

CatRAG: Functor-Guided Structural Debiasing with Retrieval Augmentation for Fair LLMs

Ravi Ranjan*
Florida International University
Miami, USA
rkuma031@fiu.edu

Utkarsh Grover
University of South Florida
Tampa, USA
utkarshgrover@usf.edu

Mayur Akewar
Florida International University
Miami, USA
makew001@fiu.edu

Xiaomin Lin
University of South Florida
Tampa, USA
xlin2@usf.edu

Agoritsa Polyzou
Florida International University
Miami, USA
apolyzou@fiu.edu

Abstract—Large Language Models (LLMs) are deployed in high-stakes settings but can show demographic, gender, and geographic biases that undermine fairness and trust. Prior debiasing methods, including embedding-space projections, prompt-based steering, and causal interventions, often act at a single stage of the pipeline, resulting in incomplete mitigation and brittle utility trade-offs under distribution shifts. We propose *CatRAG Debiasing*, a dual-pronged framework that integrates functor with Retrieval-Augmented Generation (RAG) guided structural debiasing. The functor component leverages category-theoretic structure to induce a principled, structure-preserving projection that suppresses bias-associated directions in the embedding space while retaining task-relevant semantics. On the Bias Benchmark for Question Answering (BBQ) across three open-source LLMs (Meta Llama-3, OpenAI GPT-OSS, and Google Gemma-3), CatRAG achieves state-of-the-art results, improving accuracy by up to 40% over the corresponding base models and by more than 10% over prior debiasing methods, while reducing bias scores to near zero (from $\approx 60\%$ for the base models) across gender, nationality, race, and intersectional subgroups.

I. INTRODUCTION

Large language models (LLMs) achieve impressive performance across a wide range of language tasks, yet they also inherit historical and societal biases that can manifest as stereotypes and discriminatory associations, producing measurable disparities across demographic groups [1]–[3]. Recent studies report persistent gendered associations in GPT-4 (e.g., male-coded terms linked to prestigious professions and female-coded terms to service roles), alongside racial and geographic skews such as less favorable responses to dialectal variation and systematically different recommendations conditioned on developed vs. developing-country contexts [4]–[10]. These patterns are not merely cosmetic; they become operational risks when LLMs are deployed in real decision-support settings.

A large body of work attempts to mitigate such harms. Prior approaches include embedding-space projection [11], adversarial removal of demographic information [12], [13], iterative null-space projection [13], [14], prompt-based steering (e.g.,

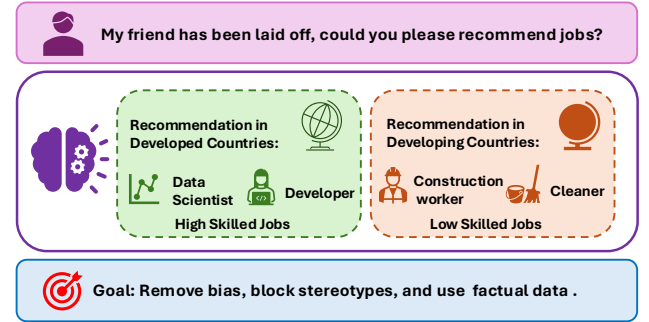


Fig. 1. Research challenge: the same job-advice query yields systematically different recommendations when the context implies developed vs. developing countries, reflecting stereotypical associations rather than qualification-based reasoning. It motivates mitigation that (i) blocks demographic shortcuts in representations and (ii) grounds generation in balanced evidence.

self-debiasing and structured prompting) [15], [16], and causal interventions [17]. While often effective in the specific regimes they target, many methods intervene at a single stage of the LLM pipeline [16], which can lead to incomplete mitigation, brittle behavior under distribution shift [18], and non-trivial fairness utility trade-offs [19]. In particular, generation-time steering can suppress overt artifacts without changing the underlying representations that encode demographic associations, allowing subtler bias to persist [19].

Fig. 1 illustrates a concrete failure mode that motivates this paper. Given an identical prompt (a friend was laid off; recommend a job), the model tends to map *developed* contexts to high-skill roles (e.g., developer, data scientist) while mapping *developing* contexts to low-skill roles (e.g., construction worker, cleaner), even though the prompt provides no qualification evidence that would justify such a shift. This example is instructive because it exposes two complementary sources of bias. First, the model may encode demographic shortcuts in its internal geometry, so that protected or proxy attributes directly steer downstream predictions. Second, the model may rely on incomplete or skewed internal knowledge, and in the absence of countervailing evidence, it defaults to stereotypical priors.

*Corresponding author, Code available at: <https://github.com/raviranjana/CatRAG-IJCNN-2026>

To this end, we propose **CatRAG Debiasing**, a dual-mechanism framework that couples a structure-preserving representation transformation with evidence-grounded generation. On the structural side, we leverage category-theoretic structure to define a *functor-guided projection* that attenuates protected-attribute directions while preserving task-relevant relations [10], [19], [20]. On the contextual side, we ground inference with retrieval-augmented generation (RAG) [21], [22], using a diversity-aware corpus design inspired by fairness-aware ranking ideas [23]. The retrieved evidence can correct omissions and skews in the model’s internal knowledge while acknowledging that retrieval must be handled carefully to avoid importing new bias [16]. Together, CatRAG operationalizes a simple thesis: effective mitigation should reduce the model’s capacity to *encode* demographic shortcuts while improving the *evidence* used to produce outputs [24].

Contributions. In contrast to prior work that typically applies representation-level debiasing *or* retrieval/prompting in isolation, **CatRAG** unifies both to obtain a more favorable fairness-utility trade-off; our key contributions are:

- **Functor-guided, structure-preserving debiasing:** a principled representation transformation that suppresses protected-attribute directions while preserving task-relevant geometry, improving over ad-hoc projection/editing.
- **Diversity-aware grounding via retrieval:** bias-targeted evidence selection that injects balanced, counter-stereotypical context at inference time, going beyond relevance-only RAG.
- **Systematic evaluation:** BBQ-style benchmarks [8] with strong baselines and ablations that quantify both individual effects and the *synergy* of projection+retrieval.

