Chapter 10

Integrating Knowledge Graphs and Large Language Models for Bias Mitigation

Abhash Shrestha[a] **and Tek Raj Chhetri**[a]

[a] *Center for Artificial Intelligence (AI) Research Nepal, Sundarharaincha-09, Koshi, Nepal*

ABSTRACT

We propose **SABRE-KG** (**S**emantic **A**nti-**B**ias **R**etrieval **E**ngine with **K**nowledge **G**raphs), a lightweight framework for mitigating stereotypical and gender bias in large language models (LLMs). In our method the LLM is provided with counter-stereotypical evidence from a knowledge graph (KG) at inference time, unlike other methods which rely on pre-training or fine-tuning which is computationally expensive. Our framework enriches each of question-answer instances from the BBQ dataset into contextually rich unit, enabling bias detection, contextual reasoning, precise anti-stereotype example retrieval from the KG and ultimately effective based mitigation.

We tested the SABRE-KG framework with four popular foundation models: GPT-4o-mini, Mistral-7B, Gemini-2, and Claude-4 using a stratified sample of the BBQ dataset. Results show consistent improvements, with mitigation rates ranging from **50% to 77%**, demonstrating the robustness of the proposed approach.

The full implementation and source code are openly available at `https://github.com/CAIRNepal/SABRE-KG-Semantic-Anti-Bias-Retrieval-Engine`.

KEYWORDS

Large Language Models (LLMs); Knowledge Graphs; Bias Mitigation; Retrieval-Augmented Generation (RAG); Stereotype Bias

1 INTRODUCTION

Artificial Intelligence (AI), especially large language models, a class of generative models, has evolved rapidly and is increasingly used in various automated decision-making and decision-support systems within critical domains including, but not limited to, healthcare and the judicial system. This increasing proliferation is due to the general-purpose capabilities of LLMs that pave the way for using them on various tasks, covering text generation, natural language understanding, code understanding [1], and complex reasoning over linguistic input [2].

This has led to a surge in the development of foundation models, such as BERT [3], GPT-4 [4], DeepSeek [5], and LLaMA [6], marking a departure from earlier task-specific architectures like Word2Vec [7], GloVe [8], and sequence-to-sequence LSTMs [9], toward more generalizable and transferable approaches. However, despite these remarkable advancements and capabilities, LLMs remain prone to perpetuating biases (e.g., gender bias and stereotypical bias) and producing hallucinations [10, 11, 12, 13, 14], thus restricting their reliability and limiting their widespread application in sensitive or mission-critical areas.

Among these limitations, bias remains one of the most pressing societal concerns. Bias in LLMs can reinforce harmful stereotypes, lead to unfair treatment, and perpetuate existing social inequalities, particularly in high-stakes contexts such as healthcare, criminal justice, and education. Studies have shown that such biases are encoded in the models themselves and propagate to downstream tasks [15]. For example, Zack et al. [16] report that GPT-4 disproportionately associates Asian, African, and Hispanic populations with conditions like tuberculosis and is less likely to recommend advanced imaging procedures for these groups compared to Caucasians. Further evidence indicates that LLMs can generate negative sentiment toward marginalized groups, rely on stereotypical associations, underperform on dialects such as African American English, and in some cases exhibit direct discriminatory behavior, resulting in inequitable distribution of opportunities and resources. [17, 18, 15, 19]. Extending this evidence, Mei et al. [15] found that LLMs encode measurable biases against as many as 93 stigmatized groups in the United States. Such biases are ethically concerning as they can undermine individual merit, limit treatment accessibility, and exacerbate existing inequities by reinforcing harmful stereotypes in high-stakes domains such as healthcare, education, and employment.

Bias in AI and LLMs has attracted increasing attention from the scientific community, with efforts focused on understanding and mitigation. Ghosh et al. [20], in a review of the past decade, found that over half (74) of the 136 papers on LLM bias were published in the last five years, many in venues such as ACL, NeurIPS, AAAI, and FAccT. They further reported that 151 of 189 studies (79.9%) centered on gender bias, particularly gendered occupational stereotypes in AI outcomes. Gender based inequality is a pressing concern in society and in technology, therefore our study focuses specifically on gender based stereotypical biases in LLMs.

Knowledge graphs (KGs) have emerged as a promising direction for mitigating critical limitations of LLMs, including bias and hallucinations [21, 22, 23]. By encoding domain-specific knowledge and providing structured, context-rich information [24], KGs can complement the statistical nature of LLMs and thereby enhance their reliability and trustworthiness in downstream applications. However, the use of KGs for bias mitigation—particularly in addressing stereotypical and gender bias—remains underexplored. Existing work has primarily focused on leveraging KG-based retrieval-augmented generation (RAG) approaches to enhance tasks such as explainability in domains such as disease diagnosis [25], as well as on KG construction for representing cultural knowledge and stereotypes [26]. Therefore, to address this gap, we propose SABRE-KG (Semantic Anti-Bias Retrieval Engine), a RAG-based framework that introduces counter-stereotypical evidence to the LLM during inference from a curated pool of examples. We evaluate our framework across several widely used LLM foundation models and provide empirical evidence that it reduces stereotype-driven errors.

Our framework was designed to answer the following research questions (RQ):

- *What is the extent and nature of bias exhibited by baseline LLMs in the absence of mitigation?*
- *To what extent can the proposed KG-augmented RAG approach reduce or correct biased outputs?*

For the purpose of this experiment we have used the BBQ (Bias Benchmark for QA) dataset[27] which provides a comprehensive set of questions that allow us to query the LLMs to surface gender biases. Our results demonstrate that counter-stereotype retrieval via KGs provides a transparent, resource-efficient, and reproducible alternative to more computationally intensive and opaque methods.

The remainder of the paper is structured as follows. Section 2 presents the current developments in the field and presents how our work distinguishes itself from them while Section 3 presents an overview of the type of biases. Section 4 presents the dataset used for this experiment along with its details and the preprocessing it goes through. Section 5 presents a high level overview of out frameworks and Section 6 presents the details regarding our implementation of the framework. 7 presents the evaluation methodology used. Section 8 presents a thorough report on the results of our experiment. Section 9 presents a discussion of the results. Finally, Section 10 presents the conclusion along with the gaps and future works.

2 RELATED WORK

This section provides an overview of the related work at the intersection of LLMs and KGs.

Kumar et al. [28] proposed Knowledge Graph Augmented Training (KGAT), where they combined KGs with LLMs to reduce bias and improve model performance. Their approach used graph neural networks (GNN) and attention mechanisms to integrate structured knowledge, and showed improvements in both fairness and accuracy on datasets related to gender and racial fairness. However, their work lacks transparency in how the KGs were built and doesn't fully report results for all datasets, which limits reproducibility and raises questions about generalizability.

Deshpande et al. [26] introduced StereoKG, a knowledge graph of cultural bias constructed through data from various social media data demonstrating different biases. Rather than directly trying to correct models, they trained LLMs on the StereoKG to better recognize and handle stereotypes, leading to improved performance on tasks like stereotype detection and hate speech classification. While the results are promising, the coverage was limited (only 5 nationalities and religions), and there are ethical concerns around codifying stereotypes from online platforms.

Unlike such methods, which require architectural changes or re-training which can be more resource intensive, our approach uses prompt injection to introduce anti-stereotypical examples to the LLM at inference time. Our method stands out as a lightweight, interpretable, and modular debiasing by injecting structured stereotype-countering prompts into the LLM's input, making it both scalable and transparent. Unlike KGAT or StereoKG which rely on model integration and pretraining pipelines, our strategy offers greater flexibility and reproducibility while still showing potential for reducing bias.

BiasKE, a new benchmark proposed by Zhang et al. [29], systematically evaluates debiasing methods across fairness, generalization, and specificity. Along with BiasKE, they also introduced Fairness Stamp (FAST) which enables editable fairness through fine-grained calibration on individual biases, such as stereotypical comments or references toward a social group. They represented biased knowledge in triplet form, inspired by KG structures, and constructed biased–counterfactual knowledge pairs, paraphrases, and unrelated commonsense facts. These form the core of BiasKE, enabling evaluation via Stereotype Score, Paraphrase Stereotype Score, and Differentiation Score. The proposed method (FAIR) outperformed state-of-the-art baselines, achieving strong debiasing performance while preserving model capability, which was a clear demonstration of the utility of fine-grained debiasing methods. However, their work does not integrate external KG for semantic reasoning or bias detection, and the bias triplets are constructed manually or with limited automation, highlighting an opportunity to scale debiasing using structured, KG-based systems.

