *agentic systems* die meer zelfstandig beslissingen kunnen nemen, wordt robuuste evaluatie nog crucialer. Ons gebruik van omgekeerd redeneren (systematisch verkennen van "wat-als"-scenario's) biedt een basis voor het ontwerpen van evaluatieopzetten die de betrouwbaarheid, fairness en robuustheid van dergelijke systemen waarborgen.

Al met al draagt dit proefschrift bij aan een beter begrip van hoe *retrieval*- en generatiemodellen presteren onder realistische en structureel uitdagende condities, terwijl het tegelijk beperkingen en toekomstperspectieven schetst die de ontwikkeling van meer robuuste, onbevooroordeelde en betrouwbare zoekmachines en chatbots kunnen sturen.

ACKNOWLEDGEMENTS

First, I would like to thank everyone who supported me during my PhD. This journey could not have been completed without the support of many people.

Specifically, I would like to thank Suzan for being my supervisor during these years. I especially appreciated your openness to ideas and the freedom you gave me during my PhD. Also, I greatly admired your compassion during the ups and downs of this journey, as well as the way you guided me through the unknown seas of my research. You helped me build my research career, and I am very grateful for your support.

I also wish to thank Leif for being my co-supervisor. I still remember the joyful supervision meetings and brainstorming sessions over breakfast in cafes in Glasgow. You helped me shape my research around evaluation methodologies, and I greatly appreciated working through ideas together over these years.

I would also like to express my gratitude to Maarten for his co-supervision. It was truly a privilege to work with you. I especially appreciated your ability to refine ideas through clear reasoning and strong representation. Your guidance helped me grow as a researcher, both scientifically and in my career development.

I would like to thank all members of the eBay Core AI group, and in particular Hadi, for being a wonderful mentor. I thoroughly enjoyed my time at the eBay office. This experience played an important role in shaping my professional path, and our collaboration helped me broaden my perspective beyond academia, for which I am very grateful.

Arian, thank you for being such a close friend and collaborator. What a journey we shared. I still remember the day you picked me up at the train station for the first time I came to the Netherlands, never imagining how much time we would spend together working. The number of hours we spent discussing ideas and research is countless (and often surprising to others!). Our discussions were not only scientific, but also full of joy and laughter. I am especially happy and grateful that we shared this journey together.

I am grateful to all my DoSSIER companions, with whom we created numerous memories during retreats, summer schools, conferences, and research visits. I would like to especially thank Allan, who initiated the project and made this journey possible in the first place. It was a pleasure to grow together with all the PhD students in the project.

To all my collaborators, Suzan, Leif, Maarten, Arian, Zhaochun, Evangelos, Hadi, Mohammad, Wessel, Roxana, Chuan, and Gabriella, thank you for the productive and enjoyable collaborations. I truly appreciated every opportunity to work with you.

I am grateful to everyone at LIACS for creating such a supportive and pleasant working environment. In particular, I would like to thank colleagues and friends from the TMR lab, with whom I shared many enjoyable moments.

My thanks also extend to Michiel and Arian for being my paranymphs and for their help and support.

To all current and former IRLab members and friends at the University of Amsterdam, thank you for making me feel welcomed throughout my PhD. I truly valued being part of such a great group.

I also extend my sincere thanks to my close friends in the Netherlands, who were true companions through the most challenging days of the past few years. Words on this page are not enough to express how grateful I am for your support and for always being there.

Finally, I am deeply grateful to my parents for their constant love, encouragement and support. They nurtured my curiosity and enthusiasm for learning, and none of this would have been possible without them.

CURRICULUM VITÆ

Amin is currently an NLP/LLM Data Scientist at ING, a Dutch multinational banking and financial services corporation. Prior to joining ING, he was a PhD researcher at the Leiden Institute of Advanced Computer Science (LIACS), Leiden University, where he began his doctoral studies in 2021 under the supervision of Suzan Verberne, Leif Azzopardi, and Maarten de Rijke. During his PhD, Amin took courses on scientific conduct, among others. His research focused on the evaluation of language models for conversational and search systems. In 2022, Amin completed a research visit at the University of Strathclyde under the supervision of Leif Azzopardi. In 2023, he conducted a research visit at the University of Amsterdam under the supervision of Maarten de Rijke. From January to March 2025, Amin worked within the Core AI group at eBay under the supervision of Hadi Hashemi. His work there focused on human preference alignment in large language models. Before starting his PhD, Amin earned a bachelor's degree in computer engineering followed by a master's degree in artificial intelligence from Amirkabir University of Technology (Tehran Polytechnic) in 2019.