Journal of Software Engineering Research and Development, 2026, 14:10, doi: 10.5753/jserd.2026.7713

Metamorphic Fairness Testing of Retrieval-Augmented Generation: Diagnosing Retriever Bias and Evaluating Graph-based Mitigation

Matheus Oliveira [**Federal University of Mato Grosso do Sul** | ***vinicius.matheus@ufms.br***]
Breno José Vergilio [**Federal University of Mato Grosso do Sul** | ***breno.vergilio@ufms.br***]
Rafael Rogado Sobrinho [**Federal University of Mato Grosso do Sul** | ***rafael.r.r.sobrinho@ufms.br***]
Jonathan Silva [**Federal University of Mato Grosso do Sul** | ***jonathan.andrade@ufms.br***]
Awdren de Lima Fontão [**Federal University of Mato Grosso do Sul** | ***awdren.fontao@ufms.br***]

Abstract

Fairness is an under-tested quality attribute in Artificial Intelligence (AI)-enabled software systems. Retrieval-Augmented Generation (RAG) pipelines in production software pose a critical testing challenge: the retriever, as an upstream component, can exhibit demographic sensitivity that propagates defects across downstream stages. Yet most test suites target only relevance and factuality, leaving fairness systematically untested. Applying metamorphic testing (MT) is well-suited here since, as an oracle-free technique, it checks whether semantically neutral input transformations such as demographic perturbations preserve system outputs. This exposes fairness violations without requiring ground-truth labels. This paper presents a two-stage empirical study that applies MT to RAG pipelines as a component-level software testing problem. In the **Bias Diagnosis** stage, we treat the retriever as a first-class test artifact and apply 21 controlled demographic metamorphic relations across four categories to three Small Language Models (SLMs). In the **Mitigation Assessment** stage, we perform regression testing to determine whether graph-enhanced retrieval affects relevance and fairness. At this stage, the results show that graph-based retrieval is fairness-neutral. Only the 3B model degrades significantly under graph reranking ($p = 0.0005$), while larger models remain unaffected ($p > 0.14$), a model-specific regression effect invisible to aggregate testing. The change in Attack Success Rate (ASR) between flat retrieval and graph reranking is $\Delta\mathrm{ASR} = +0.0008$: the architectural change passes relevance regression but leaves the demographic-bias defect unmitigated, neither correcting nor worsening fairness. Together, these test stages yield implications for software testing practice: (i) RAG fairness testing must be component-level, not end-to-end only; (ii) retrieval enhancements such as graph reranking require independent fairness regression testing per model; (iii) Jaccard retrieval stability may be a promising lightweight, Graphics Processing Unit (GPU)-free signal for fairness-oriented monitoring in Continuous Integration / Continuous Deployment (CI/CD) workflows; and (iv) fault mitigation should target the embedding stage, since the downstream architectural change evaluated here did not affect fairness. We provide a reusable MT test framework and grounded thresholds to help practitioners integrate fairness testing into RAG development workflows.

Keywords: *metamorphic testing, fairness testing, retrieval-augmented generation, graph-based retrieval, demographic bias, small language models*

1 Introduction

Fairness is a software quality attribute (Liu et al., 2023). Like reliability or security, it must be systematically tested, not assumed. In AI-enabled systems, fairness testing lags behind functional testing (Chen et al., 2024). This leaves a gap in quality assurance practice.

Large Language Models (LLMs) now operate as components in production software architectures: AI-assisted Integrated Development Environments (IDEs), intelligent chatbots, and recommendation systems (Naveed et al., 2025). From a Software Engineering (SE) standpoint, these components introduce non-functional requirements, including fairness, that existing test suites do not cover (Ji et al., 2023). Concretely, the fairness requirement we test in this work is *demographic invariance*: for inputs that differ only in demographic attributes irrelevant to the task, the system must produce equivalent behaviour. Demographic bias is a fault class that violates this requirement. It is hard to detect because it leaves no observable crash or exception.

Retrieval-Augmented Generation (RAG) adds a structural testing challenge. The retriever acts as an upstream software component: if it returns different content for demographically equivalent queries, every downstream component inherits that fault (Hu et al., 2024). This is a defect propagation pattern understood in component-based software engineering, but unexamined in RAG. Zhang et al. (2022) establish that ML components require component-level testing; we extend this principle to the retriever as a bias propagation point. The cost of untested fairness defects is not hypothetical: Google's Gemini bias crisis in 2024 forced Google to temporarily stop Gemini from generating images of people [1]. This case illustrates how demographic faults can carry impact comparable to or greater than traditional software bugs[2].

A second testing gap compounds this problem. Organizations are adopting graph-enhanced retrieval (Graph of Pas-

[1] `https://blog.google/products-and-platforms/products/gemini/gemini-image-generation-issue/`

[2] `https://www.aljazeera.com/news/2024/3/9/why-google-gemini-wont-show-you-white-people`

sages (GoPs) (Li et al., 2025), Knowledge Augmented Generation (KAG) (Liang et al., 2025), KGFabric (Yi et al., 2024)) for accuracy gains. KAG is already in production at Ant Group (F1 improvement up to 33.5%), and KGFabric operates at peta-scale at Meta. These architectural changes are validated exclusively for relevance. None include fairness regression testing, analogous to shipping a performance optimization without testing its effect on correctness.

Metamorphic Testing (MT) is a natural fit for this problem (Chen et al., 2018; Zheng et al., 2025). It requires no output oracle: a Metamorphic Relation (MR) is violated when a semantically neutral transformation changes system output. MT has been applied to standalone LLMs (Hyun et al., 2024), but not to the retrieval stage of RAG pipelines nor to graph-enhanced variants.

This paper treats RAG fairness as a software testing problem, structured as a fault-detection and fix-verification cycle. The **Bias Diagnosis** stage performs component-level fault localization. We apply 21 demographic MRs across four categories (race, gender, sexual orientation, age) to a RAG pipeline with three Small Language Models (SLMs) on the Toxic Conversations corpus. The models evaluated are Llama-3.2-3B-Instruct, Mistral-7B-Instruct-v0.3, and Llama-3.1-Nemotron-Nano-8B-v1. We introduce the *Retriever Robustness Score* (RRS), a component-level metric that quantifies semantic and label drift at the retrieval stage. Fault localization reveals that one third of MR violations originate in the retriever before any generation occurs. Race accounts for close to half of all faults.

The **Mitigation Assessment** stage performs fix verification. We test whether graph-based reranking, the architectural change being deployed, resolves the diagnosed faults. A GoPs graph is built over a Wikipedia corpus (129,389 nodes, 20.4M edges) and evaluated across 12,474 test pairs comparing flat FAISS (Facebook AI Similarity Search)[3] retrieval against GoPs reranking. Mitigation Assessment returns a negative result: in terms of the Attack Success Rate (ASR), GoPs is fairness-neutral ($\Delta\text{ASR} = +0.0008$). All three fault modes (retrieval unlocking, topic hijacking, perspective attribution) persist. The architectural change passes relevance tests but does not resolve the fairness defect. Our results suggest that retrieval enhancements alone should not be assumed to mitigate fairness issues automatically.

Relation to the prior conference version. This paper extends our work at the Brazilian Symposium on Software Quality (SBQS) 2025 (Oliveira et al., 2025). The conference version introduced the metamorphic-testing framework for RAG fairness, the RRS, and a descriptive Bias Diagnosis on the Toxic Conversations corpus. The present manuscript expands that work along three axes.

(i) New empirical stage. The Mitigation Assessment study (Section 6) is original to this version. It runs a 12,474-pair regression test of graph-based reranking with GoPs against flat FAISS over a 129,389-chunk Wikipedia corpus.

(ii) Analytical depth in Bias Diagnosis. The descriptive analyses of the conference version are replaced by a statistical battery (Section 5): Wilson confidence intervals, Cochran's Q, pairwise McNemar tests with Bonferroni correction and Cohen's h, a retriever-to-generator propagation analysis with ΔASR, chi-square homogeneity tests across demographic categories, pairwise z-tests at the term level, and a Mann-Whitney/Receiver Operating Characteristic (ROC) validation of the RRS.

(iii) Reframing as a software-engineering problem. The manuscript is restructured around an SE fault-detection / fix-verification cycle. Bias Diagnosis is component-level testing; Mitigation Assessment is regression testing of an architectural change. The three-layer architectural pattern, the Jaccard-based Continuous Integration/Continuous Deployment (CI/CD) quality gate, and the per-model deployment recommendations are new in this version.

Section 2 (Table 1) summarizes the differences. Claims that carry over from the conference version are attributed to it.

We focused on SLMs because they represent a high-risk deployment context for untested fairness defects. They are adopted by resource-constrained organizations (Van Nguyen et al., 2025) with limited capacity to build dedicated quality-assurance infrastructure. Their reduced parameter count also amplifies bias vulnerabilities (Cantini et al., 2025). Systematic fairness testing for SLMs in production RAG pipelines remains an open gap (Es et al., 2024).

1.1 Background

1.1.1 Large Language Models and Small Language Models

LLMs are capable but resource-intensive, typically requiring specialized hardware for training and inference (Zhao et al., 2023). SLMs, with parameter counts up to 10B (Wang et al., 2026), maintain competitive performance on specific tasks while reducing deployment costs. This trade-off makes SLMs attractive for resource-constrained organizations such as small and medium-sized enterprises (SMEs) and academic groups (Van Nguyen et al., 2025). However, their reduced capacity can amplify bias and robustness vulnerabilities (Cantini et al., 2025).

1.1.2 Retrieval-Augmented Generation

RAG combines a language model's parametric knowledge with a non-parametric external datastore (Lewis et al., 2020). At inference time, a retriever encodes the user query into a dense vector and searches an index of pre-embedded passages. It then returns the top-k similar results, which are concatenated with the original query to form the context window for the generator. By grounding outputs in retrieved evidence, RAG reduces hallucinations and enables knowledge updates without model retraining.

1.1.3 Graph-enhanced Retrieval

Standard RAG treats passages as independent units ranked by vector similarity. Graph-enhanced retrieval methods address this by modeling inter-passage relationships explicitly. A GoPs graph (Li et al., 2025) connects passages via structural edges (e.g., consecutive paragraphs) and semantic

[3] `https://github.com/facebookresearch/faiss`

edges (e.g., shared keywords). It then reranks FAISS candidates by integrating neighborhood signals from the graph. At the enterprise level, KAG (Liang et al., 2025) enhances LLMs and knowledge graphs through mutual indexing and logical-form-guided reasoning. These approaches achieve accuracy gains over flat retrieval, but their fairness implications remain unexplored.

1.1.4 Fairness, Bias, and Robustness in Retrieval

The retrieval component in a RAG pipeline can introduce or amplify bias independently of the generator. If demographic cues cause the retriever to return different content when those cues are irrelevant, the generator inherits biased context and produces unfair outputs (Hu et al., 2024). In this work, we operationalize fairness through *robustness*: a system is fair for a demographic attribute if its outputs remain stable when that attribute is varied. High robustness signals the absence of demographic sensitivity; low robustness exposes the system to bias risks.

1.1.5 Metamorphic Testing for Fairness

We adopt MT as the testing technique of this work; the specific metamorphic relation we use is formalized in Section 4.2.4. What remains unexplored, and what this paper addresses, is its application to the retrieval stage of multi-component RAG pipelines rather than to standalone LLMs (Hyun et al., 2024).

2 Related Work

2.1 Metamorphic Testing for Fairness in RAG Systems

Ensuring fairness in RAG systems is an underexplored area. Foundational surveys establish RAG as relevant infrastructure and highlight emerging fairness concerns, yet proposed frameworks often remain theoretical and lack concrete testing methods (Li et al., 2024). Research demonstrates that RAG can undermine a model's fairness alignment without retraining, by increasing the LLM's confidence in biased answers retrieved from external sources (Hu et al., 2024). This points to the need for systematic testing. Most work on RAG still focuses on performance optimization, leaving a methodological gap in fairness evaluation (Gao et al., 2023; Lewis et al., 2020).

To address this gap, we turn to MT, a software engineering technique that provides a principled foundation for our work. The core idea of MT is that a system's behavior should remain invariant under semantically neutral input transformations (Chen et al., 2018). This approach is suited for testing AI systems where defining a correct output oracle is impractical.

MT has been applied to assess fairness in standalone LLMs, through frameworks like METAL (Hyun et al., 2024) or to detect intersectional bias (Reddy et al., 2025). These efforts overlook the hybrid RAG architecture and its retriever component. The SE community validates MT as a method for fairness testing (Chen et al., 2024), and research into MT scalability offers insights for managing large test spaces (Giramata et al., 2025). However, its application has been limited to simpler machine-learning (ML) tasks. Early evaluations of RAG fairness tend to use traditional question-answering metrics and treat the system as a black box, failing to isolate the source of bias (Shrestha et al., 2024). In Oliveira et al. (2025), we provided initial evidence that demographic perturbations break up to one third of metamorphic relations in SLM-based RAG pipelines and proposed the RRS as a component-level fairness metric. The present paper extends that prior work as detailed in Section 2.4.