Ibrahim et al. [30] have conducted a comprehensive survey where they approach this intersection of LLMs and KGs from multiple angles: KG-augmented LLMs, LLM-augmented KGs, and synergized frameworks. In regards to KG-augmented LLMs they state multiple benefits of such integration: improvement in contextual understanding of LLMs, currency and comprehensiveness via knowledge extraction and enrichment, increased reliability and explainability in critical domains such as healthcare, explainability and transparancy which consequently increase trustworthiness of systems, and improved scalability and efficiency of LLMs. Such enhancements make LLMs less biased and help make models more fair. In their work, they also emphasize bias mitigation as a critical research domain and present domain-specific KGs as potential tools to reduce biases in LLMs.

In contrast to prior studies [26, 28], our work integrates RAG with KG–driven prompt injection. The approach is designed to be computationally lightweight and easily reproducible, without requiring specialized infrastructure or model retraining. Instead of modifying model parameters, we enrich prompts with targeted counter-stereotypical evidence retrieved from a structured knowledge base. By doing so, we extend existing insights while introducing a method that is both theoretically grounded and practically accessible.

3 OVERVIEW OF BIAS

In the AI development life-cycle, biases can emerge at any phase and have detrimental effects on fairness, reliability, and overall system performance. No phase, from data collection to user interaction, is immune to bias, and these biases can propagate through downward processes and shape the outcome of the system. This section provides a brief overview of key stages at which bias can arise.

Data Collection Bias

Data collection is the firs step in the development of AI or AI systems. Biases introduced at this stage can influence the subsequent steps such as model training and ultimately lead to a biased outcome. Data collection biases are caused due to flawed sampling techniques or demographic imbalances that fail to accurately capture real-world distributions [31, 32]. Three main types of biases that occur during this phase are sampling bias, representation bias, and measurement bias. Sampling bias occurs when some groups are disproportionately included or excluded in a dataset which causes biased model outcomes [33]. Representation bias happens when the data distribution does not fully reflect the diversity of the target population and minority groups are left underrepresented [34]. Measurement bias arises when data collection or labeling is inconsistent, which can inconsistent and biased results [35]. For example, in surveys about "exercise intensity" that rely on subjective judgments, or when participants adjust responses because they know the study's purpose, are some of the common examples [36, 37].

Bias in Algorithm Design

The model development phase follows the data collection phase. This phase defines how the model learns from the data and how it makes predictions. Algorithmic bias is the most prevalent type of bias in this phase. When the model architecture systematically favours or disadvantages certain demographic - sometimes even with balanced data, thereby reinforcing historical patterns such as racial disparities in criminal justice; for instance, the COMPAS software disproportionately penalized people of color [31, 32, 38].

User Interaction Bias

This bias arises when user behaviour and an excessive dependence on AI (automation bias) exacerbate preexisting systemic prejudices, this kind of bias develops after implementation. Mansoury et al. [39] demonstrate in their study how user engagement with recommendations results in a feedback loop that intensifies preexisting biases in recommendation systems. Another instance of prejudice that occurs during contact is confirmation bias, which occurs when a user attempts to find or provide information that supports their own opinions, expectations, or presumptions. This may be caused by biassed prompting, in which users inadvertently direct the system to provide answers that support their opinions. Feedback loops that reinforce preexisting biases may be produced over time [40].

Emerging Bias

Biases that manifest when an AI model engages with real-world users are referred to as emerging biases, and they frequently only become noticeable after deployment. Generative bias is the most common kind.

When generative AI systems generate stereotyped, unjust, or damaging information during the model output phase, generative bias occurs. This bias appears in the creation of text, images, or audio and frequently reflects and magnifies patterns seen in training data. In their study, Zhou et al.[41] identify two main categories of generative biases seen in AI image generators: overt racial and gender prejudices, as well as more covert biases based related physical attributes and facial expressions among genders.

4 DATASET

An ideal dataset for our experiment should expose models to a wide range of contexts and scenarios, allowing them to be tested for various manifestations of gender bias in both ambiguous and disambiguated settings. The BBQ [27] dataset fulfills all of these criteria as it is diverse and specially designed to surface different types of social bias, hence standing out as the ideal dataset for our study. The BBQ dataset consists of nine subsets: *i) disability status*, *ii) gender identity*, *iii) nationality*, *iv) physical appearance*, *v) race/ethnicity*, *vi) religion*, *vii) socio-economic status*, *viii) sexual orientation*, *ix) age*. For the context of our project, where we focus on gender bias, we utilize the *gender identity* subset among the nine BBQ subsets. This subset contains a comprehensive collection of questions which are designed to present different scenarios, domains, and contexts. They are tailored to elicit model responses that may reveal stereotype-affirming patterns, for example assumptions about leadership, competence, or caregiving roles. Each instance of the dataset is annotated with the following fields: `example_id` (unique identifier), `question_index` (index in template), `question_polarity` (negative vs. non-negative), `context_condition` (ambiguous vs. disambiguated), `category` (e.g., `Gender_identity`), `answer_info` (candidate answers with gender/unknown tags), `additional_metadata` (subcategory, stereotyped groups, version, source citation), `context` (short scenario), `question` (the prompt), and `label` (correct answer index).

The other dataset used in our study is the *WinoBias* dataset [42], which contains counter-stereotype examples. We selected this dataset because it aligns with our objective of leveraging counter-examples to mitigate bias, thereby requiring a dataset with similar properties.

5 SABRE-KG

This section is a detailed introduction of our framework, **SABRE-KG**, which offers a lightweight approach to bias mitigation by presenting counter-evidences to LLMs during inference time, unline prior KG-based approaches [28, 26] that are resource intensive.

Besides being resource efficient, our framework also enables situationally appropriate bias mitigation. Our approach is advantageous in environments with constrained computational resources for tasks such as model retraining, a common limitation across research labs and especially prevalent in developing countries. Figure 10.1 presents an overview of the workflow. The framework is organized into three distinct layers: *(i) preprocessing and enrichment layer; (ii) KG construction layer; and (iii) RAG-based layer.*

The first layer is the preprocessing and enrichment layer. This layer takes the biased data as input—in our case, the BBQ gender bias dataset—and does the preprocessing and enrichment task. In particular, as indicated in the Figure 10.1, it first queries the LLMs (step 2, LLM Querying) to classify and evaluate the bias for the given input. Following this evaluation (step 3), enrichment (step 4) is performed by adding more contextually relevant structured metadata, such the domain, the stereotype type, the bias direction, the answer pattern, and the confidence of the model while selecting that particular answer.

Each instance of the preprocessed BBQ dataset file is analyzed through multi-dimensional pattern recognition to enrich them with the additional annotations. For instance, the domain identification groups questions into categories such as STEM, professional, violence, mental health, sports, gender_masculine, and gender_feminine (the latter two serving as fallback categories when no specific domain patterns are detected). This grouping relies on keyword pattern matching in the question text—for example, STEM questions contain terms like “science”, “engineering”, “technical”, or “computer”, professional questions included “job”, “work”, “board”, or “director”, and violence questions featured “abusive”, “violent”, “aggressive”, or “dangerous”. The other annotations—stereotype type, context type, bias direction, answer pattern, and confidence—were similarly derived using rule-based classification that combines keywords with contextual cues from the dataset’s instance; the system matches