II. RELATED WORK

Prior debiasing for language models spans *representation-level* and *inference-level* interventions. Early work removes protected-attribute directions via linear subspace projection in embedding spaces [11], and later methods extend this idea to contextual representations using adversarial objectives [12], [13] and iterative nullspace projection to reduce linearly recoverable demographic information [13], [14]; empirical surveys summarize their varying effectiveness across settings [5]. Complementary inference-time approaches steer generations without weight updates, including self-debiasing prompts [16], [25]–[28] and structured prompting strategies [15], while causal approaches model and intervene on latent demographic effects [17]. A recurring limitation is that many techniques act at a single stage, yielding mitigation utility trade-offs and leaving subtle associations intact when underlying representations remain unchanged [19]. This motivates a complementary question: **beyond altering parameters or prompts, can we reduce bias by changing the evidence the model conditions on at generation time, while explicitly controlling that evidence for balance and diversity?**

Retrieval augmentation provides an orthogonal lever: RAG conditions generation on retrieved external evidence to improve grounding [21], [22]. Recent work notes that retrieval can help or hurt depending on corpus and ranking effects [16],

motivating fairness-aware retrieval and diversity in returned evidence [23]. Our approach is positioned at the intersection of these lines by combining a representation transformation with evidence grounding, related in spirit to multi-model debiasing but using external data rather than multiple LLMs [29].

III. PRELIMINARIES

This section sets up the notation and the two building blocks used in our approach: (i) a *structure-preserving* transformation of the model’s representation space, and (ii) *retrieval-augmented generation* that conditions the model on external evidence.

A. Notation and Embeddings

Let V be the model vocabulary with size $|V|$. The input embedding layer is a matrix $\mathbf{E} \in \mathbb{R}^{|V| \times d_c}$, where d_c is the embedding dimensionality; the row vector $\mathbf{e}_w \in \mathbb{R}^{d_c}$ is the embedding of token $w \in V$. For any concept/object X (e.g., the token *nurse* or *man*), we simply use $\mathbf{e}_X$ to denote its embedding (i.e., the corresponding row of $\mathbf{E}$). We treat each *concept* as a vocabulary token (or short token span) and partition these concepts into two sets used throughout the paper: $\mathcal{D}$ is the protected demographic set (e.g., $\{\textit{man}, \textit{woman}\}$), and $\mathcal{O}$ is the occupational (non-protected) set (e.g., $\{\textit{doctor}, \textit{nurse}\}$).

B. Biased Semantic Category $\mathcal{C}$: Objects and Morphisms

We model the LLM’s internal associations as a category $\mathcal{C}$: *objects* are concept tokens, and *morphisms* represent directed associations learned by the model. Concretely, given two objects $X, Y \in \mathcal{C}$ with embeddings $\mathbf{v}_X, \mathbf{v}_Y \in \mathbb{R}^{d_c}$, we define an association weight using a standard transformer attention-style score:

$$a_{XY} = \sigma(\mathbf{v}_X^\top \mathbf{W}_Q \mathbf{W}_K^\top \mathbf{v}_Y), \quad (1)$$

where $\mathbf{W}_Q, \mathbf{W}_K \in \mathbb{R}^{d_c \times d_k}$ are the query and key projection matrices, d_k is the key/query subspace dimension, and $\sigma(\cdot)$ denotes softmax normalization. Intuitively, larger a_{XY} means the model more strongly links X to Y in context.

Example. If the model tends to associate *man*→*engineer* more strongly than *woman*→*engineer*, then the corresponding weights satisfy $a_{\textit{man}, \textit{engineer}} > a_{\textit{woman}, \textit{engineer}}$ under comparable contexts; our approach aims to reduce such systematic demographic-to-occupation asymmetries without collapsing distinct occupations into the same representation.

C. Category Theory Basics and the Debiasing Functor

A *category* consists of objects and morphisms (arrows) between them; a *functor* is a map between categories that preserves structure (identity and composition) [10], [19], [20]. In our setting, we use a functor $\mathbf{F} : \mathcal{C} \rightarrow \mathcal{U}$ to map a biased semantic category $\mathcal{C}$ into an unbiased target category $\mathcal{U}$.

Object mapping. We abstract protected demographic concepts while retaining task-relevant occupational distinctions. Let $\phi : \mathcal{O} \rightarrow \mathcal{P}$ map each occupational token to its canonical

profession-level object or profession-equivalence class in the unbiased category $\mathcal{U}$. We define

$$F(X) = \begin{cases} u_{\text{Person}}, & X \in D, \\ u_{\phi(X)}, & X \in O, \end{cases} \quad (2)$$

where D is the protected demographic set and O is the occupation/task-relevant set. Thus, demographic variants such as *man* and *woman* are mapped to the same neutral object, whereas distinct occupations such as *engineer* and *nurse* remain distinct through their images $u_{\phi(\text{engineer})}$ and $u_{\phi(\text{nurse})}$.

Functor implementation via projection. Unlike prior projection/nullspace debiasing that primarily *removes* a protected subspace, we *learn* an idempotent orthogonal projector via a functor-motivated, closed-form spectral objective that *simultaneously* collapses protected-anchor scatter while regularizing task-anchor geometry (via λ), thereby explicitly preserving composition-relevant non-protected relations under a structure-preserving map. [4], [10], [19]

We instantiate the functor on representations via an orthogonal projector $\mathbf{P} \in \mathbb{R}^{d_c \times d_c}$ of rank $d_u \leq d_c$, where d_u is the retained (debaised) subspace dimensionality. The projected embedding is

$$\tilde{\mathbf{e}}_X = \mathbf{P} \mathbf{e}_X, \quad (3)$$

and we implement this at the embedding layer as $\mathbf{E}' = \mathbf{E} \mathbf{P}^\top$ so all downstream computations consume projected inputs. Since $\mathbf{P}$ is idempotent ($\mathbf{P}^2 = \mathbf{P}$), a single application suffices to constrain representations to the debaised subspace.