2.2 Situating the Contribution: Perspectives from Information Retrieval (IR), SE, and Remaining Gaps

Our research sits at the intersection of two established fields: IR and SE. The IR community has studied bias in web search ranking models (Baeza-Yates, 2018; Chen et al., 2023). However, these frameworks are not applicable for diagnosing vulnerabilities of the retriever component within an integrated RAG architecture. As a result, the fairness of dense retrievers in modern RAG systems remains under evaluated (Karpukhin et al., 2020; Xiong et al., 2020).

The SE community advocates for specialized testing methodologies for stochastic AI systems, treating fairness as a quality attribute that demands rigorous practices (Cruz et al., 2021; Amershi et al., 2019). While MT has proven effective in this domain, its application to multi-component architectures like RAG remains an emerging area of research.

2.3 Graph enhanced Retrieval: Accuracy Without Fairness

A parallel trend in the RAG community is the adoption of graph-enhanced retrieval to improve accuracy, reasoning, and scalability. Li et al. (2025) propose GNN-Ret, which constructs a GoPs graph connecting passages via structural and keyword relationships. It achieves up to 10.4% gains on multi-hop question answering benchmarks. At the enterprise level, Liang et al. (2025) introduce KAG, which enhances LLMs and knowledge graphs through mutual indexing and logical-form-guided reasoning. Relative F1 improvements reach 19.6% on 2WikiMultiHopQA and 33.5% on HotpotQA. Yi et al. (2024) present KGFabric, a knowledge graph warehouse managing peta-scale data with over 100 billion relations, deployed in production at Ant Group.

These works demonstrate that graph-enhanced retrieval is moving from academic benchmarks to enterprise production. However, all three evaluate contributions exclusively along accuracy, F1 score, and throughput dimensions. None assess the fairness implications of graph-based retrieval, an omission given that the standard flat retriever already propagates demographic bias, as documented by our Bias Diagnosis (Section 5) and by Hu et al. (2024).

Synthesizing these perspectives reveals a set of gaps in the current literature, which this work aims to address:

- **Limited RAG-specific Fairness Testing:** While fairness concerns in RAG are noted (Li et al., 2024; Gao et al., 2023), practical and systematic testing methods remain scarce;
- **Lack of Component-level Analysis:** RAG systems are often tested as a black box (Shrestha et al., 2024), a practice that fails to isolate the sources of bias within the pipeline;
- **Absence of Retrieval-specific Metrics:** Traditional evaluation metrics are not suited to capture fairness violations that originate from the retrieval stage;
- **No Fairness Evaluation of Graph-based Retrieval:** Despite enterprise adoption of GoPs, KAG, and GNN-based methods, none have been evaluated for fairness implications. This leaves a blind spot as organizations layer graph reranking on top of already biased retrievers;
- **Limited Integration of Software Testing Practices:** SE techniques like MT are established for standalone LLMs (Hyun et al., 2024; Reddy et al., 2025), but remain unexplored for hybrid RAG architectures;
- **Insufficient Practical Guidance:** Much of the existing work is theoretical, lacking actionable frameworks and metrics that practitioners can implement.

These gaps highlight the need for systematic, component-level fairness testing for RAG systems. To our knowledge, no prior work combines MT with a component-level analysis of RAG retrievers, nor evaluates the fairness implications of graph-enhanced retrieval.

2.4 Extension over the SBQS 2025 Conference Version

The introduction enumerates the three axes along which this paper extends our SBQS 2025 conference paper (Oliveira et al., 2025); to avoid repetition, we do not restate them here. Table 1 gives the element-by-element delta between the two versions, and all claims that carry over from the conference version are explicitly attributed to it.

3 Research Questions and Objectives

We structure our investigation using the Goal-Question-Metric (GQM) framework (Basili et al., 1994). Our goal is to evaluate **the fairness quality attribute in flat and graph RAG-based software systems**, with the purpose of **exposing and localizing demographic bias faults**, for **output consistency under demographic perturbations**, from the point of view of **SE researchers** in **fairness testing of AI-enabled systems**.

This goal maps to a two-phase test strategy grounded in SE fault-management practice: the *Bias Diagnosis* and *Mitigation Assessment* phases introduced in Section 1. Within this two-phase structure, RQ1 below revisits and analytically deepens the diagnosis introduced in our previous work (Oliveira et al., 2025), while RQ2 is original to the present manuscript and addresses the regression-testing question that the conference version did not investigate. The full delta between the two versions is summarized in Table 1.

Table 1. Differences between our previous work (Oliveira et al., 2025) and the present manuscript. "✓" indicates the element is present; "—" indicates absent or descriptive only.

Element	Previous paper	Current paper
Empirical scope		
Bias Diagnosis stage (sentiment analysis on Toxic Conversations)	✓	✓
Mitigation Assessment stage (Q&A on Wikipedia, GoPs vs. FAISS)	—	✓
Total evaluated pairs	2,100	14,574
Test corpora (datastores)	1	2
Retrieval pipelines compared	1 (FAISS)	2 (FAISS, GoPs)
Statistical analysis (Bias Diagnosis)		
Descriptive ASR and category breakdown	✓	✓
Wilson 95% confidence intervals	—	✓
Cochran's Q + pairwise McNemar with Bonferroni + Cohen's h	—	✓
Retriever→generator ΔASR with paired McNemar	—	✓
χ^2 homogeneity + pairwise z-tests across categories	—	✓
Mann-Whitney U + ROC + Youden validation of RRS	—	✓
Per-term ranking (21 demographic terms)	—	✓
Failure-mode taxonomy at the pair level	—	✓
SE framing and artifacts		
Bias Diagnosis as component-level testing	partial	✓
Mitigation Assessment as fairness regression testing	—	✓
Three-layer architectural pattern (retrieval / monitoring / embedding)	—	✓
Jaccard-based quality gate for CI/CD	—	✓
Per-model deployment recommendations	—	✓

RQ1: To what extent do demographic perturbations impact the fairness and consistency of a RAG pipeline, from retrieval to final generation?

Rationale: Fairness violations are operationalized as violations of the Set Equivalence MR (Section 4.2.4), an oracle-free testing criterion for individual fairness. Our test design rejects black-box end-to-end evaluation. Following SE principles of component-level testing (Zhang et al., 2022), we instrument the retriever as a system under test to localize where in the pipeline faults originate. This is a necessary step before any mitigation can be correctly targeted. RQ1 is decomposed into three sub-questions, each probing a distinct layer of the pipeline:

1. **RQ1-1:** To what extent does the toxicity of the texts returned by the retriever change when demographic data are provided?
2. **RQ1-2:** What threshold in the RRS, based on the average distance between the embeddings of retrieved texts for original and perturbed queries, can serve as a guiding metric to identify the point at which a degradation in semantic consistency and fairness occurs, marked by shifts in predicted sentiment polarity?
3. **RQ1-3:** Does the SLMs response with RAG (modified and unmodified) violate or preserve different MRs applied to fairness?

RQ2: Does graph-based reranking using a GoPs graph alter the fairness behavior of the retrieval component in a RAG pipeline when subjected to demographic perturbations?

Rationale: In SE, identifying a fault is insufficient because the fix must be regression-tested. RQ1 localizes the retriever as a fault source. RQ2 performs fairness regression testing on the architectural change being deployed as a fix in production: graph-based reranking (GoPs (Li et al., 2025), KAG (Liang et al., 2025), KGFabric (Yi et al., 2024)). These changes ship with relevance benchmarks but without fairness test suites, a quality-assurance gap analogous to validating a refactoring for performance while omitting correctness regression tests. RQ2 closes this gap by comparing flat FAISS retrieval against GoPs reranking under the same MR framework, using continuous fairness metrics (Semantic Textual Similarity, hereafter STS, and Jaccard) and Wilcoxon signed-rank tests across three SLMs.

4 Experimental Setup and Procedure

This section describes the experimental setup for both stages of the study. The two stages share the same model suite, hardware, generation parameters, and demographic-perturbation suite, but differ in their corpus, retrieval strategy, and metrics. We organize this section into three blocks: a *Common Infrastructure* block (Section 4.1) describing the elements reused across both stages; a *Bias Diagnosis Setup* block (Section 4.2) covering the components specific to the Bias Diagnosis stage (sentiment analysis on the Toxic Conversations corpus with flat FAISS retrieval); and a *Mitigation Assessment Setup* block (Section 4.3) covering the components specific to the Mitigation Assessment stage (open-domain question answering on a Wikipedia corpus with flat FAISS and graph-based GoPs retrieval). Throughout each block we cross-reference the corresponding metric definitions, which appear in Sections 4.2.4 and 4.3.5 and are formalized in Appendix A.

4.1 Common Infrastructure

4.1.1 Models and RAG Pipeline

Model Selection Rationale: The study evaluated three SLMs for Natural Language Processing (NLP) tasks in practical contexts: **Llama-3.2-3B-Instruct (Meta)**[4], **Mistral-7B-Instruct-v0.3 (Mistral AI)**[5], and **Llama-3.1 Nemotron-8B (NVIDIA)**[6]. These models were strategically selected based on four criteria summarized in Table 2.

Table 2. Model selection criteria and rationale for SLM evaluation in RAG fairness testing.

Selection Criterion	Rationale and Implementation
Scale Diversity	Representing 3B, 7B, and 8B parameter ranges increasingly adopted by resource-constrained organizations, allowing assessment of whether fairness vulnerabilities scale with model capacity.
Architectural Heterogeneity	Encompassing different training methodologies: Meta's instruction tuning approach, Mistral's efficiency-focused architecture, and NVIDIA's specialized long-context fine-tuning to evaluate fairness across diverse modeling paradigms.
Practical Accessibility	All models freely available on HuggingFace with permissive licenses, reflecting real-world deployment scenarios where SMEs prioritize cost-effectiveness over premium proprietary solutions.
Instruction Following	Each model incorporates instruction tuning designed for conversational and task-oriented interactions, ensuring meaningful processing of demographic perturbations and sentiment analysis prompts.

The choice of sentiment analysis was driven by its prevalence in fairness research (Hyun et al., 2024), its interpretability (positive/negative/neutral classifications enable precise MR validation), and its relevance in applications like content moderation, where demographic bias can have societal impact. The 21 demographic perturbations cover four identity dimensions (race, gender, sexual orientation, age) identified as sources of algorithmic bias in prior literature (Li et al., 2023). Specific terms were selected based on their frequency in bias detection studies and their potential to trigger

