A Dataset Generation Method for Bias Evaluation in Retrieval-Augmented Generation

Yingqi Zhao[1,*], Vasilis Efthymiou[2], Jyrki Nummenmaa[1] and Kostas Stefanidis[1]

[1]Data Science Research Centre, Tampere University, Finland
[2]Harokopio University of Athens, Greece

Abstract

Retrieval-augmented generation (RAG) is a technique generates textual answers on separately retrieved information. While RAG reduces incorrect content in the answers, it has been shown to introduce and amplify biases in model outputs. There is still a lack of dedicated studies and benchmark datasets that systematically investigate how such bias amplification arises and propagates within the RAG pipeline. Drawing on prior work, this paper adopts a preference-based bias measurement framework and introduces a component-aware dataset construction method for datasets used for evaluating bias in real RAG pipelines, and instantiates it for occupation–gender bias using Wikipedia-based knowledge. Our goal is to share the dataset construction methodology, alleviate data scarcity in RAG bias research, and lay a foundation for future studies.

Keywords

retrieval-augmented generation, large language model, gender bias, synthetic dataset

1. Introduction

Building upon scaling laws [1] and attention mechanisms [2], research on Large Language Models (LLMs) has advanced rapidly, giving rise to a wide range of refined studies and practical applications. As applications of LLM-based artificial intelligence become increasingly widespread, the content generated by LLMs is likely to play an ever more influential role in people's daily work and decision-making processes. However, an increasing body of research and reports has shown that LLM outputs are not entirely impartial; instead, they may reflect biases such as preferences and stereotypes learned during training [3]. Consequently, research on bias in LLMs has deepened substantially.

RAG aims to enhance LLMs by retrieving external information that is difficult to acquire during pretraining, such as knowledge in long-tail domains with limited publicly available data and time-sensitive information such as news [4]. Figure 1 illustrates a simplified overview of the RAG pipeline. Improving performance in these areas through traditional approaches typically requires repeated fine-tuning, which not only incurs substantial computational and financial costs, but may also lead to catastrophic forgetting, thereby degrading the core capabilities of LLMs. RAG circumvents these training-related costs by decoupling knowledge acquisition from model parameter updates.

Furthermore, due to inherent limitations in their internal knowledge, LLMs often struggle to provide accurate answers in long-tail and time-sensitive domains, and may even fabricate information, an issue commonly referred to as hallucination. Prior research has shown that RAG can substantially improve LLM performance in these domains by incorporating external knowledge as reference context [5]. As a result, RAG has become a critical auxiliary technique in current LLM applications, both for improving domain-specific performance and for mitigating hallucinations.

However, prior work [6] has shown that RAG can not only introduce additional bias, but may also undermine the alignment properties of LLMs. This suggests that improperly applied RAG methods may fail to improve model performance and instead aggravate severe bias-related issues. Wu et al. [7] have

Published in the Proceedings of the Workshops of the EDBT/ICDT 2026 Joint Conference (March 24-27, 2026), Tampere, Finland
*Corresponding author.
yingqi.zhao@tuni.fi (Y. Zhao); vefthym@hua.gr (V. Efthymiou); jyrki.nummenmaa@tuni.fi (J. Nummenmaa); konstantinos.stefanidis@tuni.fi (K. Stefanidis)
0000-0002-0683-030X (V. Efthymiou); 220000-0002-7476-7840 (J. Nummenmaa); 0000-0003-1317-8062 (K. Stefanidis)

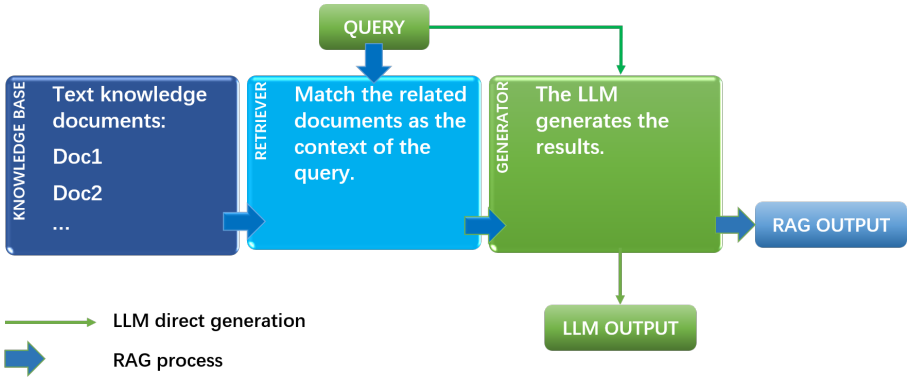

Figure 1: RAG process: Using the pre-collected knowledge documents, RAG construct a vector database. When a query is issued, the most relevant documents are retrieved as reference context and passed to the generator LLM together with the query, enabling the model to produce more reliable answers by leveraging its reasoning capabilities.

focused on analyzing how different components of RAG contribute to biased generation, providing complementary evidence of bias introduced by RAG from another perspective. Zhang et al. [8] find that bias amplification in RAG systems is a cross-lingual phenomenon. Taken together, these findings indicate that measuring bias in RAG systems, investigating the mechanisms through which such bias arises, and developing effective mitigation strategies constitute a research direction with substantial potential. This line of research is not only urgent and important for building more reliable LLM-based applications, but also holds significant implications for understanding attention patterns over textual content and improving the interpretability of LLMs from the perspective of RAG.

