Evaluation of bias and robustness in search and conversational systems

Abolghasemi, A.

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 However, the knowledge encoded in LLMs is static and may quickly become outdated [129].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:

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