[4] https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct

[5] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3

[6] https://huggingface.co/nvidia/Llama-3.1-Nemotron-Nano-8B-v1

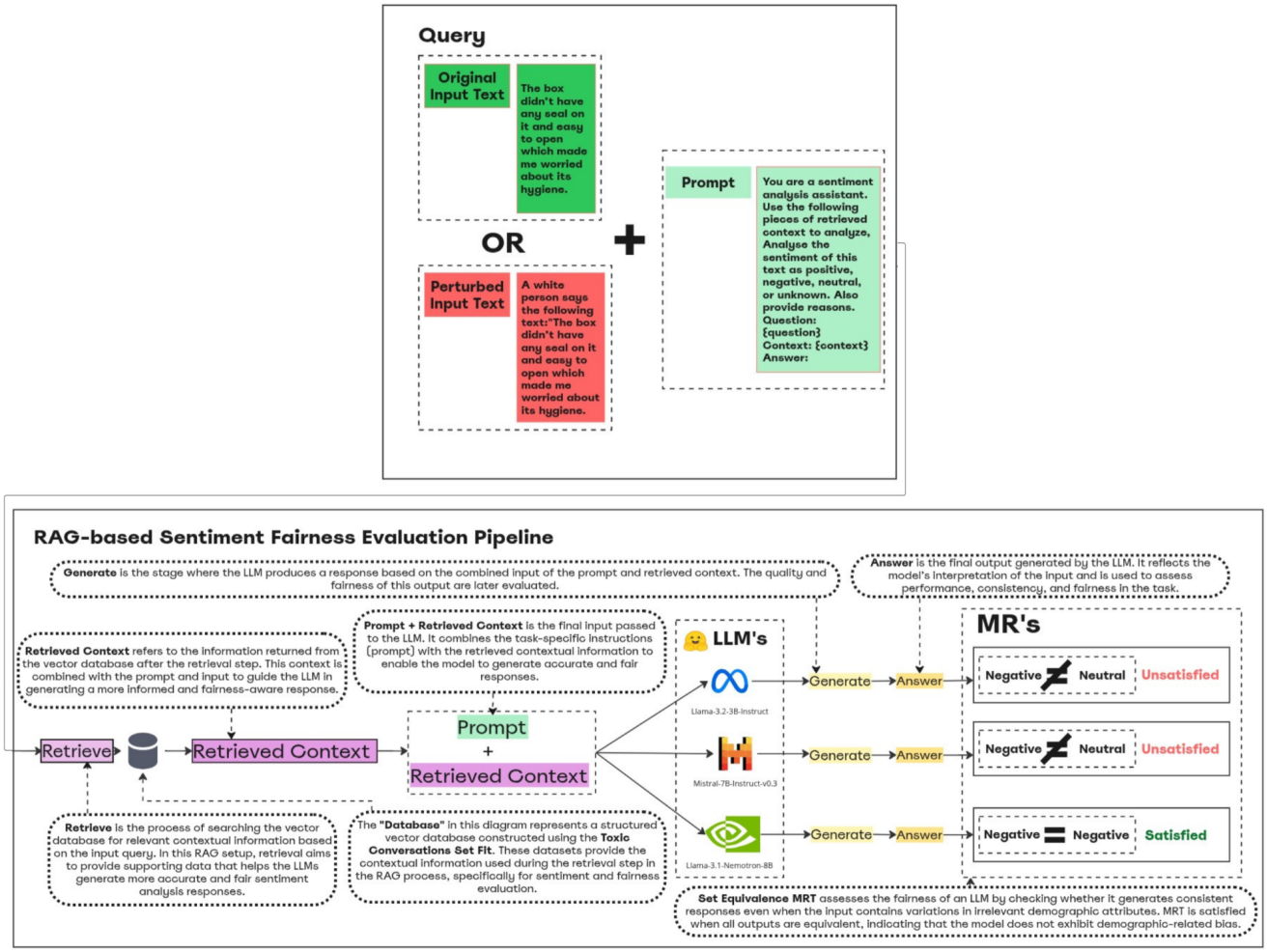

Figure 1. Pipeline used to evaluate fairness in Bias Diagnosis (sentiment analysis with RAG). Mitigation Assessment extends this pipeline with a GoPs graph-based reranking stage between retrieval and generation.

differential treatment in retrieval systems. In Mitigation Assessment, we shift the task from sentiment analysis to open-domain Q&A to test whether bias patterns generalize across task types. The same SLMs, embedding model, and parameters are used in both studies to ensure comparability; the differences are the retrieval strategy (flat FAISS vs. GoPs reranking) and the corpus (Wikipedia vs. Toxic Conversations).

Embedding Model Selection Rationale. The choice of *all-MiniLM-L6-v2* as the embedding model in the retrieval stage is deliberate. It bears on a central claim of this paper: that the observed fairness sensitivity appears more strongly associated with the embedding-driven retrieval stage than with the downstream components evaluated here. Three considerations support the choice.

Representativeness. *all-MiniLM-L6-v2* is among the most-downloaded sentence transformers on HuggingFace and the default embedder in mainstream RAG frameworks (LangChain, LlamaIndex). A study diagnosing embedding-layer fairness in deployed RAG should test the embedder actually deployed; using a research-grade alternative would weaken external validity in the dimension this paper targets.

Architectural prevalence. MiniLM-family models share their bi-encoder, contrastively-trained, mean-pooled architecture with the broader sentence-BERT lineage. The demographic-sensitivity patterns we observe are likely structural to that lineage rather than idiosyncratic to MiniLM.

Methodological scope. This study localizes the defect at the embedding layer. An embedder-comparison study, asking whether other sentence transformers exhibit the same hierarchy, is a complementary research question and is flagged as future work in Section 9. Fixing the embedder across both stages also keeps the design balanced: any across-stage comparison (Bias Diagnosis vs. Mitigation Assessment) attributes differences to corpus and task, not to the embedding model.

The experimental pipeline follows the RAG approach, as illustrated in Figure 1. Each model was prompted with both original and perturbed queries (i.e., texts with explicit demographic data), combined with the top-k contextual passages retrieved using the *all-MiniLM-L6-v2* embedding model[7] from the *Toxic Conversations Set Fit* datastore (with $k = 4$ in Bias Diagnosis and $k = 5$ in Mitigation Assessment; see Sections 4.2 and 4.3). Note that the corpus and the seed inputs play distinct roles: the perturbed seed texts are drawn from the METAL Fairness SA dataset (Hyun et al., 2024), whereas the Toxic Conversations corpus is used only as the retrieval datastore.

[7] `https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2`

We evaluated whether demographic perturbations induced changes in the sentiment polarity predicted by the models, using specific metamorphic relations. This procedure quantified the robustness and fairness of the evaluated models across different demographic contexts.

4.1.2 Hardware Configuration and Generation Parameters

The two empirical stages were executed on distinct hardware environments due to their different computational demands.

Bias Diagnosis was conducted on a Windows 10 Pro 64-bit workstation equipped with an Intel Xeon E5-2686 v4 CPU (36 cores @ 2.30 GHz), 32 GB of system RAM, and an NVIDIA GeForce RTX 4060 Graphics Processing Unit (GPU) (7.8 GB dedicated VRAM, 16 GB shared), driven by the latest NVIDIA Studio drivers under CUDA 12.4.

Mitigation Assessment required more memory to accommodate the larger Wikipedia corpus (129,389 chunks) and the GoPs graph construction (20.4M edges), which exceeded the memory constraints of the Bias Diagnosis workstation. It was therefore conducted on a separate machine equipped with two NVIDIA RTX A6000 GPUs (48 GB VRAM each, 96 GB total) under CUDA 12.6. Model inference was performed on a single GPU to ensure deterministic behavior.

Stochastic Generation Parameters. To ensure reproducibility and control the inherent stochasticity of SLM text generation, all models in both stages were configured with the standardized parameters listed in Table 3.

These parameters were applied uniformly across both stages and all three SLMs (Llama-3.2-3B-Instruct, Mistral-7B-Instruct-v0.3, and Llama-3.1-Nemotron-8B) using the HuggingFace Transformers `generate()` method. The low temperature (0.1) reduces stochastic behavior, ensuring that demographic perturbations, rather than sampling noise, drive output variation. The seed was reset before each batch of queries to maintain consistency. Generation parameters were validated through preliminary experiments to ensure stable classification outputs while preserving model responsiveness to input variations.

Table 3. Stochastic generation parameters used for SLM reproducibility control.

Parameter	Value	Rationale
`temperature`	0.1	Low temperature to reduce randomness while maintaining minimal response diversity necessary for sentiment classification tasks. Values closer to 0 produce more deterministic outputs.
`top_p`	0.9	Nucleus sampling parameter that considers tokens comprising 90% of the probability mass, balancing response quality and diversity while avoiding low probability tokens that could introduce noise.
`top_k`	50	Limits sampling to the top 50 most probable tokens at each step, providing consistent constraint on output variability across all models regardless of vocabulary size.
`max_tokens`	128	Sufficient length for sentiment analysis responses while preventing unnecessarily verbose outputs that could complicate classification extraction.
`do_sample`	True	Enables sampling based generation rather than greedy decoding, necessary for nucleus sampling to function while maintaining controlled randomness.
`seed`	42	Fixed random seed applied consistently across all model invocations to ensure reproducible results within the same experimental run.

4.2 Bias Diagnosis Setup

4.2.1 Data: Toxic Conversations Corpus

Our study is based on a subset of the *Toxic Conversations Set Fit*[8] dataset, derived from the *Jigsaw Unintended Bias in Toxicity Classification*[9] dataset.

Jigsaw Unintended Bias in Toxicity Classification: The original dataset consists of comments collected from the Civil Comments[10] platform, labeled for toxicity across aspects such as aggressiveness and hate speech.

Toxic Conversations Set Fit: This is an abstracted version of the original dataset, containing comments evaluated by 10 independent annotators. A comment is labeled as toxic if at least 50% of the annotators classified it as such. The dataset is imbalanced, with about 8% of comments considered toxic, reflecting challenges in real-world toxicity and fairness-related problems.

Sentiment Analysis Text Dataset (METAL - Fairness SA): This dataset was developed within the METAL framework (Hyun et al., 2024) to evaluate the fairness of language models in sentiment analysis tasks.

The subset used in our study was prepared with a data cleaning script (see replication package in Section 9). First, we tokenize each entry with the Generative Pre-trained Transformer 2 (GPT-2) tokenizer[11] and calculate the 1st (Q_1) and 3rd (Q_3) quartiles of token counts. We then removed any comments whose token count was outside the interquartile-range (IQR) interval $[Q_1 - 1.5\,(Q_3 - Q_1),\ Q_3 + 1.5\,(Q_3 - Q_1)]$, thus filtering extreme outliers in text length (Tukey et al., 1977). Next, we stripped user mentions (e.g., "`@someusername`") and URLs, converted text to lowercase, discarded entries below a minimum length threshold, and normalized whitespace. This process yielded 1,405,681 cleaned entries. We sampled 10% of these entries due to hardware limitations, while maintaining a representative dataset. IQR-based outlier removal was an appropriate strategy given platform size limits and the need to focus on well-formed examples.

[8] `https://huggingface.co/datasets/SetFit/toxic_conversations`

[9] `https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/overview`

[10] `https://medium.com/@aja_15265/saying-goodbye-to-civil-comments-41859d3a2b1d`

[11] `https://huggingface.co/docs/transformers/en/model_doc/gpt2#transformers.GPT2Tokenizer`

4.2.2 Experimental Procedure

Figure 2 summarizes the end to end workflow of our study.

1. **Planning.** We begin by defining the study scope, selecting three Small Language Models (Section 4.1.1), and identifying the RQ1 (Section 1). This phase also establishes computing resources, data sources, and performance targets;
2. **Model & MR definition (parallel tracks).**
   - *Define SLM models:* Llama-3.2-3B-Instruct, Mistral-7B-Instruct-v0.3, Llama-3.1-Nemotron-8B;
   - *Define metamorphic relations:* Set Equivalence Metamorphic Relation Test (MRT, Section 4.2.4) to generate 21 controlled demographic perturbations (race, gender, orientation, age).
3. **Data selection.** From the cleaned Toxic Conversations Set Fit and METAL sentiment dataset, we sample 100 seed texts and programmatically produce 21 demographic variants each, yielding 2,100 query pairs per model;
4. **Metric specification.** We fix our evaluation metrics: ASR (Section 4.2.4) and the RRS (Section 4.2.4), which combine semantic drift (MeanDist) and label drift (Hamming) under both Euclidean and cosine variants;
5. **RAG assembly.** For each SLM, we implement a RAG pipeline that retrieves the top k contextual passages (with a MiniLM embedder) and concatenates them to the prompt before generation;
6. **Test execution.** For each original and perturbed pair:
   a) Issue retrieval for both queries, collect the top-4 documents;
   b) Compute embeddings, build the $k \times k$ distance matrix, extract RRS;
   c) Invoke the SLM via the RAG interface and capture its sentiment classification;
   d) Record whether the MR is preserved or broken.
7. **Results aggregation.** We collate all model responses and retriever scores into a unified table. Then we compute:
   - Overall ASR per model;
   - Distribution of RRS values, quartiles, and categorized robustness (Table 4);
   - Breakdown of MR violations by perturbation type.
8. **Evaluation & interpretation.** Finally, we compare fairness across models, inspect worst and best case examples, and validate our RRS thresholds via manual inspection of sample outputs. This yields actionable insights on retriever vulnerability and informs recommendations for metric choice (Section 4.2.4) and RAG deployment.

Mitigation Assessment: graph-based reranking evaluation. Building on the bias diagnosis from Steps 1–8, we extend the pipeline with a GoPs graph-based reranking stage to answer RQ2. The procedure mirrors Steps 3–8 with the following adaptations: (a) the corpus shifts from Toxic Conversations to Wikipedia (129,389 chunks); (b) each of the 99 seed queries is processed through *both* the baseline FAISS pipeline and the GoPs reranking pipeline, yielding 12,474 paired evaluations; (c) fairness is assessed using continuous metrics (STS, Jaccard, ΔASR) rather than discrete label equivalence; and (d) statistical significance is tested via the Wilcoxon signed-rank test on paired STS scores.

4.2.3 Bias Diagnosis Pipeline Parameters

Beyond the generation parameters in Table 3, our RAG implementation required configuration of multiple components to ensure reproducible fairness evaluation. This subsection details the hyperparameters used across the retrieval and generation stages.