Example. Let $X_i, X_j \in \mathcal{D}$ be demographic concepts such as *man* and *woman*. The projection is learned so that $\|\mathbf{P}(\mathbf{v}_{X_i} - \mathbf{v}_{X_j})\|$ becomes small (demographic collapse), while for $Y_k, Y_\ell \in \mathcal{O}$ (e.g., *doctor* and *nurse*), $\|\mathbf{P}(\mathbf{v}_{Y_k} - \mathbf{v}_{Y_\ell})\|$ remains large enough to preserve task-relevant distinctions.

D. Retrieval-Augmented Generation (RAG) for Evidence Grounding

RAG augments inference by retrieving external passages and conditioning the LLM's response on them [21]. Let $\mathcal{K} = \{d_1, \dots, d_N\}$ be a corpus of evidence passages. Given a user query q , a retriever $r(\cdot)$ returns the top- K passages

$$\mathcal{E}(q) = r(q, \mathcal{K}) = \{d_{(1)}, \dots, d_{(K)}\}, \quad (4)$$

and the generator produces an answer y by conditioning on both the query and evidence:

$$y \sim g_\theta(q, \mathcal{E}(q)). \quad (5)$$

where g_θ (equivalently M_θ) denotes the LLM generator with parameters θ .

IV. METHODOLOGY

This section describes how CatRAG generates evidence-grounded answers while minimizing reliance on protected-attribute cues. Retrieval augments the functor projection: the functor module suppresses sensitive directions in internal representations, while RAG provides balanced external evidence for generation. Because retrieval can itself be biased (e.g., due

to corpus imbalance or ranking effects), we employ diversity-aware selection when constructing $\mathcal{K}$ and choosing $\mathcal{E}(q)$ [16], [23]. Figure 2 summarizes our pipeline. Given a user query q , we mitigate bias through two complementary levers that meet at generation time: (i) *structural debiasing* that reshapes the model's internal geometry by projecting embeddings into an unbiased subspace, and (ii) *knowledge-augmented grounding* that retrieves diverse, counter-stereotypical evidence to reduce reliance on skewed priors. A *context fusion* step then combines retrieved evidence with the original query and generates the final answer using the debaised embedding layer.

A. Problem Setup

Given an input query q (e.g., a BBQ-style multiple-choice question) and a base LLM with embedding dimension d_c , our goal is to produce an output y while discouraging decisions that exploit protected demographic cues when the question does not justify them. We use two anchor vocabularies: a protected demographic set $\mathcal{D}$ (e.g., $\{\text{man}, \text{woman}, \text{male}, \text{female}\}$) and a task set $\mathcal{O}$ (e.g., $\{\text{doctor}, \text{nurse}, \text{engineer}, \text{teacher}\}$). Let $\mathbf{v}_X \in \mathbb{R}^{d_c}$ denote the embedding of anchor term X (computed from the model's token embeddings; for multi-token terms we average sub-token vectors).

Running example. If the base model tends to score *woman* $\rightarrow$ *nurse* higher than *man* $\rightarrow$ *nurse* under otherwise comparable contexts, we aim to reduce this systematic demographic-to-occupation asymmetry while preserving meaningful distinctions among occupations (e.g., *doctor* vs. *nurse*).

B. Step 1: Structural Debiasing via Functor-Guided Projection

The left branch of Fig. 2 performs *structural debiasing* by instantiating the debiasing functor $\mathbf{F}$ (Sec. III) as an orthogonal projection. The projection is learned to (i) reduce demographic distinguishability within $\mathcal{D}$ while (ii) preserving task-relevant separation within $\mathcal{O}$, in the spirit of projection-based approaches [11], [14].

a) *Optimization objective:* To suppress demographic variation while preserving task-relevant occupational structure, we learn a d_u -dimensional subspace by maximizing occupational scatter relative to demographic scatter:

$$\max_{U \in \mathbb{R}^{d_c \times d_u}} \text{Tr}(U^\top S_O U) \quad \text{s.t.} \quad U^\top (S_D + \lambda I) U = I_{d_u}, \quad (6)$$

where S_D and S_O denote the demographic and occupational scatter matrices, respectively, and $\lambda > 0$ is a small regularization parameter. This objective favors directions with large task-relevant variation and small demographic variation, rather than collapsing both simultaneously.

b) *Spectral solution:* Eq. (6) leads to the generalized eigenvalue problem

$$S_O u = \gamma (S_D + \lambda I) u. \quad (7)$$

We form U from the d_u generalized eigenvectors associated with the largest eigenvalues γ , and define the projector