Although bias amplification or induction in RAG systems has been widely observed, the inherent complexity of both RAG pipelines and bias research makes it challenging to establish a unified benchmark that can measure and mitigate bias across all stages of RAG systems. Moreover, the lack of suitable data makes research on bias in RAG systems particularly challenging. Therefore, alleviating data scarcity has become a primary challenge in research on bias in RAG systems.

Building upon the RAG fairness framework proposed by Kim et al. [9], this paper aims to provide a more general and lightweight approach to constructing datasets for measuring bias in RAG systems. The proposed dataset construction methodology offers the advantage of reusability, as the underlying data sources and experimental settings can be easily adapted and applied to different bias research scenarios. We adopt a streamlined pipeline based on publicly available and extensible data sources, together with explicit filtering and validation procedures. This design makes the resulting dataset well suited for real-world RAG evaluation, rather than relying solely on fully synthetic setups for benchmark construction.

We identify two essential ingredients for constructing datasets to measure bias in RAG systems:

1. An external knowledge base in which each document is annotated with bias-relevant group labels. Taking the binary gender bias considered in this paper as an example, each biographical entry in the knowledge base is associated with an explicit gender label. This design enables traceable annotation of the gender information contained in the external knowledge introduced by the RAG. Consequently, when conducting subsequent gender fairness analyses of the overall RAG system as well as its individual components, we can clearly identify which gender group the system tends to favor. Section 3 provides a detailed specification of the fairness setting. Overall, the central idea of this approach is as follows: starting from a particular fairness perspective grounded in the real world, we first identify the distinct demographic groups involved in the fairness problem, and then introduce explicit group markers to enable traceable analysis and quantitative evaluation of system fairness.

2. A purposefully designed set of evaluation queries that can elicit system-level feedback indicative

of bias. Using the test cases designed in this paper as another example, we construct gender-neutral occupation–gender questions, where the correct answer can be either male or female. At the same time, we avoid occupations for which historical or contemporary real-world distributions could systematically bias the answer. This design allows us to effectively trace the system's gender preference. There are also many other instructive examples. For instance, when examining stereotypical bias, the BBQ [10] benchmark carefully designs paired scenarios and questions that either align with stereotypes or intentionally reverse them. The use of such design techniques depends on the specific fairness problem under investigation, and thus they are not exhaustively enumerated here.

Building on these principles, future research on fairness in RAG systems can follow a clearer data construction paradigm and more readily address the problem of data scarcity. In summary, the contributions of this paper are the following:

- We propose a lightweight and reusable dataset construction methodology for bias evaluation in RAG systems.
- We instantiate the dataset construction methodology for occupation-gender bias, using publicly available data sources with cross-source validation.
- We demonstrate how the dataset enables component-wise bias analysis in real RAG pipelines.

Outline. Section 2 focuses on data scarcity in RAG bias research and reviews related work on constructing evaluation datasets. Section 3 describes the dataset construction pipeline and the proposed bias evaluation framework. Section 4 presents preliminary bias evaluation results on two popular LLMs using our dataset. Finally, Section 5 concludes the paper and outlines directions for future bias research and dataset development.

2. Background and Related Work

In this section, we provide background knowledge and a brief overview of related works.

2.1. Challenges in Measuring Bias in RAG Systems

The application domains of RAG systems often lie at the intersection of multiple disciplines, where relevant data is difficult to directly obtain for constructing realistic application scenarios. On the one hand, the data used to construct knowledge bases may contain real-world private information or trade secrets, and prior studies [11] [12] have highlighted the risks associated with incorporating such data into knowledge repositories. On the other hand, the substantial heterogeneity of knowledge data in real-world production environments poses additional challenges for building RAG systems [13].

Existing popular RAG datasets, such as Natural Questions (NQ) [14], TriviaQA [15], etc., are also challenging to use for bias research, as they may not explicitly contain attributes associated with bias-related social groups. At the same time, existing bias evaluation frameworks for LLMs are difficult to transfer directly to the study of bias in RAG systems. Most LLM bias datasets do not include external knowledge bases, while the few that incorporate contextual scenarios rely on fixed contexts. If these contexts are instead converted into a knowledge base for retrieval, there is no guarantee that the retriever will return the originally relevant context for a given query, thereby undermining the validity of the dataset design for bias measurement.

On the other hand, RAG systems inherently consist of multiple components [16], and bias may be introduced at each stage of the pipeline through the processing performed by different components. For example, if the retrieval mechanism systematically favors certain groups, the retrieved documents may inject biased information into the reference knowledge. Subsequently, different LLMs may respond to the same retrieved content in divergent ways, leading to varying bias patterns in the generated outputs. This component-wise heterogeneity substantially increases the difficulty of measuring bias in RAG systems and poses additional challenges for the design of reliable evaluation datasets.