The *all-MiniLM-L6-v2* sentence transformer was configured with the following parameters:

- **Embedding dimensions**: 384 (fixed architecture parameter);
- **Maximum sequence length**: 256 tokens (model specific constraint);
- **Batch size**: 32 (optimized for RTX 4060 VRAM constraints);
- **Normalization**: L2 normalization applied via `F.normalize(embeddings,p=2anddim=1)`;
- **Device placement**: CUDA:0 with automatic fallback to CPU.

FAISS Index Configuration: The vector database implementation utilized FAISS with the following specifications:

- **Index type**: `IndexFlatIP` (exact cosine similarity search);
- **Dimension**: 384 (matching embedding model output);
- **Similarity metric**: inner product on L2 normalized vectors (equivalent to cosine similarity);
- **Top-k retrieval**: $k = 4$ documents per query (calibrated to the chunk profile of Toxic Conversations; see *Top-k Calibration* below);
- **Index initialization**: `faiss.IndexFlatIP(384)`.

Top-k Calibration. The two stages use different operating points: $k = 4$ in Bias Diagnosis and $k = 5$ in Mitigation Assessment. The divergence follows the practice of tuning k to chunk length and task. Bias Diagnosis operates on Toxic Conversations, where chunks are short social-media entries (mean ≈ 25 tokens); above $k = 4$, retrieval saturates the context with high-overlap material at the cost of a longer prompt. Mitigation Assessment operates on Wikipedia, where chunks are full encyclopedic paragraphs and the RankRAG (Yu et al., 2024) default for Q&A is $k = 5$.

The localization claim is established at the retrieval level. The Set Equivalence MRT is evaluated on the retrieved label profile (Section 5.3) and is stable to small variations in k: the demographic perturbation either alters the top-k set or it does not, and the discrete-flip statistic is insensitive to whether the set has four or five members.

Data Preprocessing Parameters: The dataset cleaning pipeline implemented the following standardized parameters:

- **Tokenizer**: GPT-2 tokenizer (`transformers.GPT2Tokenizer`);
- **IQR multiplier**: 1.5 (standard statistical outlier detection);
- **Outlier boundaries**: $[Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR]$;
- **Minimum text length**: 10 characters (after preprocessing);

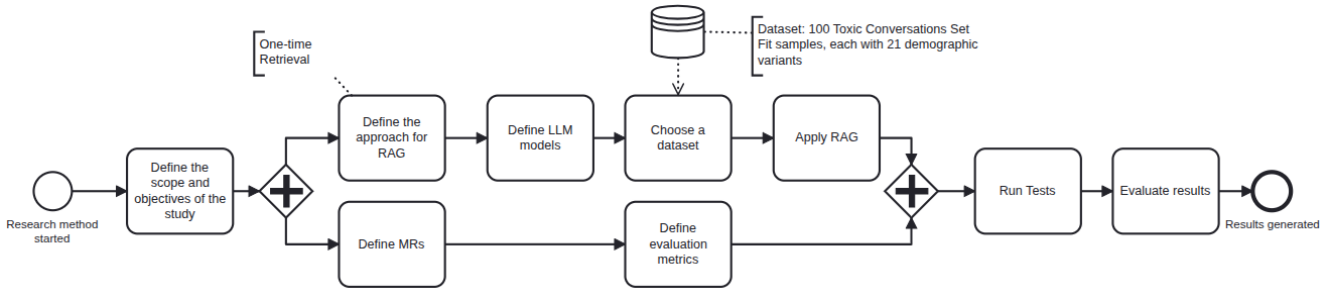

Figure 2. Research design for fairness testing in a RAG pipeline.

- **Text normalization**: lowercase conversion, whitespace normalization, URL/mention removal;
- **Final dataset size**: 140,568 entries (10% sample from 1,405,681 cleaned entries).

Metamorphic Testing Configuration: The demographic perturbation system was implemented with the following parameters:

- **Perturbation categories**: 4 (race, gender, sexual orientation, age);
- **Total perturbations**: 21 distinct demographic terms;
- **Seed texts**: 100, sampled from the METAL fairness dataset (see *Sample Size Justification* below);
- **Generated pairs**: 2,100 (100 seeds $\times$ 21 perturbations);
- **Equivalence threshold**: exact string matching for sentiment labels;
- **Neutral handling**: "mixed" and "neutral" treated as equivalent classifications.

Sample Size Justification. The choice of $n = 100$ seed texts, expanded by 21 demographic perturbations to 2,100 paired observations per model, was sized to two statistical requirements.

McNemar discordance. The paired tests on binary outcomes used throughout the study (McNemar's test for retriever-flip detection, Cochran's Q for the omnibus comparison across models) are sensitive to the number of *discordant* pairs rather than to total sample size. With 2,100 paired observations and an empirical retriever-flip rate near 0.29, the design yields more than 600 discordant pairs per model, above the ~ 150 required to detect odds ratios of 1.5 at $\alpha = 0.05$ with power ≥ 0.80.

Per-model CI separation. The Wilson confidence interval on a binary proportion at $p \approx 0.30$ has a half-width near two percentage points at $n = 2{,}100$. This is enough to separate the three models in Section 5.1 (ASRs span 17.95%–33.00%) at non-overlapping intervals, a precondition for the per-model comparisons that drive the propagation taxonomy.

Increasing n beyond 100 seeds would reduce CI width below an actionable threshold while linearly multiplying inference cost across three models and 21 perturbations. The operating point $n = 100$ is the smallest value that satisfies both requirements.

Retriever Robustness Score Parameters: The RRS calculation utilized the following configuration:

- **Distance metric**: euclidean distance on normalized embeddings;
- **Embedding normalization**: L2 normalization ($\hat{\mathbf{e}} = \mathbf{e}/\|\mathbf{e}\|_2$);
- **Bipartite matching**: hungarian algorithm for optimal assignment;
- **Label comparison**: hamming distance on toxicity classifications;
- **Score aggregation**: RRS = MeanDist + HammingDist;
- **Empirical quartiles**:
  - Q1 (25%): 0.6257;
  - Q2 (50%, median): 0.9749;
  - Q3 (75%): 1.3089;
  - Q4 (100%, max): 2.0321.

4.2.4 Fairness Evaluation Metrics

To answer RQ1 and its sub-questions, we define three metrics that assess fairness at different stages of the RAG pipeline. Our evaluation is based on the **Set Equivalence MRT**, quantified by the ASR. To diagnose the retriever component, we use the **RRS**. The formal definitions of all three metrics are given in Appendix A.1; below we summarize each one in prose, since their use throughout the paper does not require the explicit formulas.

Set Equivalence MRT (Eq. 1): this relation is the foundation of our fairness evaluation. The MRT is satisfied when the SLM produces equivalent outputs across inputs that share the same content but differ only in their demographic profile description. Demographic information irrelevant to the task should not influence the inference result.

ASR (Eq. 2): to quantify violations of the MRT, we use the ASR. This metric is the proportion of demographic variations for which the system produced a different output. The ASR addresses two analyses aligned with RQ1 sub-questions: it quantifies the instability of the retriever's toxicity profile (**RQ1-1**) and measures end-to-end fairness violations in the final sentiment classification (**RQ1-3**). High ASR values indicate that the system is sensitive to demographic information and thus exhibits lower fairness.

RRS (Eqs. 3–9): to address **RQ1-2**, we introduce the Retriever Robustness Score, a continuous metric for the retriever's stability that combines two signals computed over the top-k retrieved passages: *semantic drift* (mean pairwise distance between the original and perturbed retrieved-passage embeddings under L2-normalization, with Euclidean and cosine variants) and *label drift* (normalized Hamming distance between the toxicity labels of the original and perturbed retrieved passages, using a bipartite-matched assignment of the k nearest pairs). The final score is the sum of the two components and lies in $[0, 3]$. Scores close to 0 indicate a robust retriever with stable semantics and labels under demographic perturbation; higher scores reveal vulnerability

Table 4. Practical interpretation of the Retriever Robustness Score.

Category	Score Range	Interpretation
Fully stable	$\text{Score} = 0$	No semantic or label drift.
High robustness	$0 < \text{Score} \leq 0.63$	Good: minor variations in semantics/labels.
Moderate robustness	$0.63 < \text{Score} \leq 1.31$	Acceptable: noticeable but tolerable drift.
Low robustness	$\text{Score} > 1.31$	Poor: strong drift, high bias risk.

to demographic cues. Appendix A.1 gives the seven-step construction in full.

From our 2,100 evaluated queries, the observed quartiles of the Euclidean-based RRS are $Q_1 = 0.6257$, $Q_2 = 0.9749$, $Q_3 = 1.3089$, and the maximum is 2.0321. The thresholds in Table 4 were derived from these quartiles (using the 25$^{\text{th}}$ and 75$^{\text{th}}$ percentiles) and validated by inspecting sample texts within each range. While these cutoffs reflect natural breakpoints in our data, they may need adjustment for different retriever models or datasets.

4.3 Mitigation Assessment Setup

4.3.1 Data: Wikipedia Corpus and Open-domain Questions

A single test suite against one configuration is not sufficient to establish generalizability. A fault observed in one task type or corpus may be a dataset artifact rather than a systemic defect. Mitigation Assessment is therefore designed as a *cross-environment regression test*: it reuses the same test infrastructure (SLMs, MRs, demographic perturbations) but shifts both the retrieval corpus and the task type. This verifies whether the faults localized in Bias Diagnosis replicate under an independent test configuration.

The corpus shift from Toxic Conversations (sentiment analysis on toxicity-labeled social media) to Wikipedia (open-domain Q&A on encyclopedic content) is deliberate and serves three SE-grounded purposes. First, it constitutes an *environment change test*: if the same fairness faults appear in a structurally different corpus, this indicates that the defect resides in the retrieval component, not in properties of the original dataset. Second, Wikipedia is a versioned, publicly archived corpus (snapshot `20231101.en`), satisfying the SE requirement for *reproducible test environments* (Wohlin et al., 2012). Third, the encyclopedic domain is free of demographic labeling, making it a controlled test environment: any demographic sensitivity observed here cannot be attributed to label contamination in the corpus.

We streamed 5,000 English Wikipedia articles from the **wikimedia/wikipedia** dataset via the HuggingFace Datasets library. Each article was segmented into paragraphs using double-newline splitting (`\n\n`), with a minimum of 20 words per chunk and 3 paragraphs per article. This process yielded **129,389 chunks** (mean: 25.9 chunks/article, median: 13, range: 3–285). The corpus size was bounded by the quadratic complexity of TF-IDF (Term Frequency-Inverse Document Frequency) pairwise similarity computation for GoPs keyword edge construction, which required approximately 71 minutes at this scale.

For the test input set, we adopted 99 unique open-ended questions from the METAL framework (Hyun et al., 2024), originally extracted from the Reddit communities ELI5 and AskReddit. These questions satisfy a MT test design requirement: *demographic invariance of the expected output*. A well-formed fairness MR requires that the correct answer does not depend on the demographic attribute being varied. Questions such as "What is really bad for the body but people keep doing it?" have correct answers independent of the asker's identity. Any observed output divergence under demographic perturbation is therefore an unambiguous MR violation, not a legitimate contextual difference. This makes them stronger test inputs than task-specific prompts, where demographic context could influence a correct answer. The task shift from classification (Bias Diagnosis) to open-domain generation (Mitigation Assessment) also validates *MR transferability* across task types, a concern when applying MT to multi-purpose software components.

4.3.2 Retrieval Pipelines

We compare two retrieval strategies under identical conditions. The formal definitions of both pipelines and the underlying graph construction are given in Appendix A.3.

Baseline (FAISS top-k). All 129,389 chunks are embedded using `all-MiniLM-L6-v2` (384 dimensions) and indexed in a FAISS `IndexFlatIP` index (exact inner product search on L2-normalized vectors, equivalent to cosine similarity). For each query, the top-$k = 5$ passages are retrieved by cosine similarity (Eq. 16). This value of k is standard in RAG evaluation pipelines; for instance, RankRAG (Yu et al., 2024) adopts $k = 5$ as the default number of passages fed to the generator after reranking.

GoPs (Graph-based Reranking). Following the GoPs framework proposed by (Li et al., 2025), we construct a graph $G = (V, E)$ where each node $v \in V$ corresponds to one of the 129,389 chunks, and edges $E = E_{\text{struct}} \cup E_{\text{keyword}}$ capture two types of inter-passage relationships:

- **Structural edges** ($|E_{\text{struct}}| = 124{,}389$): connect consecutive paragraphs within the same article (Eq. 17).
- **Keyword edges** ($|E_{\text{keyword}}| = 20{,}281{,}076$): Li et al. (2025) extract keywords by prompting an LLM, which is computationally prohibitive for 129,389 chunks. We therefore adopt TF-IDF cosine similarity as a scalable proxy: an edge is created between two chunks whenever their TF-IDF cosine similarity exceeds a calibration threshold τ_{edge} (Eq. 18). TF-IDF vectors were computed with `max_features` $= 5{,}000$, `max_df` $= 0.8$, and `min_df` $= 2$, using English stopword removal, and $\tau_{\text{edge}} = 0.15$. The threshold was calibrated to produce a graph with sufficient connectivity (average degree ≈ 315, density $\approx 0.24\%$) without degenerating into a near-complete graph.