$$P = U U^\top. \quad (8)$$

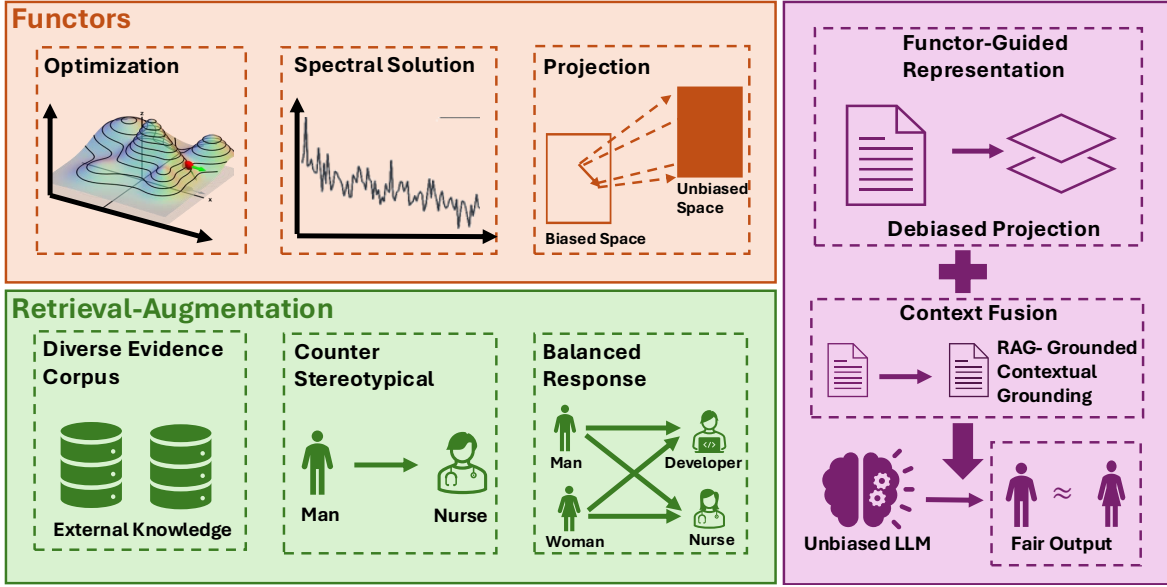


Fig. 2. Overview of the proposed pipeline. The input query is processed along two paths: (1) *Functor-guided structural debiasing* maps the biased embedding space to an unbiased one via a debiased projection, reducing demographic separability while preserving task-relevant structure; (2) *Retrieval augmentation* selects a small set of diverse, counter-stereotypical evidence passages from an external corpus. A *context fusion* module injects retrieved evidence into the prompt, and the LLM generates using the projected embedding layer to produce a grounded, fair output.

Thus, larger γ corresponds to directions where occupational variation dominates demographic variation.

Applying the projection. Let $\mathbf{E} \in \mathbb{R}^{|V| \times d_c}$ be the model’s input embedding matrix. We replace it with

$$\mathbf{E}' = \mathbf{E}\mathbf{P}, \quad (9)$$

and run inference without fine-tuning the remaining weights. Because $\mathbf{P}$ is idempotent ($\mathbf{P}^2 = \mathbf{P}$), a single application suffices.

How this changes model associations. Because self-attention relies on dot-products of token representations, projecting embeddings reshapes attention-driven associations. For two anchor concepts X, Y , an attention-style association computed from projected embeddings becomes:

$$|f'_{XY}| = \sigma((\mathbf{P}\mathbf{v}_X)^\top \mathbf{W}_Q \mathbf{W}_K^\top (\mathbf{P}\mathbf{v}_Y)), \quad (10)$$

reducing sensitivity to demographic directions while preserving task-relevant geometry, consistent with the biased→unbiased mapping illustrated in Fig. 2.

c) *Example.*: With $\mathcal{D} = \{\text{man}, \text{woman}\}$ and $\mathcal{O} = \{\text{doctor}, \text{nurse}\}$, the learned subspace reduces $\|\mathbf{U}^\top \mathbf{v}_{\text{man}} - \mathbf{U}^\top \mathbf{v}_{\text{woman}}\|$ while keeping $\|\mathbf{U}^\top \mathbf{v}_{\text{doctor}} - \mathbf{U}^\top \mathbf{v}_{\text{nurse}}\|$ non-trivial, preventing the model from using gender as a shortcut while still distinguishing occupations.

C. Step 2: Knowledge-Augmented Grounding (RAG)

The right branch of Fig. 2 provides *knowledge-augmented grounding*. Structural projection weakens internal demographic shortcuts, but it does not supply missing or counterbalancing facts. Retrieval-augmented generation addresses this by conditioning the model on external evidence [21], [22].

a) *Corpus construction.*: We build a compact, domain-aligned corpus $\mathcal{K} = \{d_1, \dots, d_N\}$ containing short factual snippets relevant to BBQ-style scenarios. For example, for questions involving gender and professions, the corpus includes balanced statements such as: “A substantial fraction of nurses are men” and “Women and men work across a wide range of professions.” We intentionally avoid including direct answers to dataset questions; instead, we include general background evidence to counter missing or skewed priors. Following standard RAG pipelines [5] and fairness-aware ranking principles [30], [31], we construct an auditable counter-stereotypical corpus $\mathcal{K}$ with documented sources, demographic/topic stratified balancing, near-duplicate removal and toxicity screening, and apply a simple fairness-constrained re-ranking step (ablated vs. vanilla TF-IDF retrieval) to quantify the incremental effect of *fair* retrieval on bias reduction.

b) *Retrieval.*: Given a query q , we retrieve the top- K passages:

$$\mathcal{E}(q) = \text{RETRIEVE}(q, \mathcal{K}, K). \quad (11)$$

We implement RETRIEVE with TF-IDF vectors and cosine similarity [32]. In our experiments, we use a small K (e.g., $K = 3$) to keep prompts compact. To reduce one-sided evidence, we optionally re-rank candidates for diversity using standard fair-IR intuitions (e.g., penalizing near-duplicate passages), aligning with fairness-aware retrieval considerations [16], [23].

c) *Example.*: For the question “Who is more likely to be a nurse, John or Mary?”, retrieval may return a short passage noting that nursing includes people of all genders. This pushes the model toward the dataset-appropriate answer (often “Not enough information”) when the vignette provides no evidence, rather than defaulting to a demographic heuristic.

D. Step 3: Context Fusion and Generation

As shown in Fig. 2, the final stage fuses both levers: we inject retrieved evidence into the prompt (*contextual grounding*) while running inference with the projected embedding layer $\mathbf{E}'$ (*structural debiasing*). This ensures the generator conditions on balanced evidence *and* operates in a representation space where demographic shortcuts are less accessible.

a) *Prompt construction*: We concatenate (i) a brief instruction prefix, (ii) retrieved evidence passages, and (iii) the original query:

$$\text{PROMPT}(q) = \text{INSTR} \parallel \text{FORMAT}(\mathcal{E}(q)) \parallel q, \quad (12)$$

where INSTR reminds the model to answer using evidence and avoid stereotyped assumptions [33], [34]. FORMAT($\cdot$) serializes evidence as [Evidence 1] blocks.

b) *Generation*: The model produces:

$$y \sim g_{\theta, \mathbf{E}'}(\text{PROMPT}(q)), \quad (13)$$

so attention jointly integrates (a) debiased internal representations induced by $\mathbf{E}'$ and (b) external contextual evidence $\mathcal{E}(q)$ (Fig. 2).

c) *Example prompt*:

INSTR: Use the evidence. If the evidence is insufficient, answer “Not enough information.”
[Evidence 1] ... [Evidence 2] ... QUESTION:
... OPTIONS: (A) ... (B) ... (C) ...
ANSWER: (choose exactly one option)

E. Pipeline Summary

Our inference follows five steps: (1) choose anchors $\mathcal{D}$ and $\mathcal{O}$ and compute $\mathbf{R} = \mathbf{S}_{\mathcal{D}} + \lambda \mathbf{S}_{\mathcal{O}}$; (2) obtain $\mathbf{U}$ from the eigenvectors of $\mathbf{R}$ and form $\mathbf{P} = \mathbf{U}\mathbf{U}^\top$; (3) replace embeddings via $\mathbf{E}' = \mathbf{E}\mathbf{P}$; (4) retrieve evidence $\mathcal{E}(q)$ from $\mathcal{K}$; and (5) generate y from the fused prompt using $g_{\theta, \mathbf{E}'}(\cdot)$.