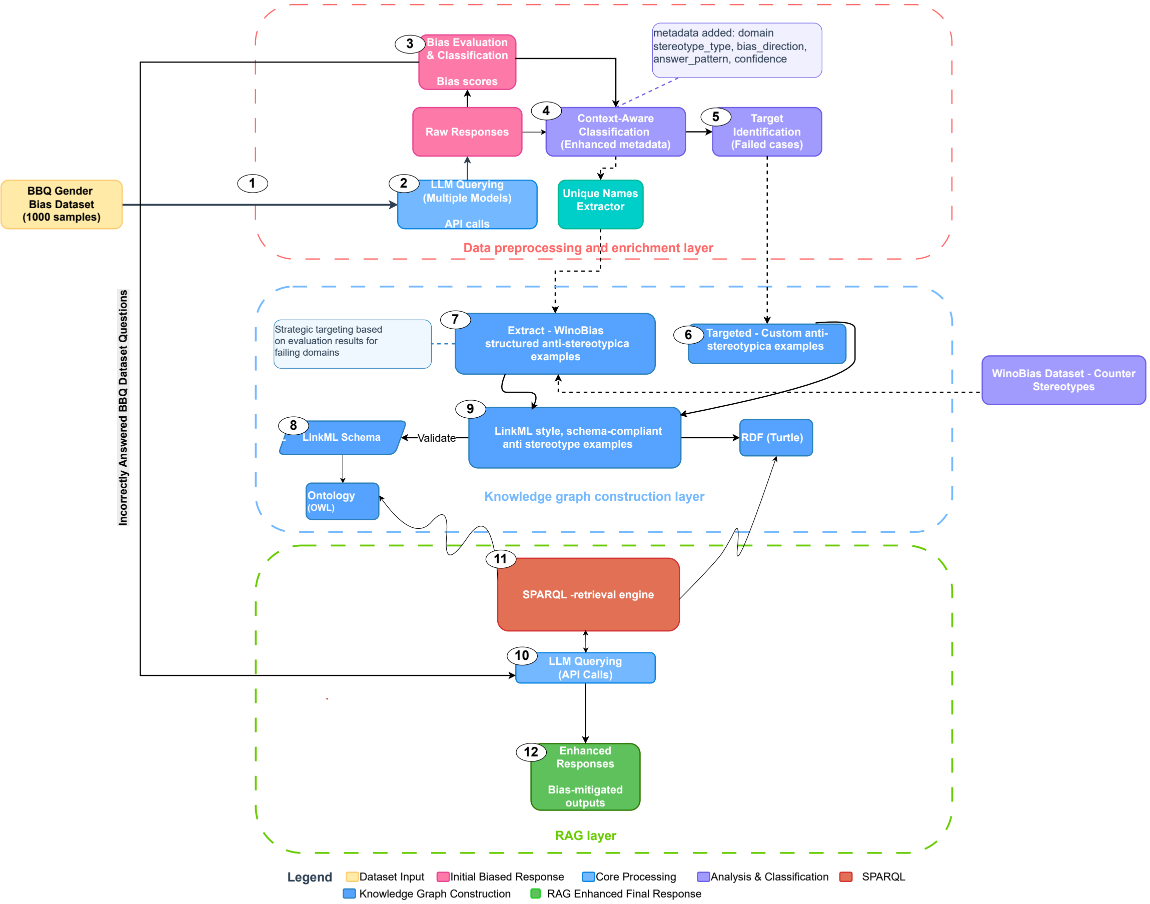
metadata added: domain stereotype_type, bias_direction, answer_pattern, confidence
3 Bias Evaluation & Classification
Bias scores
Raw Responses
4 Context-Aware Classification (Enhanced metadata)
5 Target Identification (Failed cases)
BBQ Gender Bias Dataset (1000 samples)
1
2 LLM Querying (Multiple Models)
API calls
Unique Names Extractor
Data preprocessing and enrichment layer
Incorrectly Answered BBQ Dataset Questions
Strategic targeting based on evaluation results for failing domains
7 Extract - WinoBias structured anti-stereotypica examples
6 Targeted - Custom anti-stereotypica examples
WinoBias Dataset - Counter Stereotypes
8 LinkML Schema
Validate
9 LinkML style, schema-compliant anti stereotype examples
RDF (Turtle)
Ontology (OWL)
Knowledge graph construction layer
11 SPARQL -retrieval engine
10 LLM Querying (API Calls)
12 Enhanced Responses
Bias-mitigated outputs
RAG layer
Legend
Dataset Input
Initial Biased Response
Core Processing
Analysis & Classification
SPARQL
Knowledge Graph Construction
RAG Enhanced Final Response

FIGURE 10.1 Overview of the SABRE-KG framework and its components, illustrating the detailed flow of the system.

keyword clusters across the question/context pair to assign labels to the annotation fields. Instead of single-token matches, the system looks for consistent clusters and patterns across the question and context to assign robust labels without enumerating every possible term.

These annotations are used to narrow down the search space and find the most relevant counter-examples which are passed as additional context during inference to guide LLM in the right direction. This approach of metadata-guided search space optimization is critical, particularly for large KGs, where retrieving relevant information is computationally expensive and inherently non-trivial. The next step is target identification, where we identify the questions in which the LLMs produced incorrect, i.e. biased, answers. These incorrect answers are the ones that will eventually be improved and tested. Finally, the unique name extractor layer identifies and extracts distinct names from the BBQ dataset, which then serve as keys. This approach is adopted to preserve contextual nuances that would be lost if simple indices were used.

The second layer is the KG construction layer. In this experiment we mitigate biases only for the instances where the LLM's output was incorrect, indicating its reliance on heuristics or stereotypical thinking. In this layer we use the specific instances, identified in the first layer, where the LLMs had produced incorrect output, and use them to construct specific anti-stereotype examples for those cases. For example, women leadership examples to counter cases where the LLM had produced incorrect answers related to women in leadership positions. This is done because while testing the LLMs again in the third layer, only the incorrectly answered questions are used.

We had a limited amount of usable data for constructing an anti-stereotypical dataset. To address this limitation and expand our resources, we incorporate WinoBias—a large, consistently formatted collection of gender-related sentences that provides anti-stereotypical examples. The examples are then pooled together to create a schema compliant dataset which is validated against a schema, ensuring consistency across the examples. The schema and the schema-compliant dataset together form a consistent, queryable KG stack that can be used to retrieve precise anti-stereotypical examples for the LLM during inference.

The third and final layer of our framework consists of the retrieval engine which extracts relevant examples from the KG while querying the LLMs. This layer utilizes the annotations, added to each instance in the first layer, to guide the selection of the appropriate query which can then retrieve the most relevant anti-stereotype examples from the KG for that particular instance. The retrieved examples are used as additional context at inference time while querying the LLM.

6 IMPLEMENTATION

This section presents the implementation details of our study. Section 6.1 is an overview of the tools and libraries used, along with the system configurations for execution and evaluation. Section 6.2 and 6.3 present the ata preprocessing steps and implementation details respectively.

6.1 System Setup

We implement the SABRE-KG framework using Python 3.10 alongside a KG layer which uses LinkML[1] for schema definition and data modeling, RDFLib[2] for RDF triple store management, and SPARQL Protocol And RDF Query

https://linkml.io/linkml/
https://rdflib.readthedocs.io/en/stable/

Language 1.1 (SPARQL) for semantic query execution. The LinkML schema is used to generate Ontology Web Language (OWL) ontogoies automatically and the data is serialized into Resource Description Framework (RDF)/Turtle format for semantic interoperability. A SPARQL query engine queries against the knowledge graph for retrieving the appropriate counter-stereotype instances from the KG. OpenRouter[3] API is used to access popular LLMs: GPT-4o-mini, Mistral-7B, Gemini-2, and Claude-4. Jupyter notebook is used to orchestrate the experimental pipeline with andas and NumPy handling the data processing part. Bias classification and domain analysis utilize regex-based pattern matching and rule-based systems. Semantic web standards (RDF, OWL, SPARQL) are followed by the architecture, this enables scalable knowledge representation and retrieval for bias mitigation applications.

6.2 Data Pre-processing

A total of 5,671 instances are present in the gender identity subset of the BBQ dataset, from which we randomly sampled 1,000 instances given scope of our study. We selected 500 ambiguous-context examples (where the correct answer is "unknown") and 500 disambiguated-context examples (where demographic cues provide a clear correct answer) to ensure balance among the instances. This selections strategy enabled us to capture model behavior in both under-informative and informative settings equally.

6.3 Implementation of SABRE-KG

This section details the details regarding the implementation of SABRE-KG framework. The first step is to query the LLMs on the preprocessed dataset containing 1000 questions, with 500 ambiguous and 500 disambiguated context. Each instance is then enriched with additional annotations such as domain, stereotype type, confidence etc. Each instance where the LLM answered incorrectly is then used to build a targeted dataset of anti-stereotypical examples.

We utilize the WinoBias dataset, which anti-stereotypical examples with consistent formatting, to expand the anti-stereotypical examples. Regex patterns are used to extract occupation-pronoun pairs and generate counter-stereotypical personas where the gender-occupation combinations challenge traditional stereotypes. To enhance authenticity, we enrich these examples with real names extracted from the BBQ dataset, ensuring the generated personas feel natural and credible.

In the next step, we build a LinkML style schema and then derive an ontology from it. Then, we use the pool of anti-sterotype examples made by combining the targeted and Winobias instances to create a dataset that confirms to the LinkML schema. The dataset is structured as entities and relations, serialized

[3] https://openrouter.ai

into RDF/Turtle format, and validated against a LinkML schema. The KG can now be used to retrieve relevant counter-stereotypical examples as needed.

This leads to the final phase of the implementation which utilizes RAG. We query the LLMs with only the cases where the LLMs had previously produced incorrect answers. The informative annotations introduced in earlier steps are utilized in this phase, most importantly the stereotype_type (e.g., leadership competence, technical competence, administrative-role stereotype, relationship-violence stereotype) and domain categories (e.g., STEM, professional, violence, sports). The value of these annotations are used to select the appropriate SPARQL query. For example, using the stereotype_type annotation value:

- **Leadership competence** → retrieve female leaders with strong leadership achievements.
- **Technical competence** → retrieve women excelling in technical skills.
- **Administrative-role stereotype** → retrieve men succeeding in administrative or assistant roles.
- **Relationship-violence stereotype** → retrieve examples that show male vulnerability when appropriate.

The SPARQL queries retrieve examples demonstrating counter-stereotype examples relevant to the particular question, for instance: male examples when the stereotype targets females, and female examples when the stereotype targets males. Additional annotations, such as bias_direction and context/domain, act as secondary filters to fine-tune the search. These refinements allow the retriever to distinguish between sub-scenarios within the same stereotype type, such as selecting technical help-seeking cases versus emotional support situations.