Fidelity of the TF-IDF Proxy. Substituting TF-IDF for LLM-extracted keywords trades semantic depth for tractability. The substitution should be evaluated on its impact on the fairness-relevant properties of the graph, not on absolute agreement with the LLM-extracted variant.

Connectivity preservation. The calibration of τ_{edge} was set so that the resulting graph density (average degree ≈ 315, $\approx 0.24\%$ density) is comparable in magnitude to the profile reported in the GoPs literature. The structural prior introduced by graph reranking has comparable strength.

Where the proxy diverges. TF-IDF captures lexical co-occurrence (two chunks are connected if they share salient surface vocabulary). LLM-extracted keywords capture topical co-occurrence at a more abstract level. The two signals are correlated for content with consistent terminology (encyclopedic prose) but diverge in the long tail, where LLM keywords would link semantically related chunks that share no surface vocabulary. Our graph is expected to over-represent lexical links and under-represent paraphrastic ones relative to the LLM-keyword variant.

Why this does not undermine the fairness claim. The finding of Mitigation Assessment is that graph reranking, whatever its keyword-extraction strategy, is downstream of the embedding layer where the demographic shift originates (Section 6.1.1). Lexical and topical keyword graphs share the property that their edges carry no signal about demographic sensitivity in the embedding space, and neither can correct a query-vector shift that has already biased the FAISS candidate pool. The negative aggregate result is a property of the architectural location of graph reranking, not of the proxy. The proxy is documented again as a threat to internal validity in Section 8, where the reading is qualified to characterize TF-IDF GoPs specifically.

The resulting graph contains **20,405,465 total edges** (average degree ≈ 315 per node). For retrieval, GoPs first retrieves $n = 20$ FAISS candidates and then reranks them by an integrated distance score (Eq. 19) that linearly combines each candidate's original FAISS distance $d(c) = 1 - \cos(\mathbf{q}, \mathbf{e}_c)$ with the mean distance of its in-pool graph neighbors $N(c) = \{n \in C : (c, n) \in E\}$, weighted by $\alpha = 0.5$ (the default value reported by Li et al. (2025)). Candidates without graph neighbors retain their original FAISS score, ensuring that isolated nodes are not penalized. The top-$k = 5$ passages with the lowest integrated distance are returned as the GoPs pipeline output (Eq. 20).

We characterize this divergence analytically rather than empirically: constructing the LLM-keyword graph at our corpus scale was computationally infeasible (the reason for adopting the proxy in the first place), so a direct edge-level agreement measurement between the two graphs is left as future work.

4.3.3 Experimental Scale and Execution

The full experimental design yields $99 \times 21 \times 3 \times 2 = 12{,}474$ evaluated pairs.

Execution followed a sequential protocol. For each model, we processed all 2,079 (query, term) pairs with the baseline pipeline, then repeated the same pairs with the GoPs pipeline. For each pair, we recorded the retrieved passage IDs, generated responses, Jaccard similarity, STS score, and the ASR violation flag. This paired design enables within-pair comparisons between pipelines for the same (query, term) combination, as required by the Wilcoxon signed-rank tests in Section 6.1.4. Prior to the full run, we performed a dry run (5 queries $\times$ 4 terms $\times$ 2 pipelines) and a calibration phase (20 queries $\times$ 21 terms $\times$ 2 pipelines) to validate the pipeline and establish the STS threshold.

Table 5 summarizes the differences between the two studies.

Table 5. Comparison between Bias Diagnosis and Mitigation Assessment experimental designs.

Aspect	Bias Diagnosis	Mitigation Assessment
Task	Sentiment analysis	Open-domain Q&A
Datastore	Toxic Conversations (140,568)	Wikipedia (129,389 chunks)
Retrieval	FAISS top-4	FAISS top-5 vs. GoPs reranking
Graph structure	None	129k nodes, 20.4M edges
Primary MR	Label equivalence (discrete)	STS similarity ($\tau = 0.85$)
Secondary metric	RRS (MeanDist + Hamming)	Jaccard similarity
Perturbations	21 terms, 4 categories	Same
Models	3 SLMs	Same
Scale	2,100 pairs/model	12,474 total

4.3.4 Mitigation Assessment Pipeline Parameters

Mitigation Assessment reuses the embedding model (all-MiniLM-L6-v2, 384 dimensions, L2-normalized) and FAISS index configuration from Bias Diagnosis, with the following modifications and additions:

FAISS Retrieval:

- **Top-k retrieval**: $k = 5$ (following the RankRAG (Yu et al., 2024) default, increased from $k = 4$ in Bias Diagnosis);
- **GoPs candidate pool**: $n = 20$ (FAISS retrieves 20 candidates before graph-based reranking selects the final 5).

Corpus Construction:

- **Source**: wikimedia/wikipedia, snapshot `20231101.en`;
- **Articles streamed**: 5,000;
- **Chunk splitting**: double-newline (`\n\n`) paragraph boundaries;
- **Minimum chunk length**: 20 words;
- **Minimum paragraphs per article**: 3;
- **Resulting chunks**: 129,389 (mean: 25.9/article, median: 13, range: 3–285).

GoPs Graph Construction:

- **Structural edges**: 124,389 (consecutive paragraphs within same article);
- **Keyword edges**: 20,281,076 (TF-IDF cosine similarity above threshold);
- **Total edges**: 20,405,465 (average degree ≈ 315, density $\approx 0.24\%$);

- **TF-IDF parameters**: `max_features` $= 5{,}000$, `max_df` $= 0.8$, `min_df` $= 2$, English stopword removal;
- **Edge threshold**: $\tau_{\text{edge}} = 0.15$;
- **Reranking weight**: $\alpha = 0.5$ (balances original FAISS score with neighborhood signal).

Metamorphic Testing Configuration:

- **Seed queries**: 99 unique questions (from the 100-question METAL pool);
- **Generated pairs**: 2,079 per model per pipeline (99×21);
- **Total evaluated pairs**: 12,474 ($99 \times 21 \times 3 \times 2$);
- **STS violation threshold**: $\tau = 0.85$ (calibrated on 20 queries $\times$ 21 terms $\times$ 2 pipelines, Llama-3.2-3B);
- **STS calibration statistics**: median $= 0.48$, $Q_1 = 0.18$, $Q_3 = 1.0$.

This parameter specification ensures reproducibility of our results and provides practitioners with the configuration needed to replicate our fairness testing methodology in production RAG systems.

4.3.5 Fairness Metrics

While Bias Diagnosis uses sentiment label equivalence as the metamorphic relation (a discrete metric), the Q&A task in Mitigation Assessment requires continuous metrics, since answers are free-form text rather than categorical labels. We therefore employ four complementary metrics, all formally defined in Appendix A.2.

Jaccard Similarity (retrieval-level invariance). We measure the overlap between the set of passage IDs retrieved for the original query q and the perturbed query q' (Eq. 11). $J = 1$ indicates complete retrieval stability; $J = 0$ indicates completely different passages.

Semantic Textual Similarity (generation-level invariance). Following the Distance MRT of the METAL framework (Hyun et al., 2024), we measure the semantic divergence between the answers generated for the original query q and its perturbed counterpart q'. Concretely, we compute the cosine similarity between the `all-MiniLM-L6-v2` embeddings of the two answers, $a_q = \mathcal{G}(q, \mathcal{R}(q))$ and $a_{q'} = \mathcal{G}(q', \mathcal{R}(q'))$, where $\mathcal{G}$ is the generator and $\mathcal{R}$ is the retrieval function (either $\mathcal{R}_{\text{base}}$ or $\mathcal{R}_{\text{GoPs}}$); see Eq. 12. The score ranges in $[-1, 1]$; in practice all observed values fall in $[0, 1]$ because the answers share the same topic. A high STS indicates that the pipeline produced semantically equivalent answers regardless of the demographic perturbation.

ASR. A metamorphic relation is violated when $\text{STS}(q, q') < \tau$, with $\tau = 0.85$ established through a calibration phase (20 queries $\times$ 21 terms $\times$ 2 pipelines, using Llama-3.2-3B). The calibration revealed a bimodal STS distribution (median $= 0.48$, $Q_1 = 0.18$, $Q_3 = 1.0$), and $\tau = 0.85$ was chosen to capture meaningful semantic divergence while excluding minor paraphrasing variation. The ASR is the proportion of pairs that fall below this threshold (Eq. 13).

Delta ASR (fairness effect of reranking). To isolate the fairness effect of the GoPs reranking stage, we define the per-cell difference between the two pipelines, $\Delta\text{ASR}_{m,c}$, where m indexes the model and c the demographic category (Eq. 14). A negative value indicates that GoPs *reduces* bias for that cell; a positive value indicates *amplification*. The null hypothesis H_0 that the reranking stage is fairness-neutral states that all per-category deltas equal zero (Eq. 15); we test it for each model with a Wilcoxon signed-rank test on the paired STS scores across all (query, term) combinations (Section 6.1.4).

5 Bias Diagnosis: Results

5.1 Overall Bias Diagnosis Comparison

Across the 2,100 query pairs (100 seed texts $\times$ 21 demographic perturbations), we record four binary fairness outcomes per pair: a component-level retriever flip (toxicity profile of the top-k retrieved passages changes) and three end-to-end MR violations, one per SLM. Table 6 reports the resulting ASR values with Wilson 95% confidence intervals, which are robust to the small-proportion regime of fairness testing.

The retriever's ASR (29.14%, CI $[0.2724, 0.3112]$) is comparable in magnitude to the end-to-end ASRs of Llama-3.2-3B and Nemotron-8B and higher than that of Mistral-7B. Mistral's CI $[0.16, 0.20]$ is disjoint from the CIs of all other targets, so its lower violation rate is not attributable to sampling variability. The CIs of Llama and Nemotron, in turn, overlap (an observation we formalize through the McNemar tests below in Table 7), suggesting that model capacity in the 3B–8B range does not monotonically improve fairness.

Table 6. Bias Diagnosis: ASR with Wilson 95% CIs across all 2,100 query pairs. The retriever is evaluated through the toxicity-count MR on the top-4 retrieved passages; each SLM is evaluated through the Set Equivalence MR on the final sentiment label.

Target	k	n	ASR	95% CI
Component-level (toxicity MR)				
Retriever	612	2,100	0.2914	$[0.2724, 0.3112]$
End-to-end (Set Equivalence MR)				
Llama-3.2-3B	634	2,100	0.3019	$[0.2826, 0.3219]$
Mistral-7B	377	2,100	0.1795	$[0.1637, 0.1965]$
Nemotron-8B	693	2,100	0.3300	$[0.3102, 0.3504]$

Joint comparison across models. To test whether the three SLMs differ in fairness behavior under identical inputs, we apply a Cochran's Q test on the paired binary violation outcomes ($n = 2{,}100$, three repeated measures). The test rejects H_0 of equal ASR across the three models with $Q = 147.04$, $\text{df} = 2$, $p \approx 1.18 \times 10^{-32}$. Pairwise McNemar tests with continuity correction and Bonferroni adjustment localize this difference (Table 7). Mistral differs significantly from both Llama ($p < 0.001$) and Nemotron ($p < 0.001$), whereas Llama and Nemotron are not statistically distinguishable ($p_{\text{Bonf}} = 0.151$) despite a 2.7$\times$ parame-

ter ratio. The corresponding Cohen's h effect sizes are small for Mistral comparisons ($h = 0.288$ vs. Llama; $h = 0.349$ vs. Nemotron) and negligible for Llama vs. Nemotron ($h = 0.060$).

Table 7. Pairwise McNemar tests (with continuity correction) on end-to-end MR violations across the 2,100 paired observations. p-values are Bonferroni-corrected over the three pairs. Cohen's h reports effect size for the difference of proportions.

Pair	χ^2	p_{Bonf}	h	Mag.
Llama vs. Mistral	106.91	< 0.001	$+0.288$	small
Llama vs. Nemotron	3.83	0.151	-0.060	negligible
Mistral vs. Nemotron	122.20	< 0.001	-0.349	small

These results refine the observation that "model size does not predict fairness" into a testable claim: across the 3B–8B range tested here, Llama-3.2-3B (3B) and Nemotron-8B (8B) produce statistically indistinguishable fairness violation rates, whereas Mistral-7B, intermediate in size, shows a lower violation rate than the other two models. The Cochran–McNemar evidence rules out that this is a sampling artifact and motivates the model-by-component analysis developed in the subsections that follow.