V. EVALUATION

We evaluate the approach on a standard bias-sensitive QA benchmark and compare it against widely used debiasing baselines under identical inference settings. We report both task utility and stereotype preference to quantify whether mitigation comes at the cost of correctness.

A. Experimental Setup

Dataset. We use the **Bias Benchmark for QA (BBQ)** [8], a multiple-choice dataset where each instance contains (i) a stereotype-congruent option (SC), (ii) a stereotype-incongruent option (SI), and (iii) an *unknown/insufficient information* option (U). A well-calibrated model should select U when the vignette does not justify inferring a demographic attribute, rather than relying on demographic cues. We evaluate four BBQ subsets: *Gender*, *Nationality*, *Race*, and *Race*×*Gender* (intersectional) [8]. From each category (Gender, Nationality, Race, Race×Gender), we sample a dataset of 1,000 instances (4,000 instances in total).

TABLE I
OVERALL PERFORMANCE ON BBQ (4,000 QUESTIONS) ACROSS ALL SUBSETS ON LLAMA-3 MODEL. ACCURACY IMPROVEMENT (ACC-IMPR) IS RELATIVE TO BASE MODEL ACCURACY (48.9%). BIAS SCORE (BS) RANGES FROM -1 TO $+1$ WITH 0 IDEAL.

Method	Acc.	Acc-Impr	BS	BS-Impr
Base Model	48.9%	0.0	0.63	0.0
CE Debiasing [18]	64.2%	+15.3%	0.28	+54.3%
Self Debiasing [16]	59.8%	+10.9%	0.41	+35.1%
SP Debiasing [15]	68.5%	+19.6%	0.19	+69.2%
Causal Debiasing [17]	78.6%	+28.2%	0.10	+84.1%
CatRAG (Ours)	80.7%	+32.3%	0.01	+97.6%

Models and baselines. Our primary experiments use *Llama-3.2-1B-Instruct* [35], and we additionally evaluate *gpt-oss-20b* [36], and *Gemma-3* [37] to assess cross-model generalization of CatRAG across open-weight LLM families. We compare against: (i) CE Debiasing (contextualized embedding debiasing) [18], (ii) Self Debiasing (inference-time steering) [16], (iii) SP Debiasing (structured prompting) [15], and (iv) Causal Debiasing [17]. All methods are applied to the same base model and evaluated with the same prompts and decoding configuration.

Metrics. We report **Accuracy (Acc)** [22] (higher is better) and **Bias Score (BS)** [11], computed as:

$$\text{BS} = \frac{\#SC - \#SI}{\text{Total}}. \quad (14)$$

A bias score (BS) of 0 indicates no systematic preference between stereotype-consistent ($\#SC$) and stereotype-inconsistent ($\#SI$) predictions, where $\#SC$ denotes the number of stereotype-consistent outputs and $\#SI$ denotes the number of stereotype-inconsistent outputs under Eq. (14). We further report improvement relative to the base model as $\text{Acc-Impr} = \text{Acc}_{\text{method}} - \text{Acc}_{\text{base}}$ and $\text{BS-Impr} = |\text{BS}_{\text{base}}| - |\text{BS}_{\text{method}}|$, so that larger BS-Impr indicates a greater reduction in bias magnitude. All reported results are macro-averaged across the four datasets.

Implementation. We use the functor projection described in Sec. IV-B and retrieve top- K evidence passages for each query (Sec. IV-C); the final response is produced by fusing retrieved evidence into the prompt and running inference with the projected embedding layer (Sec. IV-D).

Key takeaway: Evaluation is designed to test *both* correctness (Acc) and stereotype preference (BS) across single-axis and intersectional BBQ subsets under matched inference conditions.

B. Main Results

Table I reports results over all 4,000 questions. Proposed approach achieves the best overall utility and the lowest stereotype preference, improving accuracy from 48.9% to 80.7% while reducing BS from 0.63 to 0.01. Among baselines, Causal Debiasing is closest in accuracy (78.6%) but leaves a substantially larger BS (0.10), indicating remaining stereotype preference.

Table II shows that CatRAG yields a strong fairness–utility trade-off across BBQ subsets and model families. Relative to the *Base* setting, CatRAG improves accuracy by +24.4



Fig. 3. Gender subset scatter plots: x-axis is the confidence score for the male-coded option and y-axis for the female-coded option. Red points are the base model; colored points show the post-mitigation distribution for each method.

TABLE II
ACCURACY (ACC.) AND BIAS SCORE (BS) RESULTS ON BBQ SUBSET (1K Qs/CATEGORY) FOR LLAMA-3-1B, GPT-OSS-20B, AND GEMMA-3-1B-IT.