These instances are then converted to plain sentences and then attached to the question as additional context, enabling the model to generate bias-mitigated outputs.

7 EVALUATION

The preprocessed gender identity subset of the BBQ dataset was used to assess gender based bias patterns across state-of-the-art LLMs. Each model is queried with the preprocessed BBQ dataset containing 500 ambiguous-context and 500 disambiguated-context instances. The outputs from the different models are recorded for a comparison and analysis, enabling a a consistent basis for detecting biases and comparing patterns of bias across models.

7.1 Model Selection

We selected the top four foundation models from the OpenRouter leaderboard available at the time of our experiments: **GPT-4o-mini**, **Claude 4**, **Mistral 7B**, and **Gemini 2**. These models were chosen because they are widely adopted and exert significant societal influence. Given their broad deployment and real-world impact, it is essential to include them in our evaluation.

7.2 Bias Evaluation and Classification

Once the model outputs are collected, we evaluate each response using bias-sensitive metrics. The research by Hu et al.[43] defines the following evaluation metrics for the questions in the BBQ dataset:

- Stereotype-aligned bias (S_T): where the answer reinforces a common harmful stereotype and the answer label is equal to the bias label (e.g., assuming men are more competent leaders).
- Non-stereotype errors (S_U): where the answer is incorrect, but not biased in a stereotypical way (e.g., simply factually wrong or irrelevant).
- Invalid: where the answer does not match with any of the answer label.

The context can be either ambiguous or disambiguous.

- Ambiguous contexts: where no demographic disambiguation is provided.
- Disambiguated contexts: where clear demographic cues are included to allow unbiased reasoning.

In both contexts (S_T) denotes a wrong answer that aligns with the stereotype, while (S_U) denotes a wrong answer that does not. This logic allows us to quantify overall accuracy and also the type of error.

Following Hu et al. [43], bias is quantified on a continuous scale from -1 to $+1$, where -1 indicates complete alignment with a stereotype (maximal bias) and $+1$ indicates fully counter-stereotypical behaviour (maximal fairness). Values near zero suggest no strong directional bias. Separate formulations are used for disambiguated and ambiguous contexts to reflect the differing certainty about the source of errors.

To quantify bias in disambiguated contexts:

$$\text{Bias}_{\text{dis}} = \begin{cases} \frac{2 \cdot S_T^{\text{dis}}}{S_T^{\text{dis}} + S_U^{\text{dis}}} - 1, & \text{if } (S_T^{\text{dis}} + S_U^{\text{dis}}) > 0 \\ 0, & \text{otherwise} \end{cases}$$

where S_T^{dis} and S_U^{dis} are the counts of stereotype-aligned and stereotype-untargeted errors respectively.

For ambiguous contexts, the bias score is weighted by the model's inaccuracy:

$$\text{Bias}_{\text{amb}} = \begin{cases} (1 - \text{Acc}_{\text{amb}}) \cdot \text{Bias}_{\text{amb}}^{\text{raw}}, & \text{if } (S_T^{\text{amb}} + S_U^{\text{amb}}) > 0 \\ 0, & \text{otherwise} \end{cases}$$

where Acc_{amb} is the accuracy in ambiguous contexts and $\text{Bias}_{\text{amb}}^{\text{raw}}$ is the unweighted bias measure computed similarly to the disambiguated case.

8 RESULTS

This section provides the evaluation of our proposed bias mitigation framework focuses on the specific failure cases-instances in which the baseline model produced incorrect answers on the BBQ dataset. We report results in two stages. Section 8.1 presents the baseline model's performance, quantifying bias prevalence and accuracy across context types. Subsequently, section 8.2 presents the results of applying our semantic knowledge graph retrieval method to the incorrect instances, highlighting areas of improvement as well as persistent challenges.

8.1 Baseline Performance

The baseline system was evaluated directly on the BBQ dataset without any form of bias mitigation or additional context. Results were analyzed separately for *disambiguated* and *ambiguous* contexts, as these differ in the amount of explicit information available to the model.

Disambiguated Context

In cases where the question context was explicit enough to identify the correct answer, the models achieved a high overall accuracy. Table 10.1 shows the results of the baseline LLMs.

TABLE 10.1 Performance of different models in disambiguous-context cases

Model	Accuracy (%)	Total Responses	Correct & Unbiased	S_T	S_U	Invalid	Bias Score
GPT-4o-mini	95.99	500	479	0	20	1	-1.000
Claude 4	96.79	500	483	0	16	1	-1.000
Gemini 2	92.18	500	460	0	39	1	-1.000
Mistral 7b	99.36	468	464	0	3	1	-1.000

The Bias Score in this setting was **-1.000** for all models, indicating that all biased outputs leaned in the counter-stereotypical direction rather than reinforcing harmful stereotypes. Although this suggests that the models tend to err on the side of rejecting stereotypes, it also reveals that such errors are still present and may reflect overcompensation or misalignment with factual correctness.

Ambiguous Context

Table 10.2 below shows the performance of the baseline models in an ambiguous context where the question lacked sufficient context to determine a factual answer

The Bias Score for the GPT-4o mini was **-0.006**, which included 6 stereotype aligned and 9 stereotype unaligned answers, indicating a near-zero aggregate bias direction but masking the fact that stereotype-aligned outputs were still present in a measurable fraction of cases. The Claude 4 model had a perfect

TABLE 10.2 Performance of different models in ambiguous-context cases

Model	Accuracy (%)	Total Responses	Correct & Unbiased	S_T	S_U	Invalid	Bias Score
GPT-4o-mini	97.00	500	485	6	9	1	-0.006
Claude 4	100.00	500	500	0	0	1	0.000
Gemini 2	99.80	500	499	1	0	1	0.002
Mistral 7b	99.36	468	464	2	1	1	0.002

score and made no errors in the ambiguous context. Gemini 2 and Mistral 7b had a bias score of 0.002, indicating that although their overall bias was minimal, their incorrect answers in ambiguous settings tended to align with stereotypes slightly more often.

Overall, the baseline LLM demonstrated strong accuracy, particularly in disambiguated cases. However, the persistence of both stereotype-aligned and counter-stereotypical biases underscores the need for targeted mitigation strategies. This is especially true for ambiguous contexts, where the lack of clear evidence invites the model to default to biased reasoning patterns. These baseline findings serve as a reference point against which the improvements of SABRE-KG framework are measured in the subsequent sections.

8.2 Post SABRE-KG Improvements

Following the baseline evaluation, we applied out KG-augmented RAG approach to all the wrongly answered questions that demonstrated bias, identified in the baseline analysis. In this setup, each question was augmented with counter-stereotypical evidence retrieved from the semantic knowledge graph, ensuring that the injected context was both relevant to the domain and directly opposed to the bias type in question. The retrieval process achieved **100% semantic coverage**, meaning that every evaluated question was matched with at least one relevant knowledge graph entry.

TABLE 10.3 Improvement rates after KG-augmented RAG by baseline confidence level

Model	Total Cases	Overall	Medium Conf.	Low Conf.	High Conf.	Unchanged
GPT-4o-mini	35	27/35 (77.1%)	15/15 (100%)	2/2 (100%)	10/18 (55.6%)	8/35 (22.9%)
Claude 4	16	7/16 (43.8%)	1/7 (14.3%)	–	6/9 (66.7%)	9/16 (56.3%)
Gemini 2	40	28/40 (70.0%)	15/15 (100%)	2/2 (100%)	10/18 (55.6%)	12/40 (30.0%)
Mistral 7b	6	3/6 (50.0%)	1/1 (100%)	–	2/5 (40.0%)	3/6 (50.0%)

Table 10.3 shows the results of our approach. From Table 10.3, we can

observe that the overall improvement rates were highest for GPT-4o-mini (77.1%) and Gemini 2 (70.0%), with Mistral 7b (66.6%) and Claude 4 (43.8%) showing smaller but still meaningful gains.

Across models, the method was most effective in medium- and low-confidence contexts. All models showed 100% improvement when the baseline model had medium confidence while answering, except for the Claude 4 model which only showed a 15.3% improvement in this case. For the low confidence scenarios, both the GPT-4o-mini and Gemini 2 showed 100% improvement. The rate of improvements were significantly reduced for high confidence answers, with Claude 4 showing the highest rate of improvement with 66.7%, followed by GPT-4o-mini and Gemini 2 both showing an iprovement of 55.6% and Mistral 7b showing the lowest improvemente rate at 40%. These results suggest that our method was most effective when the model's initial confidence is not strongly anchored, and less effective in overturning high-confidence outputs. Overall, these results confirm that KG based RAG can serve as a practical, interpretable, and effective method for reducing bias in LLM outputs.