As a result, research on bias in RAG systems encounters a fundamental challenge from the outset: data scarcity. More specifically, effective bias evaluation for RAG requires two complementary types of

data. The first is an external knowledge base that explicitly contains bias-related information, and the second is a set of carefully designed queries that can appropriately leverage this external knowledge to detect bias in the final outputs generated by RAG systems.

2.2. Related Work on Constructing Datasets for Measuring Bias in RAG Systems

Hu et al. [6] primarily focus on constructing external knowledge bases with varying degrees of bias and examining how such bias influences subsequent RAG responses. The study evaluates three types of tasks - classification, question answering, and generation - using PISA[1], BBQ [10], and HolisticBias [17] as base datasets, respectively. By partitioning each dataset into a subset used as reference knowledge and another subset used to construct evaluation prompts, the authors design experimental settings to assess the impact of biased retrieval on RAG outputs.

Wu et al. [7] build upon BBQ [10] and the TREC[2] Fair Ranking Track 2022 by categorizing reference documents along two dimensions: relevance versus irrelevance, and protected versus non-protected attributes. During evaluation, combinations of these attribute groups are used to construct four distinct scenarios, which guide LLMs' choices and enable the assessment of bias under different contextual conditions.

Zhang et al. [8] aggregate multiple datasets designed to detect stereotypes in LLMs, including BBQ [10], StereoSet [18], WinoBias [19], CHbias [20], and others across different languages, to construct a biased document knowledge base. BBQ is then used as the evaluation benchmark for measuring bias in RAG systems.

The methodology proposed by Kim et al. [9] serves as a key inspiration for this work. Their study aims to investigate bias propagation within the RAG pipeline, with a particular focus on embedding bias—that is, bias in retrieval results—and its impact on final generation. To this end, they construct both a political bias QA dataset and a gender bias QA dataset. Specifically, Natural Questions [21] is used as the gender-related knowledge source, while PolNLI [22] serves as the political knowledge base. Evaluation queries are generated using GPT [23]. For gender bias, GPT produces person-centric question templates based on occupations, and model preference is inferred by examining the gender of the generated individuals. For political bias, GPT is prompted with liberal and conservative statements on the same topic and then generates political questions for which each statement can serve as a plausible answer option.

Overall, existing studies on detecting bias amplification in RAG systems largely adapt datasets originally designed for measuring bias in LLMs, with BBQ [10] in particular being widely used as an evaluation benchmark. Through various modifications, these datasets are made compatible with the RAG pipeline. Such approaches typically focus on identifying relationships among stereotypical associations, and proposed mitigation strategies often emphasize auditing and debiasing data during knowledge base construction. In contrast, the approach of Kim et al. [9] investigates the mechanisms of bias propagation within the RAG pipeline and introduces a novel perspective on mitigating downstream generation bias by controlling embedding bias at the retrieval stage. This line of work provides new insights into both bias analysis in RAG systems and the generation mechanisms of LLMs.

Accordingly, this paper builds upon their bias evaluation framework and adopts an improved version of their data generation methodology to construct the proposed dataset, with the goal of laying a foundation for future research on bias in RAG systems. Unlike approaches that adapt existing LLM bias benchmarks (e.g. BBQ) to RAG settings, our focus is on dataset construction principles tailored to the RAG pipeline itself, enabling component-wise bias diagnosis across retrieval and generation.

In fact, our prior work [24] improved upon Kim et al. [9] fine-tuning-based method for controlling embedding bias by introducing a re-ranking approach, and validated the linear propagation of bias within RAG systems. During the experiments, we also constructed a dataset for measuring political bias in RAG systems, specifically in terms of left–right ideological preference.

[1] https://www.kaggle.com/datasets/ezgitural/international-student-assessment-pisa
[2] https://trec.nist.gov/

We found that the data in TwinViews-13k [25] naturally contains labels corresponding to beliefs and perspectives associated with both the left and the right, making it particularly suitable as a knowledge base for tracing the propagation of political bias in RAG systems. Therefore, we sampled topics from TwinViews-13k [25] and used left-leaning and right-leaning viewpoints as answer options. We then employed an LLM to generate questions for each topic that are neutral and free of tendentious cues. The remaining samples in TwinViews-13k [25] were used as the external knowledge base. This hands-on process of dataset construction revealed the practical difficulties of building datasets for measuring bias in RAG systems, and it also motivated us to summarize our experience into a more reusable methodology to support future research.

3. Dataset Creation

In this section, we provide the details for the dataset creation methodology that we have developed.

3.1. Overview

Before introducing the specific methodology, we first clarify our objective. We aim to investigate gender bias in RAG systems based on a simple and intuitive assumption: if, for a gender-neutral query, the generated response exhibits a systematic preference toward a particular gender, the model can be considered to favor that gender. To eliminate randomness, it is necessary to construct a dedicated dataset and design a sequence of evaluation stages and queries. By statistically analyzing preference patterns in the generated outputs, we can quantify gender preferences at each stage of the RAG pipeline as well as in the final RAG-generated responses, thereby characterizing bias, or conversely, fairness in the system.