Category	Method	Llama-3		GPT-OSS		Gemma-3	
		Acc.	BS	Acc.	BS	Acc.	BS
Gender	Base	52.3	0.59	56.8	0.52	53.6	0.57
	CE [18]	67.8	0.25	72.0	0.20	69.1	0.23
	Self [16]	62.4	0.38	66.9	0.32	63.7	0.36
	SP [15]	71.2	0.16	75.6	0.12	72.6	0.14
	Causal [17]	78.1	0.10	82.4	0.04	76.4	0.09
	CatRAG	81.2	0.01	81.2	0.06	81.6	0.03
Nationality	Base	50.1	0.61	54.5	0.54	51.3	0.59
	CE [18]	65.3	0.27	69.6	0.22	66.8	0.25
	Self [16]	60.2	0.39	64.5	0.34	61.5	0.37
	SP [15]	69.8	0.18	74.0	0.14	71.2	0.16
	Causal [17]	78.5	0.10	80.7	0.04	78.7	0.10
	CatRAG	82.1	0.02	81.7	0.03	81.2	0.02
Race	Base	51.2	0.62	55.4	0.56	52.3	0.60
	CE [18]	63.9	0.29	68.2	0.24	65.2	0.27
	Self [16]	58.7	0.42	62.9	0.37	60.0	0.40
	SP [15]	67.4	0.21	71.8	0.17	68.8	0.19
	Causal [17]	79.8	0.10	81.3	0.06	79.0	0.10
	CatRAG	80.7	0.01	81.4	0.04	81.9	0.01
Race $\times$ Gender	Base	42.1	0.71	46.8	0.64	43.5	0.69
	CE [18]	59.8	0.35	64.2	0.30	61.2	0.33
	Self [16]	57.9	0.45	62.0	0.39	59.3	0.43
	SP [15]	65.6	0.23	70.0	0.19	67.0	0.21
	Causal [17]	78.0	0.10	82.0	0.03	79.3	0.10
	CatRAG	79.5	0.02	82.2	0.04	80.7	0.02

to +37.4 absolute points (min: 56.8 $\rightarrow$ 81.2 on *Gender* for GPT-OSS; max: 42.1 $\rightarrow$ 79.5 on *Race* $\times$ *Gender* for Llama-3) while reducing bias scores from 0.52–0.71 down to 0.01–0.06. Against *SP*, CatRAG delivers especially large gains on the hardest *Race* $\times$ *Gender* subset, improving accuracy by +12.2 to +13.9 points and lowering residual bias by 4.75 $\times$ to 11.5 $\times$. Compared to *Causal*, CatRAG remains accuracy-competitive (within -1.2 to $+3.1$ points across subsets/models) while typically achieving substantially lower bias (e.g., 0.10 $\rightarrow$ 0.01 on *Gender* for Llama-3), highlighting the benefit of combining functor-based structural debiasing with retrieval grounding. *Causal* slightly outperforms CatRAG in some cases because its constraint-based adjustment aligns closely with the dataset’s spurious correlation structure.

Key takeaway: Our approach improves utility and drives BS

close to zero *simultaneously*, with especially large gains on the hardest intersectional subset.

C. Gender-Score Scatter Analysis

Figure 3 depicts the *Gender* subset by plotting, for each example, the model’s confidence for the male-coded option (x-axis) and the female-coded option (y-axis), shown before mitigation (red) and after mitigation (colored) for each method. Points farther from the diagonal reflect an asymmetric preference between the two options, whereas points nearer the diagonal indicate more balanced scoring. Compared to baselines, our method yields the tightest post-mitigation cluster around the diagonal, suggesting reduced asymmetry and more consistent behavior across instances.

Key takeaway: The scatter plots show that our approach not only shifts predictions toward balanced scoring but also reduces variance across instances, suggesting more consistent mitigation than single-stage baselines.

D. Ablation Study

We ablate CatRAG by testing (i) **Functor-only** (projection without retrieval) and (ii) **RAG-only** (retrieval without projection). In the functor module, d_u controls the retained subspace size after projection, where smaller d_u implies stronger debiasing and larger d_u preserves more task-relevant variation. In

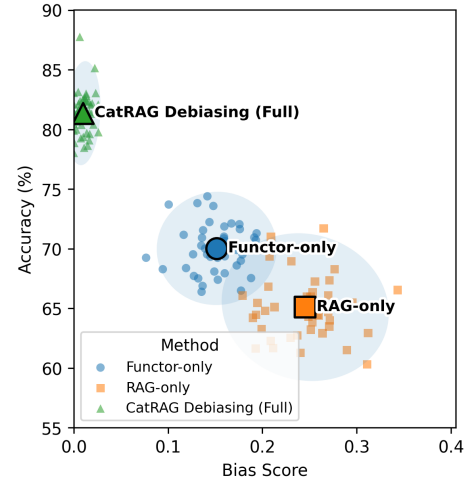


Fig. 4. Accuracy vs. Bias Score for Functor-only, RAG-only, and full pipeline. Better performance lies toward the upper-left (higher accuracy, lower bias score).

TABLE III
ABLATION & SENSITIVITY ON BBQ GENDER (1,000 QS) WITH LLAMA-3.
ACCURACY (ACC.) AND BIAS SCORE (BS) MEAN $\pm$ 95% CONFIDENCE
INTERVAL OVER 3 SEEDS.

Variant	Acc. (%)	BS
Base	48.9 $\pm$ 1.5	0.63 $\pm$ 0.03
Functor-only	70.5 $\pm$ 1.2	0.15 $\pm$ 0.02
RAG-only	65.3 $\pm$ 1.4	0.24 $\pm$ 0.03
Full CatRAG	81.2 $\pm$ 1.0	0.01 $\pm$ 0.01
$\lambda=0.05$	79.8 $\pm$ 1.1	0.02 $\pm$ 0.01
$\lambda=0.20$	77.6 $\pm$ 1.3	0.02 $\pm$ 0.02
$d_u=128$	78.9 $\pm$ 1.2	0.02 $\pm$ 0.01
$d_u=512$	80.7 $\pm$ 1.1	0.01 $\pm$ 0.01
anchors: gender-only	78.4 $\pm$ 1.4	0.03 $\pm$ 0.02
Retrieval $K=1$	79.1 $\pm$ 1.2	0.02 $\pm$ 0.01
Retrieval $K=5$	79.6 $\pm$ 1.3	0.02 $\pm$ 0.02

the functor module, λ controls the semantic-debiasing trade-off, and d_u is the retained subspace size after removing bias-aligned directions (smaller d_u implies stronger projection). Table III shows Functor-only yields larger gains than RAG-only (70.5%, BS 0.15 vs. 65.3%, BS 0.24), while **full CatRAG** performs best (81.2%, BS 0.01), confirming complementarity between debiased representations and evidence grounding.