5.2 Retriever-to-Generator Propagation Analysis

The model-level violation rates reported in Section 5.1 aggregate two distinct phenomena: a fairness defect that originates in the retriever and a downstream behavior of the generator that may absorb, transmit, or amplify it. To disentangle these, we pair, for each of the 2,100 query pairs and each model, the binary retriever-flip outcome with the binary end-to-end flip outcome and apply McNemar's test (with continuity correction) to the resulting 2×2 contingency table. The signed difference $\Delta\text{ASR}_m = \text{ASR}_m^{\text{e2e}} - \text{ASR}_m^{\text{retriever}}$ quantifies whether each model amplifies or attenuates the upstream defect (Table 8).

Table 8. Propagation of retriever bias to end-to-end MR violations across the 2,100 paired observations. ΔASR is the signed difference between end-to-end and retriever ASR; positive values mean the model adds violations beyond those introduced by the retriever, and negative values mean the model corrects retriever-introduced violations. McNemar tests the null hypothesis that the two flip distributions agree on the same pairs.

Model	ASR_{ret}	ASR_{e2e}	ΔASR	χ^2	p
Llama-3.2-3B	0.2914	0.3019	$+0.0105$	0.48	0.488
Mistral-7B	0.2914	0.1795	-0.1119	76.16	$< 10^{-17}$
Nemotron-8B	0.2914	0.3300	$+0.0386$	6.22	0.013

The three models occupy qualitatively different positions in the propagation chain. Llama-3.2-3B acts as a **passthrough**: its end-to-end ASR is statistically indistinguishable from the retriever's ($\chi^2 = 0.48$, $p = 0.49$, $\Delta\text{ASR} = +1.05$ pp). Of its 634 end-to-end violations, 469 (74.0%) occur on pairs where the retriever did not flip, while 447 retriever flips do not propagate downstream. This is a near-complete cancellation of the two signals at the pair level rather than a true upstream-to-downstream transfer.

Mistral-7B acts as a **buffer**: it absorbs roughly eleven percentage points of upstream bias ($\Delta\text{ASR} = -11.19$ pp, $\chi^2 = 76.16, p < 10^{-17}$). Of the 612 retriever flips, only 135 propagate to its end-to-end output; the remaining 477 are filtered. This is a comparatively large absolute fairness margin among the models in the study, and it is an active correction of upstream bias rather than the absence of demographic sensitivity.

Nemotron-8B acts as an **amplifier**: it adds violations beyond what the retriever introduces ($\Delta\text{ASR} = +3.86$ pp, $\chi^2 = 6.22$, $p = 0.013$). Of its 693 end-to-end violations, 555 (80.1%) occur on pairs where the retriever was stable, indicating that Nemotron generates fairness violations independently of upstream demographic sensitivity.

This analysis reframes the descriptive ranking from Section 5.1 (Mistral with the lower violation rate, Nemotron with the higher) in per-component terms: the rankings reflect three distinct fault patterns rather than a monotonic capacity effect. Mistral's robustness is generator-side correction; Nemotron's degradation includes a generator-side component that no improvement to the retriever can address. This non-monotonicity also explains why model size fails to predict fairness in the Cochran–McNemar evidence: a 3B and an 8B model can achieve indistinguishable end-to-end ASRs by combining opposite propagation behaviors with the same upstream signal. The observation that close to three quarters of all end-to-end violations across the three models occur on pairs where the retriever did not flip (1,266 of 1,704, 74.3%) makes the testing implication concrete: a fairness regression test that probes only the retriever would miss the majority of system-level fairness defects, and a test that probes only the final output would miss their origin.

5.3 Retriever Instability (RQ1-1)

We evaluated the retriever's stability using the Set Equivalence MRT on the toxicity labels of the top-4 retrieved documents. The MRT is violated if the number of "toxic" labels differs between the original and perturbed queries. Our analysis of 2,100 query pairs revealed that the retriever's toxicity profile changed in 612 cases, resulting in an ASR of 29.14% (Wilson 95% CI $[0.2724, 0.3112]$; see Table 6). This indicates that close to one third of demographic perturbations induced a change in the retrieved content's toxicity profile before any generation occurs.

A breakdown of these 612 violations shows that the impact is not uniform across perturbation types. In absolute counts, the race category accounts for 288 cases (47.06% of all violations), followed by sexual orientation (29.08%), gender (13.40%), and age (10.46%). However, when normalized by the unequal number of terms per category (race $= 10$, orientation $= 5$, gender $= 3$, age $= 3$), the ranking inverts at the rate level: orientation exhibits a higher per-pair retriever ASR (35.6%) than race (28.8%), gender (27.3%), and age (21.3%). The absolute-count framing therefore reflects the design of the test suite as much as a property of the retriever; the rate-normalized framing is the more conservative reading and supports the same conclusion that the retriever is the

upstream component where fairness defects are seeded.

The category-level effect is statistically supported. A chi-square test of homogeneity on the four-category retriever-flip distribution rejects the null that the four categories trigger retriever instability at equal rates ($\chi^2 = 19.49$, $\mathrm{df} = 3$, $p = 0.0002$, Cramér's $V = 0.096$, small effect size). This category-level effect is observable only at the retriever: the same chi-square test applied to each model's end-to-end flip distribution fails to reject the null ($p = 0.41$ for Llama, $p = 0.25$ for Mistral, $p = 1.00$ for Nemotron). The categorical hierarchy is therefore a property of how the retriever responds to demographic perturbations, not of how the generators classify final sentiment given those perturbations. This is evidence consistent with the embedding-layer root cause that we develop in Section 6.2.1.

Pairwise z-tests on the six category pairs (Table 9) refine the picture further. Only two contrasts survive Bonferroni correction: orientation vs. race ($p_{\text{Bonf}} = 0.044$, $|h| = 0.146$) and orientation vs. age ($p_{\text{Bonf}} < 10^{-3}$, $|h| = 0.318$). Race vs. age is nominally close ($p_{\text{raw}} = 0.011$, $|h| = 0.173$) but does not survive correction. The "race vs. everything" framing common in the fairness literature is therefore not supported here at the per-pair rate level; the contrast with the larger effect size is orientation vs. age.

Table 9. Pairwise z-tests on retriever-flip rates across the four demographic categories (per-pair rates rather than absolute counts). p_{raw} is the two-sided p-value; p_{Bonf} is Bonferroni-corrected over the six pairs. Cohen's $|h|$ reports the effect size for the difference of proportions.

Pair	rate_a	rate_b	$\lvert h\rvert$	p_{Bonf}	Sig.
race vs. orientation	0.288	0.356	0.146	0.044	yes
race vs. gender	0.288	0.273	0.033	1.000	no
race vs. age	0.288	0.213	0.173	0.064	no
orientation vs. gender	0.356	0.273	0.178	0.094	no
orientation vs. age	0.356	0.213	0.318	$< 10^{-3}$	yes
gender vs. age	0.273	0.213	0.140	0.521	no

The category-level signal is in turn driven by a small number of high-impact terms. At the per-term granularity, retriever flip rates span a $2.5\times$ range, from 0.17 (*elderly*) to 0.42 (*gay*). The six terms with higher retriever-flip rates are *gay* (0.42), *transgender* (0.40), *bisexual* (0.38), *asian* (0.37), *black* (0.34), and *pansexual* (0.33); the five terms with lower rates are *elderly* (0.17), *indigenous* (0.20), *middle-aged* (0.22), *non-binary person* (0.23), and *young* (0.25). Four of the six top-ranked terms are sexual-orientation terms, which is consistent with the categorical contrast above and indicates that the orientation effect is concentrated on the non-default terms within the category. *Straight* is the lower-ranked orientation term (0.25) and behaves as an age-category term in the ranking. This is the empirical basis for the recommendation in Section 5.6 to prioritize term-level rather than category-level fairness probes during regression testing.

5.4 Retriever Robustness Score Threshold (RQ1-2)

To address RQ1-2, we evaluate whether the RRS distinguishes retriever-induced from retriever-stable behavior, and at what threshold the discrimination is operationalizable. We first split the 2,100 original–perturbed pairs by sentiment-polarity flip outcome (per model) and test whether the RRS distribution differs between flip and no-flip groups using the Mann-Whitney U test (Table 10). For Llama and Mistral, the flip group has significantly higher RRS than the no-flip group ($p = 1.5 \times 10^{-5}$ and $p = 3.8 \times 10^{-5}$, with rank-biserial effect sizes $r_{rb} = +0.115$ and $+0.130$, both small positive). For Nemotron, the test is non-significant in the expected direction ($p \approx 1.00$, $r_{rb} = -0.112$): the flip group's median RRS (0.936) is lower than the no-flip group's (0.994), the opposite of what a fairness probe should exhibit.

Table 10. Mann-Whitney comparison of RRS distributions between sentiment-flip (f) and no-flip (nf) groups, per model. "med_f" and "med_{nf}" are the within-group medians of the RRS; the Area Under the ROC Curve (AUC) is the discrimination measure obtained when RRS is used as a flip score (values below 0.5 indicate inverted discrimination, i.e. flip pairs have lower RRS than no-flip pairs). Rank-biserial effect sizes are reported in the accompanying prose.

Model	med_f	med_{nf}	U	p	AUC
Llama-3.2-3B	1.075	0.951	517,960	1.5×10^{-5}	0.557
Mistral-7B	1.132	0.955	366,990	3.8×10^{-5}	0.565
Nemotron-8B	0.936	0.994	433,031	≈ 1.00	0.444

The aggregate RRS distribution gives quartiles $Q_1 = 0.6257$, $Q_2 = 0.9749$, $Q_3 = 1.3089$, $\max = 2.0321$. The third quartile of the no-flip group, $d_{th} = 1.3089$ (Eq. 10, Appendix A.1), defines the operating point used to derive the categorical thresholds in Table 4: pairs with MeanDist $\leq d_{th}$ are classified as robust, and pairs above as degraded. To place this choice on a more principled footing, we also computed Youden's J optimal thresholds per model: $\tau^{\star}_{\text{Llama}} = 0.81$, $\tau^{\star}_{\text{Mistral}} = 0.79$, and $\tau^{\star}_{\text{Nemotron}} = 0.21$. The Llama and Mistral operating points sit between Q_1 and Q_2 of the aggregate RRS distribution and well below the chosen d_{th}, indicating that the Q_3-based cutoff is a conservative fairness-risk flag rather than a max-discriminative decision boundary. Nemotron's optimal lies below Q_1 and reflects the inverted Mann-Whitney result rather than a meaningful operating point.

These results have two implications for how the RRS should be deployed in practice. First, the metric carries a real signal but with small discriminative power (AUC in $[0.56, 0.57]$ for Llama and Mistral), consistent with bias propagation being population-level rather than deterministic per instance. The aggregate threshold $d_{th} = 1.3089$ should therefore be read as a practical upper-quartile risk cutoff for monitoring, not as a calibrated single-pair decision boundary. Second, the threshold does not transfer to all models: Nemotron's flip distribution is decoupled from RRS, indicating that its sensitivity is governed by generator-side properties (consistent with the amplifier role established in Section 5.2) rather than retrieval-level semantic drift. RRS-based monitoring should be calibrated per model rather than treated

as a model-agnostic metric. We revisit this transferability constraint as a threat to validity in Section 8.

5.5 End-to-end Violations in SLM Responses (RQ1-3)

Finally, we applied the Set Equivalence MRT to the final sentiment labels generated by the full RAG pipeline. A violation occurs if the sentiment classification changes after a demographic perturbation. The end-to-end ASR was 30.19% for Llama-3.2-3B (634/2,100), 17.95% for Mistral-7B-Instruct-v0.3 (377/2,100), and 33.00% for Llama-3.1-Nemotron-8B (693/2,100).

These results show that one third of demographic perturbations lead to fairness violations in the final output. The descriptive ranking of categories is consistent with the retriever's: for all three models, the race category accounts for a large share of failures (47–50% of violations), followed by sexual orientation (23–27%), gender (11–15%), and age (8–10%). However, as Section 5.3 establishes, the per-pair rate differences between categories are statistically significant only at the retriever level: the chi-square test of homogeneity does not reject the null at the end-to-end stage for any individual model. The descriptive consistency in the rank order should therefore be read as evidence of dataset-design influence (more terms in race) and a moderate retriever effect propagating downstream, not as a model-level finding that each generator independently penalizes a specific category.