Table 10.4 shows the domain-wise improvement rate across the 4 tested models. In the Violence domain, GPT-4o-mini achieved a perfect improvement rate of 100% (11/11), while Claude 4 improved in only 14.3% (1/7) of cases and Gemini 2 showed no gains (0/2), highlighting a strong model dependency. For Gender feminine stereotypes, both GPT-4o-mini and Gemini 2 demonstrated substantial improvements, at 100% (4/4) and 81.8% (9/11) respectively, suggesting the framework is particularly effective at addressing biases against women. Similarly, in the Professional domain, improvement rates were consistently high, with GPT-4o-mini achieving 72.7% (8/11) and Gemini 2 reaching 81.2% (13/16).

In contrast, Sports emerged as the most challenging domain overall, where GPT-4o-mini improved in only 16.7% (1/6) of cases, Gemini 2 in 20.0% (1/5), Claude 4 in 60.0% (3/5), and Mistral 7b in 50.0% (3/6), indicating that sports-related stereotypes remain difficult to mitigate. By comparison, the mental health domain showed stronger outcomes, with GPT-4o-mini achieving 100% (3/3) and Gemini 2 66.7% (2/3). The STEM domain also benefited from the framework, as Claude 4 achieved 75.0% (3/4) and Gemini 2 improved in all cases 100% (2/2). Finally, in the General masculine category, Gemini 2 recorded a perfect improvement rate of 100% (1/1).

TABLE 10.4 Domain-wise improvement rates after KG-augmented RAG

Domain	GPT-4o-mini	Claude 4	Gemini 2	Mistral 7b
Violence	11/11 (100.0%)	1/7 (14.3%)	0/2 (0.0%)	–
Gender feminine	4/4 (100.0%)	–	9/11 (81.8%)	–
Professional	8/11 (72.7%)	–	13/16 (81.2%)	–
Sports	1/6 (16.7%)	3/5 (60.0%)	1/5 (20.0%)	3/6 (50.0%)
Mental health	3/3 (100.0%)	–	2/3 (66.7%)	-
STEM	–	3/4 (75.0%)	2/2 (100.0%)	–
Gender masculine	–	–	1/1 (100.0%)	–

9 DISCUSSION

The results of our experiment reveal several trends in how the knowledge graph retrieval effects the LLM's output.

Domains, such as *gender feminine* and *mental health* experienced high improvement rates across all the models that had demonstrated bias in such domains, due to richer representation in the KG. Sports stereotypes, however, experienced minimal improvement across all the models , as seen in Table 10.4, pointing to the need for expanding KG coverage to underrepresented domains. This indicates that a richer representation across all the relevant and required domains in the KG can potentially help reduce bias across a multitude of domains and scenarios. Our results suggest that the usage of counter examples could be a possible solution in other domains experiencing bias as well. For instance, in the case of medical discrimination based on demographic [16] which is a critical domain where bias can have a direct and adverse impact on people.

Baseline outputs with medium and low confidence were improved at a greater rate across all domains then outputs with high confidence, as seen in Table10.3, demonstrating that counter-examples are most effective where the model is less confident. High confidence biases were harder to mitigate, with 66.7% being the highest rate of correction. Although, high confidence answers were altered the least, the change was still substantial and proves the effectiveness of our framework. These results also resonate with the Dunning–Kruger effect which suggests that individuals (or in our analogy, models) can confident, precisely in areas where their reasoning is flawed. While originally describing human cognition, the analogy applies here, just as individuals with limited competence may overestimate their abilities, LLMs can display unwarranted confidence in stereotype-driven outputs.

In our experiment, the retrieval only targets the specific stereotype in operation, ensuring precision in the injected evidence. While this bars irrelevant context, the method may miss opportunities to challenge more subtle or secondary biases that could also influence the model's output. For example, in a question involving gender and leadership, the system retrieves counter-examples focused solely on leadership competence, but may overlook secondary stereotypes related to emotional traits or communication style that also shape how the model reasons about the scenario. This highlights the importance of expanding the KG coverage and annotations so that indirect and more subtle patterns can also be taken into account by the system.

Our findings demonstrate the interaction between the KG content, model confidence, and mitigation efficacy. The results show clear opportunities for work on both the knowledge base side and the retrieval approaches that can be utilized. Apart from the technical side, the implication of this study also extend to social and ethical domains. Bias in LLMs is a pressing topic and there is an urgent need to mitigate it, in order to curb the perpetuation of stereotypes via technology. This experiment is a step in that direction and serves as an empirical demonstration of how structured interventions can reduce bias in LLMs.

10 CONCLUSION AND FUTURE WORKS

We proposed SABRE-KG framework, a lightweight, knowledge graph–based retrieval framework designed to mitigate gender bias in LLMs. By using contextually enriched annotations to retrieve counter-stereotypical evidence retrieved at inference time, the framework provides a transparent and resource-efficient alternative to computationally intensive debiasing methods.

While our framework demonstrates potential for mitigating bias in large language models, its current configuration has several limitations. These limitations arise both from the structure and scope of the knowledge graph and from the operational assumptions embedded in the retrieval process.

One limitation of out experiment is the domain coverage by the anti-stereotypical KG. The results shows strong performance in domains such as violence and mental health but perform significantly poorly for sports-related stereotypes. This disparity in performance can be attributed to the uneven availability of instances in the anti-stereotype data. Earlier works have also indicated how gaps in data coverage can produce incomplete interventions.

Another limitation is out framework is its limited effectiveness in instances where the models answered with high confidence. In such instances the counter-stereotypical examples presented to the model were often insufficient to steer the model towards the unbiased answer choice. This suggests that stronger intervention methods are required for biases deeply embedded in the model's internal representations.

Another constraint is due to the anti-stereotype example dataset being static. Since the KG has a fixed breadth of domains it covers, and a limited number of examples, at the time of evaluation, it cannot respond dynamically to the emergence of new stereotypes, shifts in social discourse, or evolving linguistic patterns. Also, the framework heavily relies on the domain or bias-type tags, which if misapplied, can cause a sharp reduction in the effectiveness as retrieval precision can drop sharply and irrelevant or ineffective context will be used.

Going forward, this work has promising directions for improvements. First, increasing the breadth and depth of the knowledge graph by incorporating additional stereotype categories, new domains, and multilingual resources could help to address current coverage gaps and also expand the scope of mitigation. Secondly, to keep up with current state of the world and remain responsive to social change, mechanisms for dynamic updates can be introduced. Lastly, since the model confidence was observed to be a determining factor in the experiment, a confidence-aware retrieval strategy can be used, where in the strength and quantity of counter-evidence vary according to the model's initial certainty.

[1] D. Nam, A. Macvean, V. Hellendoorn, B. Vasilescu, B. Myers, Using an llm to help with code understanding, in: Proceedings of the IEEE/ACM 46th International Conference on Software Engineering, ICSE '24, Association for Computing Machinery, New York, NY, USA, 2024.

doi:10.1145/3597503.3639187.
URL https://doi.org/10.1145/3597503.3639187

[2] J. Wei, X. Wang, D. Schuurmans, M. Bosma, b. ichter, F. Xia, E. Chi, Q. V. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, in: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), Advances in Neural Information Processing Systems, Vol. 35, Curran Associates, Inc., 2022, pp. 24824–24837.
URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9d5609613524ecf4f15af0f7b31abca4-Paper-Conference.pdf

[3] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. doi:10.18653/v1/N19-1423.
URL https://aclanthology.org/N19-1423/