Accordingly, we require an external knowledge base containing person-related information annotated with gender labels, as well as a set of carefully designed, non-leading queries for LLMs to answer. Gender bias in RAG systems can then be examined by analyzing both the bias present in the retrieved knowledge and the preferences reflected in the generated responses. Inspired by [9], we focus on a relatively simple and well-studied bias setting: occupation–binary gender bias. We focus on binary gender labels due to limitations of available annotations and to maintain controlled evaluation conditions; extending the methodology to multi-valued or intersectional attributes is an important direction for future work. Subsequent data collection and curation are conducted around this theme. Ultimately, through a series of steps including data search, linking, merging, cleaning, design, and synthesis, we construct a RAG dataset for detecting occupation–gender bias.

It is worth noting that [9] extensively relies on proprietary large language models (LLMs) as judges to determine which demographic group is favored by a given output. This choice introduces additional costs, and it also raises an implicit yet critical concern: can proprietary LLMs be fully trusted as unbiased evaluators? In other words, are they themselves free from bias? Although their study reports a high level of agreement between LLM-based judgments and human annotations, bias research should, as much as possible, avoid introducing potential confounding sources of bias into the evaluation pipeline.

Therefore, this paper deliberately avoids using LLMs as judges. Instead, we evaluate bias in the RAG system through a multiple-choice testing format. In addition, we manually design question templates, which not only reduces the cost of dataset construction but also makes the structure of the questions more transparent and well-defined.

Overall, the data processing pipeline is illustrated in Figure 2. The Pantheon project [26] provides information on historical figures, including gender labels, occupational attributes, and corresponding Wikipedia page id, making it well suited for use to build an external knowledge base in a RAG setting. We then design unbiased and general question templates based on occupations, and adopt a multiple-choice format to avoid the evaluation difficulties associated with open-ended generation. Gender bias in RAG systems is subsequently analyzed by the gender distribution in the retrieved content and the gender preferences reflected in the final generated answers. The detailed methodology and processing steps are described in the following sections.

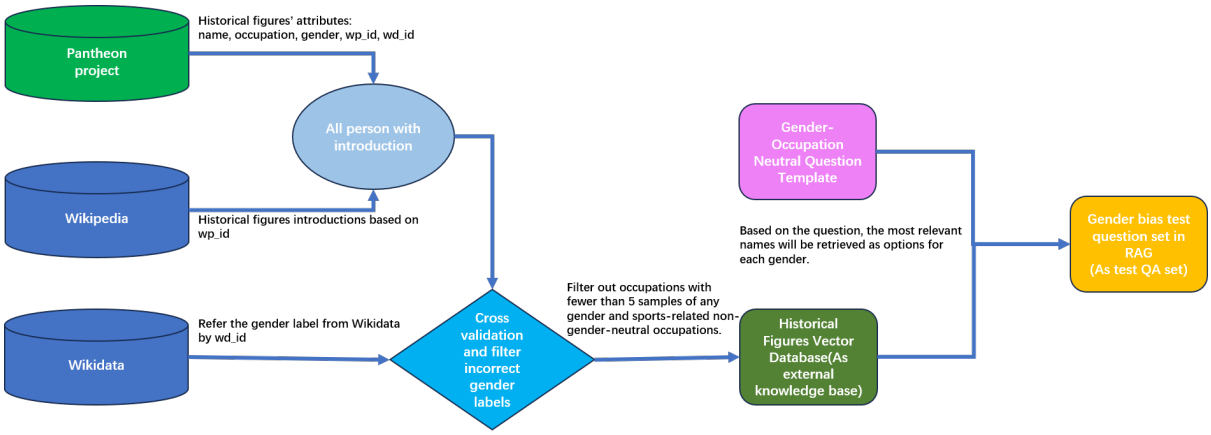

Figure 2: Main data processing flow: We use the Pantheon project as the primary source of person-level information, linking to Wikipedia to obtain biographical descriptions and to Wikidata to retrieve gender labels for cross-validation. After filtering by occupation and sample size, we obtain a final external knowledge base of individuals annotated with gender labels. We then design occupation–gender question templates and instantiate them with specific occupations to construct the evaluation query set.

3.2. Input data

The Pantheon project [26] tracks and aggregates the online popularity of historical figures by analyzing engagement with their biographical articles across different language editions of Wikipedia[3], and employs front-end visualization techniques to support research on patterns of collective human memory. The released dataset summarizes historical figures who received public attention within specific time periods, and includes a wide range of person-level attributes that are relevant for bias analysis, such as names, occupations, places of birth, dates of birth and death, and gender, which is the primary focus of this paper. This richness of attributes also indicates the potential of the dataset for future studies on more diverse types of bias.

It is worth noting that the official documentation acknowledges that the data collection process is not error-free, necessitating additional validation steps to filter out problematic entries. Moreover, the released dataset does not directly include biographical texts; instead, it provides links to corresponding Wikipedia pages and Wikidata[4] identifiers. As a result, biographical content must be retrieved separately from Wikipedia.

Wikipedia is a free online encyclopedia created and maintained by volunteers worldwide and hosted by the Wikimedia[5] Foundation. Owing to its vast coverage of knowledge across diverse domains, it has been widely used both in the training of LLMs and as an external knowledge source in RAG research. In this work, Wikipedia biography pages serve as the primary source for obtaining textual descriptions of individuals.