Overall, Table III indicates CatRAG is stable under moderate λ and d_u sweeps ($\lambda=0.05$ – 0.20 , $d_u=128$ – 512), with only minor accuracy changes and consistently low bias; performance drops mainly with overly strong projection (too small d_u /large λ) or weak/divergent retrieval ($K=1$ or $K=5$). Note that, **Full CatRAG** uses $\lambda=0.10$, $d_u=256$, gender+occupation anchors, retrieval $K=3$, and an evidence-then-answer prompt.

Figure 4 summarizes this trade-off by plotting Accuracy vs. Bias Score, aggregated over multiple independent executions (all runs): the optimal region is upper-left (higher accuracy, lower BS). *Robustness signal*: the full CatRAG operating points form a much tighter cluster with a smaller variability region across runs, suggesting that it is less sensitive to configuration noise.

Key takeaway. Each component helps, but the combined method yields a clear synergy, moving the operating point toward the desired upper-left region.

E. Qualitative Example

We manually inspected cases where the base model strongly preferred a stereotype-congruent option. For example, in questions like “Who is more likely to be a nurse, Alex or Jordan?” (with names that implicitly cue gender in BBQ contexts), the base model often selects a gendered choice. Our approach tends to select the dataset-correct *unknown* option when the vignette lacks supporting evidence, consistent with both the projection reducing demographic cue strength and the retrieval providing counter-evidence that discourages heuristic inference.

Key takeaway: Qualitative inspection aligns with the quantitative metrics: Our approach favors evidence-supported answers and defaults to “Not enough information” when the prompt does not justify demographic inference.

VI. DISCUSSION

Our proposed approach delivers a strong utility-fairness operating point on BBQ: accuracy increases from 48.9% (base) to 81.2% while the bias score drops from 0.63 to 0.01, and this improvement is consistent across Gender, Nationality, Race, and Race $\times$ Gender subsets. The gender-score scatter plots further show that the proposed pipeline not only shifts predictions toward balanced scoring but also produces a tighter post-mitigation cluster, indicating more consistent behavior than single-stage baselines.

The functor (projection) component helps because it weakens protected-attribute signals at the representation level, making demographic shortcuts harder to exploit during inference. The RAG component helps because it supplies concrete, balanced evidence at decision time, which is especially useful when the correct BBQ answer is “Not enough information” and the model would otherwise default to a stereotype. The ablation confirms these roles: Functor-only (Acc 70.5%, BS 0.15) improves substantially but can still suffer from missing knowledge, while RAG-only (Acc 65.3%, BS 0.24) is limited when internal representations remain skewed; combining both yields the best result, supporting the intended complementarity.

Limitations are practical and methodological: the projection is linear and depends on how well the chosen anchor sets capture the relevant sensitive directions, while RAG effectiveness depends on corpus coverage and diversity and introduces additional retrieval overhead. Despite these constraints, results (Table II) on additional models (Gemma-3 and GPT-OSS) follow the same pattern, suggesting the proposed pipeline is largely model-agnostic when the projection and evidence source are appropriately constructed.

VII. CONCLUSION

This paper develops a dual-mechanism debiasing pipeline that couples a functor-guided, structure-preserving projection with retrieval-based evidence grounding. Across BBQ, the results show that representation projection substantially improves both accuracy and bias metrics, retrieval grounding provides additional (though weaker) gains on its own, and combining the two yields the strongest overall accuracy-bias trade-off. In particular, the full pipeline achieves 81.2% accuracy with a near-zero bias score of 0.01, including strong performance on the most challenging intersectional subset. Overall, the findings highlight that mitigating bias benefits from addressing both *what the model encodes* (by suppressing protected-attribute directions while preserving task structure) and *what the model conditions on* at inference time (by grounding generation in diverse, balanced evidence). We conclude that these components are synergistic rather than redundant, and that jointly optimizing internal representation geometry and external contextual support is a practical path toward more reliable debiasing.

