Evaluation of bias and robustness in search and conversational systems

Abolghasemi, A.

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 While these advances bring new opportunities, they also introduce challenges such as outdated knowledge, hallucinations, bias, and failures across multi-stage information systems.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.