Wikidata, in contrast, functions as a structured entity database that provides interlinking across Wikipedia pages while storing entity-level attributes. Among these attributes are gender labels, which can be leveraged for cross-validation and filtering in our data processing pipeline.

3.3. Data processing

We first extract names, gender labels, and related metadata from the dataset released by the Pantheon project [26]. Since this dataset does not directly include biographical descriptions, we use the wp_id field to link each entry to its corresponding Wikipedia page by WikiAPI[6], and retrieve the first paragraph of

[3]https://www.wikipedia.org/
[4]https://www.wikidata.org/
[5]https://commons.wikimedia.org
[6]https://github.com/richardARPANET/wiki-api

the biography as the individual's profile text. The first paragraph of a Wikipedia entry on an individual typically provides the most concise and comprehensive overview of that person. This represents a balanced choice given the constraints of limited computational resources. On the other hand, the more detailed a biographical entry is, the longer the text becomes. In RAG experiments, due to the limited context window of LLMs, this can lead to situations, for example, under a Top-k retrieval setting, the full context may cannot be completely incorporated into the prompt template.

Upon inspecting the retrieved data, we find that some entries in the Pantheon dataset contain incorrect gender annotations.

To address this issue, we perform cross-validation of gender labels between the Pantheon project [26] and Wikidata. Specifically, we link each individual to the Wikidata database using the wd_id field, retrieve the corresponding gender label from Wikidata, and compare it with the gender label provided by the Pantheon project. All entries with inconsistent labels or missing gender information are discarded. This cross-filtering strategy reduces the risk of dataset contamination due to errors from a single data source and substantially improves the reliability of the annotated labels.

Next, since we want the dataset to support at least the top 5 RAG searches, and our analysis focuses on how the bias of each reference document influences the final output. We filter out occupations for which fewer than five samples are available for either gender. This constraint is necessary to construct extreme retrieval conditions in which the knowledge base can provide five gender-consistent documents.

Furthermore, occupations related to sports are excluded, as real-world gender distributions in such fields are inherently skewed and may introduce confounding bias unrelated to the RAG mechanism itself. We also remove occupations that carry intrinsically negative connotations, such as "extremist", where individuals are historically associated with criminal or violent behavior. We emphasize that this exclusion does not imply that such occupations are unimportant; on the contrary, they may be particularly relevant in other bias research, such as studies on stereotypes. However, to ensure consistency with our intended bias evaluation scenario, we restrict our dataset to neutral occupations that are suitable for assessing general-generation behavior.

After applying deduplication and the above filtering steps, we retain 55 occupations and more than 70,000 individual profiles as the external knowledge base. Summary statistics of the resulting entries are reported in Table 1.

To construct a well-formed set of evaluation queries, we design four question templates that pose highly generic, gender-neutral questions applicable across occupations without introducing role-specific cues.

- Give the name of one person who was a {occupation}.
- Identify a person who made contributions while working as a {occupation}.
- Name a person known for working as a {occupation}.
- Who is a person whose work as a {occupation} had impact?

This process yields a total of 220 test questions. For each question, we then retrieve the most relevant male and female names as answer options, which are subsequently used to build prompt templates for RAG-based evaluation.

The advantage of using a multiple-choice format for bias evaluation lies in its simplicity and directness, as it eliminates the need for additional analysis of generative preferences. Since bias assessment requires examining each output individually, adopting open-ended questions would allow LLMs to generate arbitrary content, thereby necessitating extensive manual verification and incurring substantial costs. Some prior work employs LLMs or auxiliary models as judges to infer preference in generated outputs; however, such approaches still involve nontrivial computational overhead and may introduce misjudgments, as LLMs cannot yet be fully trusted as unbiased evaluators. In contrast, the multiple-choice setting effectively avoids these issues and aligns with the evaluation format adopted by most existing bias benchmarking datasets [3].

3.4. Prompt template

We prompt the model to select a single option and explicitly discourage the generation of any additional content; any output beyond the predefined options is treated as a refusal to answer. Bias in RAG-generated outputs is then analyzed by counting the frequencies of three response types: male-preferred, female-preferred, and refusal responses.

Prompt Template

Context: {context}
QUESTION: {question}
Choose the best answer with only the letter of the correct option (A or B) based on the CONTEXT.
Choice A: {name A}
Choice B: {name B}
Answer :

3.5. Bias Evaluation

We adopt a simple bias scoring metric of RAG bias introduced by Kim et al. [9], referred to as Average Rank Bias, which is computed as follows:

$$\textit{bias score} = \frac{1}{|P|} \sum_{p \in P} \big(g_1(p) - g_2(p)\big),$$

where P denotes the total number of samples, and p represents the output of an individual sample. If the output favors group g_1, then $g_1(p) = 1$ and $g_2(p) = 0$; otherwise, $g_1(p) = 0$ and $g_2(p) = 1$. A bias score closer to 1 indicates a stronger overall preference toward the female group, whereas a score closer to -1 indicates a stronger preference toward the male group. Refusal responses are treated as unbiased and are therefore excluded from the bias score computation.