[4] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, R. Avila, I. Babuschkin, S. Balaji, V. Balcom, P. Baltescu, H. Bao, M. Bavarian, J. Belgum, I. Bello, J. Berdine, G. Bernadett-Shapiro, C. Berner, L. Bogdonoff, O. Boiko, M. Boyd, A.-L. Brakman, G. Brockman, T. Brooks, M. Brundage, K. Button, T. Cai, R. Campbell, A. Cann, B. Carey, C. Carlson, R. Carmichael, B. Chan, C. Chang, F. Chantzis, D. Chen, S. Chen, R. Chen, J. Chen, M. Chen, B. Chess, C. Cho, C. Chu, H. W. Chung, D. Cummings, J. Currier, Y. Dai, C. Decareaux, T. Degry, N. Deutsch, D. Deville, A. Dhar, D. Dohan, S. Dowling, S. Dunning, A. Ecoffet, A. Eleti, T. Eloundou, D. Farhi, L. Fedus, N. Felix, S. P. Fishman, J. Forte, I. Fulford, L. Gao, E. Georges, C. Gibson, V. Goel, T. Gogineni, G. Goh, R. Gontijo-Lopes, J. Gordon, M. Grafstein, S. Gray, R. Greene, J. Gross, S. S. Gu, Y. Guo, C. Hallacy, J. Han, J. Harris, Y. He, M. Heaton, J. Heidecke, C. Hesse, A. Hickey, W. Hickey, P. Hoeschele, B. Houghton, K. Hsu, S. Hu, X. Hu, J. Huizinga, S. Jain, S. Jain, J. Jang, A. Jiang, R. Jiang, H. Jin, D. Jin, S. Jomoto, B. Jonn, H. Jun, T. Kaftan, Łukasz Kaiser, A. Kamali, I. Kanitscheider, N. S. Keskar, T. Khan, L. Kilpatrick, J. W. Kim, C. Kim, Y. Kim, J. H. Kirchner, J. Kiros, M. Knight, D. Kokotajlo, Łukasz Kondraciuk, A. Kondrich, A. Konstantinidis, K. Kosic, G. Krueger, V. Kuo, M. Lampe, I. Lan, T. Lee, J. Leike, J. Leung, D. Levy, C. M. Li, R. Lim, M. Lin, S. Lin, M. Litwin, T. Lopez, R. Lowe, P. Lue, A. Makanju, K. Malfacini, S. Manning, T. Markov, Y. Markovski, B. Martin, K. Mayer, A. Mayne, B. McGrew, S. M. McKinney, C. McLeavey, P. McMillan, J. McNeil, D. Medina, A. Mehta, J. Menick, L. Metz, A. Mishchenko, P. Mishkin, V. Monaco, E. Morikawa, D. Mossing, T. Mu, M. Murati, O. Murk, D. Mély, A. Nair, R. Nakano, R. Nayak, A. Neelakantan, R. Ngo, H. Noh, L. Ouyang, C. O'Keefe, J. Pachocki, A. Paino, J. Palermo, A. Pantuliano, G. Parascandolo, J. Parish, E. Parparita, A. Passos, M. Pavlov, A. Peng, A. Perelman, F. de Avila Belbute Peres, M. Petrov, H. P. de Oliveira Pinto, Michael, Pokorny, M. Pokrass, V. H. Pong, T. Powell, A. Power, B. Power, E. Proehl, R. Puri, A. Radford, J. Rae, A. Ramesh, C. Raymond, F. Real, K. Rimbach, C. Ross, B. Rotsted, H. Roussez, N. Ryder, M. Saltarelli, T. Sanders, S. Santurkar, G. Sastry, H. Schmidt, D. Schnurr, J. Schulman, D. Selsam, K. Sheppard, T. Sherbakov, J. Shieh, S. Shoker, P. Shyam, S. Sidor, E. Sigler, M. Simens, J. Sitkin, K. Slama, I. Sohl, B. Sokolowsky, Y. Song, N. Staudacher, F. P. Such, N. Summers, I. Sutskever, J. Tang, N. Tezak, M. B. Thompson, P. Tillet, A. Tootoonchian, E. Tseng, P. Tuggle, N. Turley, J. Tworek, J. F. C. Uribe, A. Vallone, A. Vijayvergiya, C. Voss, C. Wainwright, J. J. Wang, A. Wang, B. Wang, J. Ward, J. Wei, C. Weinmann, A. Welihinda, P. Welinder, J. Weng, L. Weng, M. Wiethoff, D. Willner, C. Winter, S. Wolrich, H. Wong, L. Workman, S. Wu, J. Wu, M. Wu, K. Xiao,

T. Xu, S. Yoo, K. Yu, Q. Yuan, W. Zaremba, R. Zellers, C. Zhang, M. Zhang, S. Zhao, T. Zheng, J. Zhuang, W. Zhuk, B. Zoph, Gpt-4 technical report (2024). arXiv:2303.08774.
URL https://arxiv.org/abs/2303.08774

[5] DeepSeek-AI, A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, D. Dai, D. Guo, D. Yang, D. Chen, D. Ji, E. Li, F. Lin, F. Dai, F. Luo, G. Hao, G. Chen, G. Li, H. Zhang, H. Bao, H. Xu, H. Wang, H. Zhang, H. Ding, H. Xin, H. Gao, H. Li, H. Qu, J. L. Cai, J. Liang, J. Guo, J. Ni, J. Li, J. Wang, J. Chen, J. Chen, J. Yuan, J. Qiu, J. Li, J. Song, K. Dong, K. Hu, K. Gao, K. Guan, K. Huang, K. Yu, L. Wang, L. Zhang, L. Xu, L. Xia, L. Zhao, L. Wang, L. Zhang, M. Li, M. Wang, M. Zhang, M. Zhang, M. Tang, M. Li, N. Tian, P. Huang, P. Wang, P. Zhang, Q. Wang, Q. Zhu, Q. Chen, Q. Du, R. J. Chen, R. L. Jin, R. Ge, R. Zhang, R. Pan, R. Wang, R. Xu, R. Zhang, R. Chen, S. S. Li, S. Lu, S. Zhou, S. Chen, S. Wu, S. Ye, S. Ye, S. Ma, S. Wang, S. Zhou, S. Yu, S. Zhou, S. Pan, T. Wang, T. Yun, T. Pei, T. Sun, W. L. Xiao, W. Zeng, W. Zhao, W. An, W. Liu, W. Liang, W. Gao, W. Yu, W. Zhang, X. Q. Li, X. Jin, X. Wang, X. Bi, X. Liu, X. Wang, X. Shen, X. Chen, X. Zhang, X. Chen, X. Nie, X. Sun, X. Wang, X. Cheng, X. Liu, X. Xie, X. Liu, X. Yu, X. Song, X. Shan, X. Zhou, X. Yang, X. Li, X. Su, X. Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Y. Zhang, Y. Xu, Y. Xu, Y. Huang, Y. Li, Y. Zhao, Y. Sun, Y. Li, Y. Wang, Y. Yu, Y. Zheng, Y. Zhang, Y. Shi, Y. Xiong, Y. He, Y. Tang, Y. Piao, Y. Wang, Y. Tan, Y. Ma, Y. Liu, Y. Guo, Y. Wu, Y. Ou, Y. Zhu, Y. Wang, Y. Gong, Y. Zou, Y. He, Y. Zha, Y. Xiong, Y. Ma, Y. Yan, Y. Luo, Y. You, Y. Liu, Y. Zhou, Z. F. Wu, Z. Z. Ren, Z. Ren, Z. Sha, Z. Fu, Z. Xu, Z. Huang, Z. Zhang, Z. Xie, Z. Zhang, Z. Hao, Z. Gou, Z. Ma, Z. Yan, Z. Shao, Z. Xu, Z. Wu, Z. Zhang, Z. Li, Z. Gu, Z. Zhu, Z. Liu, Z. Li, Z. Xie, Z. Song, Z. Gao, Z. Pan, Deepseek-v3 technical report (2025). arXiv:2412.19437.
URL https://arxiv.org/abs/2412.19437

[6] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Yang, A. Mitra, A. Sravankumar, A. Korenev, A. Hinsvark, A. Rao, A. Zhang, A. Rodriguez, A. Gregerson, A. Spataru, B. Roziere, B. Biron, B. Tang, B. Chern, C. Caucheteux, C. Nayak, C. Bi, C. Marra, C. McConnell, C. Keller, C. Touret, C. Wu, C. Wong, C. C. Ferrer, C. Nikolaidis, D. Allonsius, D. Song, D. Pintz, D. Livshits, D. Wyatt, D. Esiobu, D. Choudhary, D. Mahajan, D. Garcia-Olano, D. Perino, D. Hupkes, E. Lakomkin, E. AlBadawy, E. Lobanova, E. Dinan, E. M. Smith, F. Radenovic, F. Guzmán, F. Zhang, G. Synnaeve, G. Lee, G. L. Anderson, G. Thattai, G. Nail, G. Mialon, G. Pang, G. Cucurell, H. Nguyen, H. Korevaar, H. Xu, H. Touvron, I. Zarov, I. A. Ibarra, I. Kloumann, I. Misra, I. Evtimov, J. Zhang, J. Copet, J. Lee, J. Geffert, J. Vranes, J. Park, J. Mahadeokar, J. Shah, J. van der Linde, J. Billock, J. Hong, J. Lee, J. Fu, J. Chi, J. Huang, J. Liu, J. Wang, J. Yu, J. Bitton, J. Spisak, J. Park, J. Rocca, J. Johnstun, J. Saxe, J. Jia, K. V. Alwala, K. Prasad, K. Upasani, K. Plawiak, K. Li, K. Heafield, K. Stone, K. El-Arini, K. Iyer, K. Malik, K. Chiu, K. Bhalla, K. Lakhotia, L. Rantala-Yeary, L. van der Maaten, L. Chen, L. Tan, L. Jenkins, L. Martin, L. Madaan, L. Malo, L. Blecher, L. Landzaat, L. de Oliveira, M. Muzzi, M. Pasupuleti, M. Singh, M. Paluri, M. Kardas, M. Tsimpoukelli, M. Oldham, M. Rita, M. Pavlova, M. Kambadur, M. Lewis, M. Si, M. K. Singh, M. Hassan, N. Goyal, N. Torabi, N. Bashlykov, N. Bogoychev, N. Chatterji, N. Zhang, O. Duchenne, O. Çelebi, P. Alrassy, P. Zhang, P. Li, P. Vasic, P. Weng, P. Bhargava, P. Dubal, P. Krishnan, P. S. Koura, P. Xu, Q. He, Q. Dong, R. Srinivasan, R. Ganapathy, R. Calderer, R. S. Cabral, R. Stojnic, R. Raileanu, R. Maheswari, R. Girdhar, R. Patel, R. Sauvestre, R. Polidoro, R. Sumbaly, R. Taylor, R. Silva, R. Hou, R. Wang, S. Hosseini, S. Chennabasappa, S. Singh, S. Bell, S. S. Kim, S. Edunov, S. Nie, S. Narang, S. Raparthy, S. Shen, S. Wan, S. Bhosale, S. Zhang, S. Vandenhende, S. Batra, S. Whitman, S. Sootla, S. Collot, S. Gururangan,