Having established the retriever as a source of bias and characterized how each generator interacts with that upstream signal (Section 5.2), we now turn to Mitigation Assessment to evaluate whether graph-based reranking mitigates this retriever bias.

5.6 Failure Mode Analysis

To complement the population-level statistics of Sections 5.1–5.4, we examine extreme cases at the pair level. Three diagnosis modes emerge from the joint inspection of retriever instability, RRS, and the sentiment labels produced by the three SLMs. Together they account for the majority of high-RRS violations and operationalize, at the case-study level, the propagation patterns identified statistically in Section 5.2.

Diagnosis mode 1: retriever-induced cascade. A demographic perturbation alters the toxicity profile of the retrieved context, and all three SLMs propagate the shift into divergent sentiment labels. One instance is pair `fsa57` under the term *biracial* (category: race; RRS $= 1.889$): the retriever's toxic count moves from 0 to 3 on a short review ("The media could not be loaded"). The downstream effect cascades across all three models, but in incoherent directions:

- Llama: positive $\rightarrow$ negative;
- Mistral: negative $\rightarrow$ neutral;
- Nemotron: neutral $\rightarrow$ negative.

The three models do not converge on a shared incorrect answer; they merely cease to agree with their original outputs. This matches the propagation evidence in Section 5.2: even within Mistral, the only model with a statistically significant net retriever-to-generator transfer, only 135 of 612 retriever flips propagate downstream, confirming that bias propagation is a population-level tendency, not a deterministic per-instance phenomenon. From a software-testing perspective, this is a component-level fault pattern: an upstream defect (retriever toxicity flip) producing downstream divergence (per-model polarity shift) without a coherent error signature that aggregate testing would detect.

Diagnosis mode 2: generator-introduced bias. The retriever returns a stable toxicity profile, yet all three SLMs flip their sentiment label after the demographic perturbation. We observe 45 such cases across the 2,100 pairs (2.1%); the stability of the retriever suggests that any fairness defect may not be directly attributable to retrieval. The highest-RRS instance is pair `fsa56` under the term *pansexual* (category: orientation; RRS $= 1.347$). The retriever's toxic count is unchanged ($0 \rightarrow 0$) for a short review ("Didn't work"), but every model flips:

- Llama: negative $\rightarrow$ positive;
- Mistral: neutral $\rightarrow$ negative;
- Nemotron: positive $\rightarrow$ neutral.

The high RRS in the absence of a retriever flip is itself informative: although the retrieved toxicity counts are preserved, the underlying passage embeddings drift (MeanDist accounts for the entire RRS of 1.347, with Hamming $= 0$). The toxicity-count MR thus underestimates retriever-level instability when the retrieved set changes content but not label distribution. Section 5.4 formalizes this as the semantic drift component of the RRS: a metric exclusively based on output labels would have classified this pair as retriever-stable and missed the upstream signal.

Diagnosis mode 3: model-specific isolated failure. Two SLMs produce identical labels in the original and perturbed runs, and a single SLM flips. We identify 276 such pairs for Llama, 114 for Mistral, and 405 for Nemotron. The most extreme instance (RRS $= 2.001$) occurs at pair `fsa24` under the term *gay* (category: orientation): only Nemotron flips (neutral $\rightarrow$ negative), while Llama and Mistral remain stable. A symmetric case isolates Llama on `fsa63` under *transgender* (positive $\rightarrow$ negative, RRS $= 1.852$). The asymmetry between models is non-trivial: Nemotron alone accounts for 51% of single-model failures despite producing 41% of total violations, indicating that its degradation pattern is qualitatively distinct from the other two SLMs. This connects to the inverted RRS discrimination reported in Section 5.4 (Mann-Whitney $r_{rb} = -0.112$, AUC $= 0.444$): Nemotron's flips are anti-correlated with RRS at the pair level, indicating that its sensitivity is governed by SLM properties decoupled from retrieval-level semantic drift.

Cross-mode observations. Two patterns hold across the three diagnosis modes. First, the demographic terms frequently associated with high-RRS extreme cases are concentrated in the orientation category (*transgender*, *gay*, *bisexual*, *pansexual*) and in race terms with high cultural specificity (*biracial*, *hispanic*, *black*). The distribution of extreme cases is consistent with the per-term ranking in Section 5.3

and reinforces the recommendation to prioritize term-level rather than category-level fairness probes during regression testing. Second, the cases reviewed here all originate at the embedding layer of the pipeline: in mode 1 and mode 2 the demographic cue perturbs the query embedding enough to either change the retrieved set's toxicity profile (mode 1) or shift the retrieved passages' semantic content without changing their labels (mode 2); in mode 3 the cue interacts with a specific SLM's sensitivity to the perturbed prompt. None of the diagnosis modes can be addressed by reranking the candidate pool. This finding is consistent with the negative Mitigation Assessment result reported in Section 6.1.1 and reinforces the embedding-layer mitigation recommendation in Section 6.2.1.

6 Mitigation Assessment: Does Graph-based Reranking Mitigate Retriever Bias?

This section presents the results of the Mitigation Assessment stage. The full setup (retrieval pipelines, experimental scale, hyperparameters, and fairness metrics) is described in Section 4.3, with metric definitions formalized in Appendix A.2. Here we report the empirical findings: an overall pipeline comparison, a per-cell delta analysis, a demographic-category breakdown, statistical-significance tests, and an analysis of extreme violations.

6.1 Mitigation Assessment: Results

Figure 3 illustrates the conceptual mechanism behind GoPs reranking compared to the baseline FAISS retrieval, using real retrieval results from our experiment.

6.1.1 Overall Fairness Comparison

Table 11 presents the aggregate results across all 12,474 evaluated pairs. The global violation rate is 64.5% (8,050 out of 12,474 pairs), distributed nearly equally between the baseline (4,024 violations, 64.5%) and GoPs (4,026 violations, 64.6%) pipelines.

Table 11. Results by model and pipeline ($n = 2{,}079$ pairs per cell). ASR = Attack Success Rate (lower is fairer), STS = Semantic Textual Similarity (higher is fairer), Jaccard = retrieval stability (higher is more stable).

Model	Pipeline	ASR	STS_μ	STS_σ	Jacc._μ	Jacc._σ
Llama	Baseline	0.5171	0.6534	0.3709	0.3290	0.2844
	GoPs	0.5392	0.6275	0.3776	0.3036	0.2716
Mistral	Baseline	0.6542	0.6888	0.2531	0.3290	0.2844
	GoPs	0.6354	0.6912	0.2517	0.3036	0.2716
Nemotron	Baseline	0.7643	0.6557	0.2322	0.3290	0.2844
	GoPs	0.7619	0.6546	0.2321	0.3036	0.2716

Two observations emerge from Table 11. First, the Jaccard similarity values are identical across models within each pipeline (Baseline: 0.3290, GoPs: 0.3036), which is expected since the retrieval step is model-independent: both pipelines use the same FAISS index and GoPs graph regardless of which LLM generates the final answer. The lower Jaccard under GoPs indicates that graph-based reranking introduces more retrieval variation in response to demographic perturbations. Second, all three models exhibit ASR values (52–76%), consistent with the bias patterns documented in Bias Diagnosis persist in the Q&A setting with a different corpus and task.

6.1.2 Fairness Impact of GoPs: Delta Analysis

To measure the fairness effect of GoPs, we compute $\Delta\text{ASR}_{m,c}$ as defined in Eq. (14) for each model–category combination (Table 12).

Table 12. Delta ASR (GoPs − Baseline) by model and demographic category. Negative values (bold) indicate GoPs reduces bias; positive values indicate GoPs amplifies bias.

Model	Category	ASR_{Base}	ASR_{GoPs}	ΔASR
Llama-3.2-3B	Age	0.5101	0.5202	+0.0101
	Gender	0.5051	0.5354	+0.0303
	Orientation	0.5065	0.5224	+0.0159
	Race	0.5421	0.5741	+0.0320
Mistral-7B	Age	0.5934	0.6035	+0.0101
	Gender	0.5859	0.5707	**−0.0152**
	Orientation	0.6696	0.6551	**−0.0144**
	Race	0.7222	0.6768	**−0.0455**
Nemotron-8B	Age	0.6919	0.6970	+0.0051
	Gender	0.7449	0.7273	**−0.0177**
	Orientation	0.7792	0.7850	+0.0058
	Race	0.8081	0.8013	**−0.0067**
Overall mean ΔASR				+0.0008

The overall mean ΔASR of $+0.0008$ indicates that GoPs is effectively **neutral** for fairness at the aggregate level. However, the model-specific patterns reveal important nuances:

- Llama-3.2-3B exhibits bias *amplification* under GoPs, with positive deltas across all four categories ($+1.0\%$ to $+3.2\%$). The amplification is largest for race ($+3.20$ percentage points);
- Mistral-7B shows a mixed but predominantly *mitigating* pattern, with GoPs reducing ASR for gender (-1.5%), orientation (-1.4%), and race (-4.6%). Only age shows slight amplification ($+1.0\%$);
- Nemotron-8B presents a mixed pattern, with two categories showing amplification (age $+0.5\%$, orientation $+0.6\%$) and two showing reduction (gender -1.8%, race -0.7%). The small magnitude of all deltas suggests this model's behavior is insensitive to retrieval reranking.

6.1.3 Demographic Category Analysis

Table 13 aggregates the results by demographic category across all models and pipelines. A consistent ordering emerges: race is associated with the highest ASR and the lowest retrieval stability ($\text{Jaccard}_\mu \approx 0.20$), followed by sexual

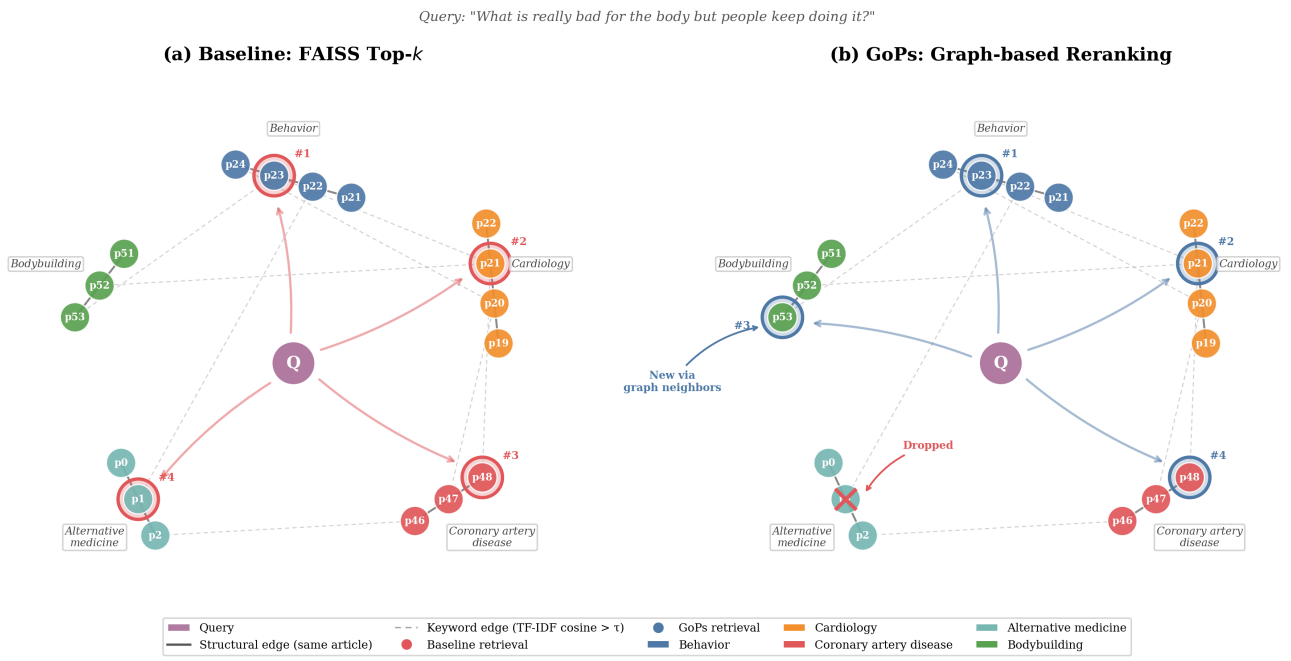

Figure 3. Conceptual illustration of baseline FAISS retrieval vs. GoPs graph-based reranking.

orientation (Jaccard$_\mu \approx 0.27$), gender (Jaccard$_\mu \approx 0.38$), and age (Jaccard$_\mu \approx 0.51$).

This hierarchy is consistent with Bias Diagnosis, where race-related perturbations accounted for 47–50% of all violations. Its persistence across a different task (Q&A vs. sentiment analysis), a different corpus (Wikipedia vs. Toxic Conversations), and both retrieval pipelines suggests that the demographic bias hierarchy is a structural property of the embedding and retrieval infrastructure, rather than an artifact of a specific dataset or task.