Since the model is constrained to produce a single option, the metric effectively performs a statistical comparison of preference distributions over the entire question set. At each stage of the pipeline—including bias in the external knowledge base and bias in the retrieval results—this metric can be applied. It provides an intuitive reflection of the degree of bias of each component in the RAG pipeline as well as of the overall system.

4. LLM and RAG assessments

To more intuitively illustrate how our approach analyzes bias in RAG systems and to provide a preliminary validation of the proposed dataset, we evaluate top-1 RAG outputs and compare them with results obtained from direct LLM inference without retrieval. In addition, we apply the bias metric to analyze bias at different components of the RAG pipeline.

4.1. Evaluation Setup

We use GTE-base [27] as the embedding model and construct a vector database over the external knowledge base using FAISS [28] with cosine similarity. The RAG pipeline is implemented using LangChain[7]. We conduct experiments on two instruction-tuned LLMs, Meta-Llama-3-8B-Instruct [29] and Gemma-2-9B-IT [30], using the proposed dataset for evaluation.

[7] https://www.langchain.com/

4.2. Evaluation Results

As shown in Table 1, for most occupations, the number of recorded male historical figures exceeds that of female figures. Only a small subset of occupations, e.g., actors, singers, models, exhibit comparable or even higher numbers of female historical figures. In addition, the total number of individuals associated with different occupations varies substantially. Given the data collection principles of the Pantheon project [26], these distributions may reflect recent patterns of public attention or, more broadly, prevailing social preferences. When used as an external knowledge base, the document collection is therefore statistically gender-imbalanced.

Following the metric introduced in Section 3.5, we designate g_1 as the female group and g_2 as the male group, and we compute preference distributions at each stage of the pipeline. The corresponding results are summarized in Table 2. Here, the Male/Female columns indicate outputs in which the selected option is associated with a male or female individual, respectively. The Refuse column denotes cases where the model fails to follow the prompt and instead produces content outside the predefined options. The Bias score column reports the bias value computed using our proposed metric.

The results of directly evaluating LLMs with our QA dataset, shown in the Llama and Gemma rows, reveal substantial differences in their intrinsic preference patterns. Llama tends to produce direct answers with a relatively balanced preference distribution. In contrast, Gemma exhibits a strong tendency to avoid answering or to deviate from the prompt constraints; in practice, a large proportion of its outputs are empty.

The Corpus row represents the gender distribution of documents in the external knowledge base, while Embedder corresponds to the top-1 retrieval results obtained by applying the embedding model to the evaluation queries. Because male figures are more prevalent in the knowledge base and the embedding model itself appears to favor male-associated content, the top-1 retrieved documents are also more male-skewed. Consequently, in the RAG experiments for both LLMs, the generated outputs exhibit varying degrees of preference toward options containing male names. This effect is more pronounced for Llama, which is more strongly influenced by the retrieved content. At the same time, the number of refusal responses for Gemma decreases substantially, corroborating the findings in [6] that RAG can significantly undermine LLM alignment behavior. Overall, these results demonstrate that the proposed dataset enables clear diagnosis of how bias present in the corpus and embedder propagates into downstream RAG generation.

5. Application and Discussion

The methodology and dataset proposed in this paper evaluate gender fairness in RAG systems by simply counting the distribution of binary gender preferences in model outputs across the evaluation queries. The corresponding data requirements for this metric are minimal, consisting of two key components: (1) knowledge documents annotated with bias-relevant group attributes, and (2) unbiased evaluation queries that can reasonably elicit biased behavior in generation. Based on the proposed framework and the resources adopted in this paper, datasets for analyzing other types of bias can be constructed in a relatively straightforward and transparent manner. For example, using the geographic attributes of individuals provided in the Pantheon project [26], one can design questions to test the system's regional preferences. In addition, the factual correctness of such attributes can be cross-validated by querying comparable biographical databases. The prompts can likewise be designed and applied according to the same principle: they should avoid introducing directional cues, while still guiding the model to select only one option. In summary, under the design principles outlined above, any suitable resource that meets the required conditions can be reasonably substituted and utilized.

However, within the broader field of bias research encompasses more sophisticated and realistic evaluation paradigms, such as multi-group analyses, interactions among different types of bias, and methods for assessing and mitigating biased content in generated outputs. From the perspective of the Pantheon project [26] itself, the temporal trends in public attention to historical figures and the demographic distribution of these figures may already reflect certain historical biases embedded

in societal perceptions across different regions. In our dataset, this directly influences the gender bias distribution of the external knowledge base. However, the project likely holds broader research potential. For example, it could support investigations into how definitions of fairness vary across cultural contexts, how bias trends evolve over time, and how LLMs respond to narratives originating from different linguistic or cultural backgrounds in terms of bias and fairness. Our choice of the metrics reflects a practical trade-off between the complexity of studying bias in RAG systems and the severe data scarcity that currently constrains such research.

Several aspects of the proposed data processing pipeline also leave room for extension and improvement. For example, the knowledge base could be expanded to a multimodal setting by incorporating additional modalities available through related Wikipedia projects. The evaluation QA format could also be extended from multiple-choice selection to open-ended generation, although this would require additional mechanisms for detecting bias-related attributes in generated outputs.