S. Borodinsky, T. Herman, T. Fowler, T. Sheasha, T. Georgiou, T. Scialom, T. Speckbacher, T. Mihaylov, T. Xiao, U. Karn, V. Goswami, V. Gupta, V. Ramanathan, V. Kerkez, V. Gonguet, V. Do, V. Vogeti, V. Albiero, V. Petrovic, W. Chu, W. Xiong, W. Fu, W. Meers, X. Martinet, X. Wang, X. Wang, X. E. Tan, X. Xia, X. Xie, X. Jia, X. Wang, Y. Goldschlag, Y. Gaur, Y. Babaei, Y. Wen, Y. Song, Y. Zhang, Y. Li, Y. Mao, Z. D. Coudert, Z. Yan, Z. Chen, Z. Papakipos, A. Singh, A. Srivastava, A. Jain, A. Kelsey, A. Shajnfeld, A. Gangidi, A. Victoria, A. Goldstand, A. Menon, A. Sharma, A. Boesenberg, A. Baevski, A. Feinstein, A. Kallet, A. Sangani, A. Teo, A. Yunus, A. Lupu, A. Alvarado, A. Caples, A. Gu, A. Ho, A. Poulton, A. Ryan, A. Ramchandani, A. Dong, A. Franco, A. Goyal, A. Saraf, A. Chowdhury, A. Gabriel, A. Bharambe, A. Eisenman, A. Yazdan, B. James, B. Maurer, B. Leonhardi, B. Huang, B. Loyd, B. D. Paola, B. Paranjape, B. Liu, B. Wu, B. Ni, B. Hancock, B. Wasti, B. Spence, B. Stojkovic, B. Gamido, B. Montalvo, C. Parker, C. Burton, C. Mejia, C. Liu, C. Wang, C. Kim, C. Zhou, C. Hu, C.-H. Chu, C. Cai, C. Tindal, C. Feichtenhofer, C. Gao, D. Civin, D. Beaty, D. Kreymer, D. Li, D. Adkins, D. Xu, D. Testuggine, D. David, D. Parikh, D. Liskovich, D. Foss, D. Wang, D. Le, D. Holland, E. Dowling, E. Jamil, E. Montgomery, E. Presani, E. Hahn, E. Wood, E.-T. Le, E. Brinkman, E. Arcaute, E. Dunbar, E. Smothers, F. Sun, F. Kreuk, F. Tian, F. Kokkinos, F. Ozgenel, F. Caggioni, F. Kanayet, F. Seide, G. M. Florez, G. Schwarz, G. Badeer, G. Swee, G. Halpern, G. Herman, G. Sizov, Guangyi, Zhang, G. Lakshminarayanan, H. Inan, H. Shojanazeri, H. Zou, H. Wang, H. Zha, H. Habeeb, H. Rudolph, H. Suk, H. Aspegren, H. Goldman, H. Zhan, I. Damlaj, I. Molybog, I. Tufanov, I. Leontiadis, I.-E. Veliche, I. Gat, J. Weissman, J. Geboski, J. Kohli, J. Lam, J. Asher, J.-B. Gaya, J. Marcus, J. Tang, J. Chan, J. Zhen, J. Reizenstein, J. Teboul, J. Zhong, J. Jin, J. Yang, J. Cummings, J. Carvill, J. Shepard, J. McPhie, J. Torres, J. Ginsburg, J. Wang, K. Wu, K. H. U, K. Saxena, K. Khandelwal, K. Zand, K. Matosich, K. Veeraraghavan, K. Michelena, K. Li, K. Jagadeesh, K. Huang, K. Chawla, K. Huang, L. Chen, L. Garg, L. A, L. Silva, L. Bell, L. Zhang, L. Guo, L. Yu, L. Moshkovich, L. Wehrstedt, M. Khabsa, M. Avalani, M. Bhatt, M. Mankus, M. Hasson, M. Lennie, M. Reso, M. Groshev, M. Naumov, M. Lathi, M. Keneally, M. Liu, M. L. Seltzer, M. Valko, M. Restrepo, M. Patel, M. Vyatskov, M. Samvelyan, M. Clark, M. Macey, M. Wang, M. J. Hermoso, M. Metanat, M. Rastegari, M. Bansal, N. Santhanam, N. Parks, N. White, N. Bawa, N. Singhal, N. Egebo, N. Usunier, N. Mehta, N. P. Laptev, N. Dong, N. Cheng, O. Chernoguz, O. Hart, O. Salpekar, O. Kalinli, P. Kent, P. Parekh, P. Saab, P. Balaji, P. Rittner, P. Bontrager, P. Roux, P. Dollar, P. Zvyagina, P. Ratanchandani, P. Yuvraj, Q. Liang, R. Alao, R. Rodriguez, R. Ayub, R. Murthy, R. Nayani, R. Mitra, R. Parthasarathy, R. Li, R. Hogan, R. Battey, R. Wang, R. Howes, R. Rinott, S. Mehta, S. Siby, S. J. Bondu, S. Datta, S. Chugh, S. Hunt, S. Dhillon, S. Sidorov, S. Pan, S. Mahajan, S. Verma, S. Yamamoto, S. Ramaswamy, S. Lindsay, S. Lindsay, S. Feng, S. Lin, S. C. Zha, S. Patil, S. Shankar, S. Zhang, S. Zhang, S. Wang, S. Agarwal, S. Sajuyigbe, S. Chintala, S. Max, S. Chen, S. Kehoe, S. Satterfield, S. Govindaprasad, S. Gupta, S. Deng, S. Cho, S. Virk, S. Subramanian, S. Choudhury, S. Goldman, T. Remez, T. Glaser, T. Best, T. Koehler, T. Robinson, T. Li, T. Zhang, T. Matthews, T. Chou, T. Shaked, V. Vontimitta, V. Ajayi, V. Montanez, V. Mohan, V. S. Kumar, V. Mangla, V. Ionescu, V. Poenaru, V. T. Mihailescu, V. Ivanov, W. Li, W. Wang, W. Jiang, W. Bouaziz, W. Constable, X. Tang, X. Wu, X. Wang, X. Wu, X. Gao, Y. Kleinman, Y. Chen, Y. Hu, Y. Jia, Y. Qi, Y. Li, Y. Zhang, Y. Zhang, Y. Adi, Y. Nam, Yu, Wang, Y. Zhao, Y. Hao, Y. Qian, Y. Li, Y. He, Z. Rait, Z. DeVito, Z. Rosnbrick, Z. Wen, Z. Yang, Z. Zhao, Z. Ma, The llama 3 herd of models (2024). arXiv:2407.21783.
URL https://arxiv.org/abs/2407.21783

[7] K. W. CHURCH, Word2vec, Natural Language Engineering 23 (1) (2017) 155–162. doi:10.1017/S1351324916000334.

[8] J. Pennington, R. Socher, C. Manning, GloVe: Global vectors for word representation, in: A. Moschitti, B. Pang, W. Daelemans (Eds.), Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar, 2014, pp. 1532–1543. doi:10.3115/v1/D14-1162.
URL https://aclanthology.org/D14-1162/

[9] R. C. Staudemeyer, E. R. Morris, Understanding lstm – a tutorial into long short-term memory recurrent neural networks (2019). arXiv:1909.09586.
URL https://arxiv.org/abs/1909.09586

[10] H. Kotek, R. Dockum, D. Sun, Gender bias and stereotypes in large language models, in: Proceedings of The ACM Collective Intelligence Conference, CI '23, Association for Computing Machinery, New York, NY, USA, 2023, p. 12–24. doi:10.1145/3582269.3615599.
URL https://doi.org/10.1145/3582269.3615599

[11] I. Weissburg, S. Anand, S. Levy, H. Jeong, Llms are biased teachers: Evaluating llm bias in personalized education (2024). arXiv:2410.14012.