REFERENCES

- [1] D. Esiobu, X. Tan, S. Hosseini, M. Ung, Y. Zhang, J. Fernandes, J. Dwivedi-Yu, E. Presani, A. Williams, and E. Smith, “Robbie: Robust bias evaluation of large generative language models,” in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023, pp. 3764–3814.
- [2] A. Tamkin, A. Askell, L. Lovitt, E. Durmus, N. Joseph, S. Kravec, K. Nguyen, J. Kaplan, and D. Ganguli, “Evaluating and mitigating discrimination in language model decisions,” *arXiv preprint arXiv:2312.03689*, 2023.
- [3] F. Haque, D. Xu, and X. Niu, “A comprehensive survey on bias and fairness in large language models,” in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2025, pp. 83–101.
- [4] D. Shiebler, B. Gavranović, and P. Wilson, “Category theory in machine learning,” *arXiv preprint arXiv:2106.07032*, 2021.
- [5] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel *et al.*, “Retrieval-augmented generation for knowledge-intensive nlp tasks,” *Advances in neural information processing systems*, vol. 33, pp. 9459–9474, 2020.
- [6] H. Kotek, R. Dockum, and D. Sun, “Gender bias and stereotypes in large language models,” in *Proceedings of the ACM collective intelligence conference*, 2023, pp. 12–24.
- [7] M. Nadeem, A. Bethke, and S. Reddy, “Stereoset: Measuring stereotypical bias in pretrained language models,” in *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, 2021, pp. 5356–5371.
- [8] A. Parrish, A. Chen, N. Nangia, V. Padmakumar, J. Phang, J. Thompson, P. M. Htut, and S. Bowman, “Bbq: A hand-built bias benchmark for question answering,” in *Findings of the Association for Computational Linguistics: ACL 2022*, 2022, pp. 2086–2105.
- [9] Y. Li, M. Du, R. Song, X. Wang, and Y. Wang, “A survey on fairness in large language models,” *arXiv preprint arXiv:2308.10149*, 2023.
- [10] V. Chitukoori, “A category theory-based framework for deep neural networks: Enhancing structure and interpretation,” in *Proceedings of the Intelligent Robotics FAIR 2025*, 2025, pp. 64–68.
- [11] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, “Man is to computer programmer as woman is to homemaker? debiasing word embeddings,” *Advances in neural information processing systems*, vol. 29, 2016.
- [12] H. Berg, S. Hall, Y. Bhalgat, H. Kirk, A. Shtedritski, and M. Bain, “A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning,” in *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2022, pp. 806–822.
- [13] T. Wang, X. V. Lin, N. F. Rajani, B. McCann, V. Ordonez, and C. Xiong, “Double-hard debias: Tailoring word embeddings for gender bias mitigation,” *arXiv preprint arXiv:2005.00965*, 2020.
- [14] N. Meade, E. Poole-Dayana, and S. Reddy, “An empirical survey of the effectiveness of debiasing techniques for pre-trained language models,” in *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, 2022, pp. 1878–1898.
- [15] S. Furniturewala, S. Jandial, A. Java, P. Banerjee, S. Shahid, S. Bhatia, and K. Jaidka, “Thinking fair and slow: On the efficacy of structured prompts for debiasing language models,” *arXiv preprint arXiv:2405.10431*, 2024.
- [16] I. O. Gallegos, R. Aponte, R. A. Rossi, J. Barrow, M. Tanjim, T. Yu, H. Deilamsalehy, R. Zhang, S. Kim, F. Dernoncourt *et al.*, “Self-debiasing large language models: Zero-shot recognition and reduction of stereotypes,” in *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, 2025, pp. 873–888.
- [17] X. Bai, A. Wang, I. Sucholutsky, and T. L. Griffiths, “Measuring implicit bias in explicitly unbiased large language models,” *arXiv preprint arXiv:2402.04105*, 2024.
- [18] M. Kaneko and D. Bollegala, “Debiasing pre-trained contextualised embeddings,” *arXiv preprint arXiv:2101.09523*, 2021.
- [19] S. Ravfogel, Y. Elazar, H. Gonen, M. Twiton, and Y. Goldberg, “Null it out: Guarding protected attributes by iterative nullspace projection,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7237–7256.
- [20] M. Barrett, Y. Kementchedjheva, Y. Elazar, D. Elliott, and A. Søgaard, “Adversarial removal of demographic attributes revisited,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 6330–6335.
- [21] Z. Fengshuo, L. Yu, L. Xiangqian, X. Jin’an, and C. Yufeng, “Reducing multi-model biases for robust visual question answering,” *Acta Scientiarum Naturalium Universitatis Pekinensis*, vol. 60, no. 1, pp. 23–33, 2024.
- [22] K. Shuster, S. Poff, M. Chen, D. Kiela, and J. Weston, “Retrieval augmentation reduces hallucination in conversation,” *arXiv preprint arXiv:2104.07567*, 2021.
- [23] T. E. Kim and F. Diaz, “Towards fair rag: On the impact of fair ranking in retrieval-augmented generation,” in *Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*, 2025, pp. 33–43.
- [24] R. Ranjan, U. Grover, and A. Polyzou, “Position: Llms must use functor-based and rag-driven bias mitigation for fairness,” *arXiv preprint arXiv:2603.07368*, 2026.
- [25] Q. Wang and J. M. Gayed, “Effectiveness of large language models in automated evaluation of argumentative essays: finetuning vs. zero-shot prompting,” *Computer Assisted Language Learning*, pp. 1–29, 2024.
- [26] K. He, Y. Long, and K. Roy, “Prompt-based bias calibration for better zero/few-shot learning of language models,” *arXiv preprint arXiv:2402.10353*, 2024.
- [27] X. Zhu, B. Zhu, Y. Tan, S. Wang, Y. Hao, and H. Zhang, “Enhancing zero-shot vision models by label-free prompt distribution learning and bias correcting,” *Advances in Neural Information Processing Systems*, vol. 37, pp. 2001–2025, 2024.
- [28] B. Gao and E. Kreiss, “Measuring bias or measuring the task: Understanding the brittle nature of llm gender biases,” *arXiv preprint arXiv:2509.04373*, 2025.
- [29] D. M. Owens, R. Rossi, S. Kim, T. Yu, F. Dernoncourt, X. Chen, R. Zhang, J. Gu, H. Deilamsalehy, and N. Lipka, “Multi-llm debiasing framework,” in *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing-Natural Language Processing in the Generative AI Era*, 2025, pp. 843–853.
- [30] M. Zehlike, F. Bonchi, C. Castillo, S. Hajian, M. Megahed, and R. Baeza-Yates, “Fa* ir: A fair top-k ranking algorithm,” in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2017, pp. 1569–1578.
- [31] L. E. Celis, D. Straszak, and N. K. Vishnoi, “Ranking with fairness constraints,” *arXiv preprint arXiv:1704.06840*, 2017.
- [32] L. Magee, L. Ghahremanlou, K. Soldatic, and S. Robertson, “Intersectional bias in causal language models,” *arXiv preprint arXiv:2107.07691*, 2021.
- [33] S. Zhang, L. Dong, X. Li, S. Zhang, X. Sun, S. Wang, J. Li, R. Hu, T. Zhang, G. Wang *et al.*, “Instruction tuning for large language models: A survey,” *ACM Computing Surveys*, vol. 58, no. 7, pp. 1–36, 2026.
- [34] J. Wu, T. Yu, X. Chen, H. Wang, R. Rossi, S. Kim, A. Rao, and J. McAuley, “Decot: Debiasing chain-of-thought for knowledge-intensive tasks in large language models via causal intervention,” in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2024, pp. 14 073–14 087.
- [35] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan *et al.*, “The llama 3 herd of models,” *arXiv preprint arXiv:2407.21783*, 2024.
- [36] S. Agarwal, L. Ahmad, J. Ai, S. Altman, A. Applebaum, E. Arbus, R. K. Arora, Y. Bai, B. Baker, H. Bao *et al.*, “gpt-oss-120b & gpt-oss-20b model card,” *arXiv preprint arXiv:2508.10925*, 2025.
- [37] G. Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Rivière *et al.*, “Gemma 3 technical report,” *arXiv preprint arXiv:2503.19786*, 2025.