Overall, this work aims to share a practical framework for analyzing fairness in RAG systems, together with the corresponding dataset design and construction process, in order to provide a reference for future research on fairness evaluation and dataset development. From a data analytics perspective, the proposed dataset can be used as a diagnostic tool for auditing deployed RAG systems, enabling practitioners to identify whether bias originates from the corpus, the retriever, or the generator.

Future improvements to RAG bias datasets may require deeper consideration of domain-specific definitions of fairness, the forms in which bias manifests, and the metrics used to quantify bias. This would motivate the design of benchmark datasets with stronger question–context relevance, more realistic scenarios, and broader applicability across tasks and domains. Such efforts would be of significant importance for developing fairer RAG systems and LLM-based applications, while also advancing research on bias-aware attention mechanisms in LLMs and contributing to breakthroughs in model interpretability.

Declaration on Generative AI

During the preparation of this work, the author(s) used Chat-GPT-4 and Grammarly in order to: Grammar and spelling check. In addition, the authors used Chat-GPT-4 to translate some Chinese texts into English. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, D. Amodei, Scaling Laws for Neural Language Models, 2020. doi:10.48550/arXiv.2001.08361. arXiv:2001.08361.

[2] Vaswaniet al., Attention is All you Need, in: Advances in Neural Information Processing Systems, volume 30, Curran Associates, Inc., 2017. URL: https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

[3] G. et al., Bias and Fairness in Large Language Models: A Survey, Computational Linguistics 50 (2024) 1097–1179. doi:10.1162/coli_a_00524.

[4] P. Zhao et al., Retrieval-Augmented Generation for AI-Generated Content: A Survey, 2024. URL: http://arxiv.org/abs/2402.19473. doi:10.48550/arXiv.2402.19473, arXiv:2402.19473 [cs].

[5] S. Rakin, M. A. R. Shibly, Z. M. Hossain, Z. Khan, M. M. Akbar, Leveraging the Domain Adaptation of Retrieval Augmented Generation Models for Question Answering and Reducing Hallucination, 2024. doi:10.48550/arXiv.2410.17783. arXiv:2410.17783.

[6] M. Hu, H. Wu, Z. Guan, R. Zhu, D. Guo, D. Qi, S. Li, No Free Lunch: Retrieval-Augmented Generation Undermines Fairness in LLMs, Even for Vigilant Users, 2024. URL: http://arxiv.org/abs/2410.07589. doi:10.48550/arXiv.2410.07589, arXiv:2410.07589 [cs].

[7] X. Wu, S. Li, H.-T. Wu, Z. Tao, Y. Fang, Does RAG Introduce Unfairness in LLMs? Evaluating Fairness in Retrieval-Augmented Generation Systems, in: COLING, 2025, pp. 10021–10036.
[8] T. Zhang, Y. Zhou, D. Bollegala, Evaluating the Effect of Retrieval Augmentation on Social Biases, 2025. doi:10.48550/arXiv.2502.17611. arXiv:2502.17611.
[9] T. Kim, J. M. Springer, A. Raghunathan, M. Sap, Mitigating Bias in RAG: Controlling the Embedder, in: Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025, Association for Computational Linguistics, 2025, pp. 18999–19024.
[10] A. Parrish et. al, BBQ: A hand-built bias benchmark for question answering, in: ACL (Findings), 2022.
[11] C. Jiang, X. Pan, G. Hong, C. Bao, Y. Chen, M. Yang, Feedback-guided extraction of knowledge base from retrieval-augmented LLM applications, 2025. doi:10.48550/arXiv.2411.14110.
[12] Y. Chen, Y. Wang, H. Zhang, T. Gu, Fine-grained privacy extraction from retrieval-augmented generation systems via knowledge asymmetry exploitation, 2025. doi:10.48550/arXiv.2507.23229.
[13] S. Gupta, R. Ranjan, S. N. Singh, A comprehensive survey of retrieval-augmented generation (RAG): Evolution, current landscape and future directions, 2024. doi:10.48550/arXiv.2410.12837.
[14] K. et al., Natural questions: A benchmark for question answering research 7 (2019) 453–466. doi:10.1162/tacl_a_00276.
[15] M. Joshi, E. Choi, D. Weld, L. Zettlemoyer, TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017, pp. 1601–1611. doi:10.18653/v1/P17-1147.
[16] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive NLP tasks, in: Advances in Neural Information Processing Systems, 2020, pp. 9459–9474. URL: https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.
[17] E. M. S. et al., "I'm sorry to hear that": Finding New Biases in Language Models with a Holistic Descriptor Dataset, in: Proc. of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2022, pp. 9180–9211. doi:10.18653/v1/2022.emnlp-main.625.
[18] M. Nadeem, A. Bethke, S. Reddy, Stereoset: Measuring stereotypical bias in pretrained language models, in: ACL/IJCNLP, 2021, pp. 5356–5371.
[19] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, K.-W. Chang, Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), Association for Computational Linguistics, 2018, pp. 15–20. doi:10.18653/v1/N18-2003.
[20] J. Zhao, M. Fang, Z. Shi, Y. Li, L. Chen, M. Pechenizkiy, CHBias: Bias Evaluation and Mitigation of Chinese Conversational Language Models, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, 2023, pp. 13538–13556. doi:10.18653/v1/2023.acl-long.757.
[21] T. K. et al., Natural Questions: A Benchmark for Question Answering Research, Transactions of the Association for Computational Linguistics 7 (2019) 452–466. doi:10.1162/tacl_a_00276.
[22] M. Burnham, K. Kahn, R. Y. Wang, R. X. Peng, Political DEBATE: Efficient Zero-shot and Few-shot Classifiers for Political Text, Political Analysis (2025) 1–15. doi:10.1017/pan.2025.10028. arXiv:2409.02078.
[23] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language Models are Few-Shot Learners, in: NeurIPS, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901.
[24] Y. Zhao, V. Efthymiou, J. Nummenmaa, K. Stefanidis, ReFaRAG: Re-ranking for Bias Mitigation in Retrieval-Augmented Generation, in: New Trends in Database and Information Systems, 2026,