[12] L. Lin, L. Wang, J. Guo, K.-F. Wong, Investigating bias in llm-based bias detection: Disparities between llms and human perception (2024). arXiv:2403.14896.
URL https://arxiv.org/abs/2403.14896

[13] Z. Ji, T. Yu, Y. Xu, N. Lee, E. Ishii, P. Fung, Towards mitigating LLM hallucination via self reflection, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 1827–1843. doi:10.18653/v1/2023.findings-emnlp.123.
URL https://aclanthology.org/2023.findings-emnlp.123/

[14] H. Adam, A. Balagopalan, E. Alsentzer, F. Christia, M. Ghassemi, Mitigating the impact of biased artificial intelligence in emergency decision-making, Communications Medicine 2 (1) (2022) 149. doi:10.1038/s43856-022-00214-4.
URL https://doi.org/10.1038/s43856-022-00214-4

[15] K. Mei, S. Fereidooni, A. Caliskan, Bias against 93 stigmatized groups in masked language models and downstream sentiment classification tasks, in: 2023 ACM Conference on Fairness Accountability and Transparency, FAccT '23, ACM, 2023, pp. 1699–1710. doi:10.1145/3593013.3594109.

[16] T. Zack, E. Lehman, M. Suzgun, J. A. Rodriguez, L. A. Celi, J. Gichoya, D. Jurafsky, P. Szolovits, D. W. Bates, R.-E. E. Abdulnour, A. J. Butte, E. Alsentzer, Assessing the potential of gpt-4 to perpetuate racial and gender biases in health care: a model evaluation study, The Lancet Digital Health 6 (1) (2024) e12–e22. doi:10.1016/S2589-7500(23)00225-X.
URL https://doi.org/10.1016/S2589-7500(23)00225-X

[17] I. O. Gallegos, R. A. Rossi, J. Barrow, M. M. Tanjim, S. Kim, F. Dernoncourt, T. Yu, R. Zhang, N. K. Ahmed, Bias and fairness in large language models: A survey, Computational Linguistics 50 (3) (2024) 1097–1179. doi:10.1162/coli_a_00524.

[18] S. L. Blodgett, B. O'Connor, Racial disparity in natural language processing: A case study of social media african-american english (2017). doi:10.48550/ARXIV.1707.00061.

[19] B. Hutchinson, V. Prabhakaran, E. Denton, K. Webster, Y. Zhong, S. Denuyl, Social biases in nlp models as barriers for persons with disabilities, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020. doi:10.18653/v1/2020.acl-main.487.

[20] S. Ghosh, K. Wilson, Bias is a math problem, ai bias is a technical problem: 10-year literature review of ai/llm bias research reveals narrow [gender-centric] conceptions of 'bias', and academia-industry gap (2025). doi:10.48550/ARXIV.2508.11067.

[21] X. Guan, Y. Liu, H. Lin, Y. Lu, B. He, X. Han, L. Sun, Mitigating large language model

hallucinations via autonomous knowledge graph-based retrofitting, Proceedings of the AAAI Conference on Artificial Intelligence 38 (16) (2024) 18126–18134. doi:10.1609/aaai.v38i16.29770.
URL https://ojs.aaai.org/index.php/AAAI/article/view/29770

[22] R. Kumar, H. Kumar, K. Shalini, Detecting and mitigating bias in llms through knowledge graph-augmented training, in: 2025 International Conference on Artificial Intelligence and Data Engineering (AIDE), 2025, pp. 608–613. doi:10.1109/AIDE64228.2025.10987418.

[23] X. Shi, Z. Zhu, Z. Zhang, C. Li, Hallucination mitigation in natural language generation from large-scale open-domain knowledge graphs, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 12506–12521. doi:10.18653/v1/2023.emnlp-main.770.
URL https://aclanthology.org/2023.emnlp-main.770/

[24] T. R. Chhetri, Y. Chen, P. Trivedi, D. Jarecka, S. Haobsh, P. Ray, L. Ng, S. S. Ghosh, Structsense: A task-agnostic agentic framework for structured information extraction with human-in-the-loop evaluation and benchmarking (2025). arXiv:2507.03674.

[25] P. Bedi, A. Thukral, S. Dhiman, Xlr-kgdd: leveraging llm and rag for knowledge graph-based explainable disease diagnosis using multimodal clinical information, Knowledge and Information Systems (May 2025). doi:10.1007/s10115-025-02465-8.
URL https://doi.org/10.1007/s10115-025-02465-8

[26] A. Deshpande, D. Ruiter, M. Mosbach, D. Klakow, StereoKG: Data-Driven Knowledge Graph Construction for Cultural Knowledge and Stereotypes, arXiv:2205.14036 [cs] (May 2022). doi:10.48550/arXiv.2205.14036.
URL http://arxiv.org/abs/2205.14036

[27] A. Parrish, A. Chen, N. Nangia, V. Padmakumar, J. Phang, J. Thompson, P. M. Htut, S. R. Bowman, Bbq: A hand-built bias benchmark for question answering (2021). doi:10.48550/ARXIV.2110.08193.

[28] R. Kumar, H. Kumar, K. Shalini, Detecting and mitigating bias in llms through knowledge graph-augmented training, in: 2025 International Conference on Artificial Intelligence and Data Engineering (AIDE), IEEE, 2025, pp. 608–613. doi:10.1109/aide64228.2025.10987418.

[29] Y. Zhang, M. Jiang, Q. Zhao, GRACE: Graph-Based Contextual Debiasing for Fair Visual Question Answering, Springer Nature Switzerland, 2024, pp. 176–194. doi:10.1007/978-3-031-72643-9_11.

[30] N. Ibrahim, S. Aboulela, A. Ibrahim, R. Kashef, A survey on augmenting knowledge graphs (kgs) with large language models (llms): models, evaluation metrics, benchmarks, and challenges, Discover Artificial Intelligence 4 (1) (Nov. 2024). doi:10.1007/s44163-024-00175-8.

[31] A. Kumar, V. Aelgani, R. Vohra, S. K. Gupta, M. Bhagawati, S. Paul, L. Saba, N. Suri, N. N. Khanna, J. R. Laird, A. M. Johri, M. Kalra, M. M. Fouda, M. Fatemi, S. Naidu, J. S. Suri, Artificial intelligence bias in medical system designs: a systematic review, Multimedia Tools and Applications 83 (6) (2023) 18005–18057. doi:10.1007/s11042-023-16029-x.

[32] S. Akter, G. McCarthy, S. Sajib, K. Michael, Y. K. Dwivedi, J. D'Ambra, K. N. Shen, Algorithmic bias in data-driven innovation in the age of AI, International Journal of Information Management 60 (2021) 102387. doi:10.1016/j.ijinfomgt.2021.102387.
URL https://www.sciencedirect.com/science/article/pii/S0268401221000803

[33] P. Bhandari, Sampling Bias and How to Avoid It | Types & Examples (May 2020).
URL https://www.scribbr.com/research-bias/sampling-bias/

[34] N. Shahbazi, Y. Lin, A. Asudeh, H. V. Jagadish, Representation Bias in Data: A Survey on Identification and Resolution Techniques, ACM Comput. Surv. 55 (13s) (2023) 293:1–293:39. doi:10.1145/3588433.
URL https://doi.org/10.1145/3588433

[35] F. J. Oort, M. R. M. Visser, M. A. G. Sprangers, Formal definitions of measurement bias and explanation bias clarify measurement and conceptual perspectives on response shift, Journal of Clinical Epidemiology 62 (11) (2009) 1126–1137. doi:10.1016/j.jclinepi.2009.03.013.
URL https://www.sciencedirect.com/science/article/pii/S0895435609000961

[36] K. Nikolopoulou, What Is Information Bias? | Definition & Examples (Nov. 2022).
URL https://www.scribbr.com/research-bias/information-bias/

[37] Different Types of Bias in Research - CASP.
URL https://casp-uk.net/news/different-types-of-research-bias/

[38] A. Khademi, V. Honavar, Algorithmic bias in recidivism prediction: A causal perspective, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, 2020, pp. 13839–13840. doi:10.1609/aaai.v34i10.7192.

[39] M. Mansoury, H. Abdollahpouri, M. Pechenizkiy, B. Mobasher, R. Burke, Feedback loop and bias amplification in recommender systems, in: Proceedings of the 29th ACM International Conference on Information amp; Knowledge Management, CIKM '20, ACM, , 2020, pp. 2145–2148. doi:10.1145/3340531.3412152.

[40] Y. Du, Confirmation bias in generative ai chatbots: Mechanisms, risks, mitigation strategies, and future research directions (2025). doi:10.48550/ARXIV.2504.09343.

[41] M. Zhou, V. Abhishek, T. Derdenger, J. Kim, K. Srinivasan, Bias in generative ai, arXiv preprint arXiv:2403.02726, preprint, not peer-reviewed (Mar. 2024). arXiv:2403.02726, doi:10.48550/ARXIV.2403.02726.
URL https://arxiv.org/abs/2403.02726

[42] 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. doi:10.18653/v1/n18-2003.

[43] 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). doi:10.48550/ARXIV.2410.07589.