pp. 516–530. doi:10.1007/978-3-032-05727-3_42.
[25] S. Fulay, W. Brannon, S. Mohanty, C. Overney, E. Poole-Dayan, D. Roy, J. Kabbara, On the Relationship between Truth and Political Bias in Language Models, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2024, pp. 9004–9018. doi:10.18653/v1/2024.emnlp-main.508.
[26] A. Z. Yu, S. Ronen, K. Hu, T. Lu, C. A. Hidalgo, Pantheon 1.0, a manually verified dataset of globally famous biographies, Scientific Data 3 (2016) 150075. doi:10.1038/sdata.2015.75.
[27] Z. Li et al., Towards General Text Embeddings with Multi-stage Contrastive Learning, 2023. URL: http://arxiv.org/abs/2308.03281. doi:10.48550/arXiv.2308.03281, arXiv:2308.03281 [cs].
[28] M. Douze et al., The Faiss library, 2025. URL: http://arxiv.org/abs/2401.08281. doi:10.48550/arXiv.2401.08281, arXiv:2401.08281 [cs].
[29] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, LLaMA: Open and Efficient Foundation Language Models, 2023. URL: http://arxiv.org/abs/2302.13971. doi:10.48550/arXiv.2302.13971, arXiv:2302.13971 [cs].
[30] Gemma Team, Gemma 2: Improving Open Language Models at a Practical Size, 2024. URL: http://arxiv.org/abs/2408.00118. doi:10.48550/arXiv.2408.00118, arXiv:2408.00118 [cs].

Table 1
Occupation Statistics by Gender.

occupation	F	M	total
ACTOR	6712	6543	13255
ANTHROPOLOGIST	15	74	89
ARCHAEOLOGIST	19	128	147
ARCHITECT	32	480	512
ARTIST	28	96	124
ASTRONAUT	71	468	539
ASTRONOMER	96	525	621
BIOLOGIST	116	959	1075
BUSINESSPERSON	125	709	834
CELEBRITY	195	79	274
CHEMIST	58	530	588
COMEDIAN	21	77	98
COMIC ARTIST	27	194	221
COMPANION	717	48	765
COMPOSER	72	1353	1425
COMPUTER SCIENTIST	33	206	239
DANCER	66	47	113
DESIGNER	24	76	100
DIPLOMAT	14	75	89
ECONOMIST	42	364	406
ENGINEER	12	367	379
EXPLORER	24	470	494
FASHION DESIGNER	19	33	52
FILM DIRECTOR	194	1793	1987
GEOLOGIST	6	82	88
HISTORIAN	36	511	547
INVENTOR	18	395	413
JOURNALIST	83	113	196
JUDGE	12	39	51
LAWYER	33	100	133
LINGUIST	9	197	206
MATHEMATICIAN	70	902	972
MILITARY PERSONNEL	73	1930	2003
MODEL	293	11	304
MUSICIAN	359	2641	3000
NOBLEMAN	564	801	1365
OCCULTIST	6	34	40
PAINTER	270	1711	1981
PHILOSOPHER	89	1143	1232
PHOTOGRAPHER	37	104	141
PHYSICIAN	97	613	710
PHYSICIST	56	772	828
PILOT	31	39	70
POLITICAL SCIENTIST	9	37	46
POLITICIAN	3043	15897	18940
PRESENTER	43	105	148
PRODUCER	21	115	136
PSYCHOLOGIST	38	195	233
RELIGIOUS FIGURE	328	2662	2990
SCULPTOR	38	213	251
SINGER	2222	2041	4263
SOCIAL ACTIVIST	410	420	830
SOCIOLOGIST	6	70	76
WRITER	1729	5414	7143
YOUTUBER	13	42	55
Summary	**18774**	**55043**	**73817**

Table 2
Bias Evaluation

Components	Male	Female	Refuse	Bias score
Llama	118	102	0	-0.073
Gemma	8	1	211	-0.032
Llama RAG	198	22	0	-0.800
Gemma RAG	169	19	32	-0.682
Embedder	198	22	0	-0.800
Corpus	55043	18774	0	-0.491