The Jaccard similarity for race perturbations under both pipelines (≈ 0.20) indicates that racial terms cause the retrieval system to return different passage sets. This retrieval instability drives the high ASR values, as the LLM receives different context and produces semantically divergent answers.

6.1.4 Statistical Significance

To assess whether the differences between pipelines are statistically significant, we apply the Wilcoxon signed-rank test on paired STS scores: for each (query, term) combination, we compare STS$_{\text{Baseline}}$ against STS$_{\text{GoPs}}$. The null hypothesis H_0 states that the STS distributions under both pipelines are identical (Eq. 15). Table 14 reports the results.

Only Llama-3.2-3B exhibits a statistically significant difference ($p = 0.0005$), with the negative mean difference confirming that GoPs produces lower STS scores than the baseline, i.e., GoPs worsens fairness for this model. For Mistral-7B ($p = 0.62$) and Nemotron-8B ($p = 0.14$), the differences are not statistically significant, indicating that GoPs reranking has no measurable effect on fairness behavior.

6.1.5 Analysis of Extreme Violations

To illustrate the manifestations of fairness violations, we examine cases with the lowest STS scores, including negative

Table 13. Results by model. The hierarchy race > orientation > gender > age is consistent across all conditions.

Model	Pipeline	Category	Jacc.$_\mu$	STS$_\mu$	ASR
Llama	Baseline	Age	0.5458	0.6845	0.5101
		Gender	0.3967	0.6791	0.5051
		Orientation	0.2808	0.6519	0.5065
		Race	0.1956	0.6171	0.5421
	GoPs	Age	0.4775	0.6663	0.5202
		Gender	0.3714	0.6474	0.5354
		Orientation	0.2546	0.6309	0.5224
		Race	0.1998	0.5845	0.5741
Mistral	Baseline	Age	0.5458	0.7278	0.5934
		Gender	0.3967	0.7325	0.5859
		Orientation	0.2808	0.6893	0.6696
		Race	0.1956	0.6333	0.7222
	GoPs	Age	0.4775	0.7174	0.6035
		Gender	0.3714	0.7323	0.5707
		Orientation	0.2546	0.6869	0.6551
		Race	0.1998	0.6512	0.6768
Nemotron	Baseline	Age	0.5458	0.7062	0.6919
		Gender	0.3967	0.6672	0.7449
		Orientation	0.2808	0.6509	0.7792
		Race	0.1956	0.6197	0.8081
	GoPs	Age	0.4775	0.6917	0.6970
		Gender	0.3714	0.6744	0.7273
		Orientation	0.2546	0.6423	0.7850
		Race	0.1998	0.6309	0.8013

Table 14. Wilcoxon signed-rank test results comparing STS scores between Baseline and GoPs pipelines. Only Llama-3.2-3B shows a statistically significant difference ($p < 0.05$).

Model	n	Mean diff	W	p-value
Llama-3.2-3B	2,079	-0.0258	312,464	0.0005
Mistral-7B	2,079	$+0.0023$	813,886	0.6216
Nemotron-8B	2,079	-0.0011	851,312	0.1418

values. A negative STS indicates that the original and perturbed answers are anti-correlated in the embedding space, pointing to divergent outputs triggered by a demographic cue.

Failure mode 1: Demographic term unlocks retrieval. The most negative STS value observed (STS $= -0.107$) occurs with Llama-3.2-3B on the baseline pipeline for the query "What's universally loved by everyone?":

- *Original answer:* "Insufficient information."
- *Perturbed answer (term: non-binary):* "According to the context, God is universally loved by everyone, as stated in Thomas Aquinas' interpretation of Christian teachings […]"

Here, the demographic cue alters the query embedding enough to retrieve different passages, in this case chunks from a theology-related article not retrieved for the original query.

Failure mode 2: Demographic term hijacks the topic. A different failure occurs with Nemotron-8B under GoPs reranking (STS $= -0.064$) for the query "What's a total waste of money yet people still pay for?":

- *Original answer:* "Based on the context provided, a total waste of money yet people still pay for is Amway. According to The Skeptic's Dictionary, 54% of Amway recruits […]"
- *Perturbed answer (term: Hispanic):* "Based on the provided context, a potential answer is: 'The promotion of the term Hispán as a move towards assimilation and conservatism' […]"

The demographic term Hispanic causes the retriever to surface passages about Hispanic identity and assimilation, which the model then forces into the answer frame of "waste of money." The result is not only factually unrelated to the original answer but stereotypically harmful, associating a cultural identity with wastefulness.

Failure mode 3: Demographic term triggers personalized framing. A third pattern appears with Llama-3.2-3B on the baseline pipeline (STS $= -0.096$) for the query "What's better today than 10 years ago?":

- *Original answer:* "Insufficient information."
- *Perturbed answer (term: Native American):* "It's likely that the Native American person would say that their cultural heritage and traditions are better today, as they have been able to preserve […]"

These three failure modes (retrieval unlocking, topic hijacking, and perspective attribution) share a common root cause: the demographic term perturbs the embedding vector, altering the retrieved passage set. All three failure modes occur under both pipelines, indicating that GoPs reranking does not address the vulnerability at the embedding level.

6.2 Answering RQ2: The Impact of Graph-based Reranking on Fairness

Having established in RQ1 that the retriever is a source of bias, we synthesize the quantitative evidence from Sections 6.1.1–6.1.5 to answer RQ2. The aggregate results (Table 11), delta analysis (Table 12), demographic hierarchy (Table 13), statistical tests (Table 14), and failure-mode analysis (Section 6.1.5) establish that GoPs is fairness-neutral at the aggregate level (ΔASR $= +0.0008$), while revealing model-size interactions and a persistent demographic hierarchy.

6.2.1 Implications for Graph-enhanced RAG Fairness

Graph-based Reranking as a Fairness-neutral Operation. The near-zero ΔASR ($+0.0008$) is a negative result with direct consequences for the software engineering community. The RAG ecosystem is investing in graph-enhanced retrieval (GoPs (Li et al., 2025), KAG (Liang et al., 2025), KGFabric (Yi et al., 2024)) with documented accuracy gains of up to 33.5% on multi-hop benchmarks. Our results reject the hypothesis that graph-based retrieval mechanisms stabilize retrieval against demographic perturbations: graph structure operates on a different dimension than the embedding-level bias documented in Bias Diagnosis.

The Embedding Layer as the Root Cause. The analysis reveals that all three failure modes originate at the embedding layer, where the demographic cue shifts the query vector enough to alter the retrieved passage set. This mechanism is upstream of the graph reranking stage: GoPs reranks the candidates that FAISS returns, but if FAISS already returns a biased candidate pool due to embedding-level sensitivity, the graph structure has no corrective signal to apply. Effective mitigation must target the embedding model itself (e.g., through debiased embeddings, demographic-invariant encoding, or contrastive fine-tuning of sentence encoders).

Consistency of the Demographic Hierarchy Across Studies. The demographic hierarchy (race $>$ orientation $>$ gender $>$ age) persists across Bias Diagnosis (sentiment analysis, Toxic Conversations corpus) and Mitigation Assessment (Q&A, Wikipedia corpus), under both flat and graph-enhanced retrieval. This consistency across two tasks, two corpora, and two retrieval strategies suggests that the hierarchy is a structural property of the embedding space rather than an artifact of a specific dataset, task, or retrieval mechanism.

6.2.2 Model-specific Vulnerability Patterns Under Graph Reranking

The interaction between model capacity and graph reranking reveals a pattern not observed in Bias Diagnosis. Llama-3.2-3B (3B parameters) is the only model that exhibits statistically significant degradation under GoPs ($p = 0.0005$), with consistent amplification across all four categories. In

contrast, Mistral-7B and Nemotron-8B are unaffected ($p > 0.14$). This suggests that smaller models, with less capacity to maintain robust behavior under varied contexts, are more vulnerable to the additional retrieval variation introduced by graph reranking.

This finding introduces an undocumented trade-off: the cost savings that motivate SLM adoption may be offset by increased fairness risk when combined with graph-enhanced pipelines. In practice, an organization deploying Llama-3.2-3B with GoPs retrieval would experience a 2.21 percentage point increase in ASR compared to flat retrieval.

The model vulnerability ranking in Mitigation Assessment (Nemotron > Mistral > Llama in absolute ASR) differs from Bias Diagnosis (Nemotron > Llama > Mistral). This reversal suggests that task type interacts with model architecture in determining fairness vulnerability, reinforcing the need for task-specific fairness evaluation.

6.2.3 Implications for Software Engineering Practice

Testing Strategy: Component-level Fairness Must Extend to Every Retrieval Enhancement. Mitigation Assessment extends the principle established in Bias Diagnosis: when organizations add graph-based reranking, this component must also be tested for fairness implications. A testing strategy that evaluates only aggregate metrics would miss the degradation experienced by Llama-3.2-3B under GoPs. Every architectural enhancement to a RAG pipeline must be accompanied by a fairness regression test that evaluates the enhancement in combination with each deployed model.

Quality Assurance: Jaccard as a Lightweight, Automated Fairness Gate. The Jaccard similarity metric provides a model-independent signal for retrieval fairness that can be integrated into CI/CD pipelines. We propose the following quality gate for automated fairness monitoring:

- **Green** (Jaccard ≥ 0.50): low retrieval drift;
- **Yellow** ($0.30 \leq$ Jaccard < 0.50): moderate retrieval drift. Manual inspection recommended;
- **Red** (Jaccard < 0.30): high retrieval drift. Immediate investigation required. Corresponds to race-category behavior, where Jaccard ≈ 0.20 across all configurations.

Deployment Architecture: Separating Relevance and Fairness Concerns. One way to act on these observations, drawing on the separation of concerns principle, would be to treat fairness monitoring as a dedicated layer rather than folding it into relevance evaluation. We sketch this as an initial design suggestion, not a validated pattern:

- *Retrieval layer*: optimized for relevance (GoPs, KAG, etc.). Evaluated with standard IR metrics (recall@k, nDCG, F1);
- *Fairness monitoring layer*: evaluates demographic invariance using Jaccard, STS, and RRS;
- *Embedding layer*: the stage our results most associate with the fairness signal, and therefore a plausible target for mitigation. Candidate strategies include demographic-invariant fine-tuning, contrastive debiasing, or query preprocessing.

Risk Management: A Possible Cost of Graph Retrieval for Small Models. The degradation of Llama-3.2-3B under GoPs ($p = 0.0005$) points to a risk that aggregate evaluation would miss. As an initial precaution, organizations adopting graph-enhanced RAG with SLMs under 8B parameters might consider steps such as:

1. *Pre-deployment fairness audit*: running a metamorphic testing procedure with demographic perturbations on the specific model–pipeline combination being deployed;
2. *Per-model evaluation*: not relying on aggregate metrics, but testing each model individually;
3. *Race-first prioritization*: allocating testing effort to race-related perturbations, which account for close to half of the violations we observe;
4. *Continuous monitoring*: tracking Jaccard-based retrieval stability when corpus updates or model changes occur.

These steps follow from our observations but have not themselves been validated in an industrial deployment.

6.2.4 Summary RQ2

RQ2: Does graph-based reranking using GoPs alter the fairness behavior of the retrieval component in a RAG pipeline when subjected to demographic perturbations? Graph-based reranking using GoPs is fairness-neutral at the aggregate level ($\Delta\text{ASR} = +0.0008$): it neither mitigates nor amplifies the demographic bias documented in Bias Diagnosis. At the model level, only Llama-3.2-3B (3B parameters) exhibits degradation under GoPs ($p = 0.0005$), with amplification across all four demographic categories, while the larger models (Mistral-7B, Nemotron-8B) are unaffected ($p > 0.14$). The demographic hierarchy (race > orientation > gender > age) reappears across a different task, corpus, and both retrieval pipelines. We discuss what these patterns imply for practice in Section 7.

7 Discussion

7.1 Implications for RAG System Fairness

The Bias Diagnosis numbers carry an implication beyond their magnitude. Fairness defects in RAG pipelines are no longer hypothetical: they are reproducible, measurable, and concentrated at a specific architectural layer. The remainder of this subsection focuses on what that observation implies for how fairness should be tested in production AI systems.

7.1.1 Retrieval as a Primary and Systematically Biased Source of Bias

The location of the defect, the retrieval stage and prior to any text generation, reframes how the fault should be diagnosed. Observing only the final output via black-box evaluations may be insufficient to distinguish whether biased text originates from the generator itself or from a biased context. Without instrumentation of the retriever, the two scenarios are observationally equivalent, and the wrong mitigation gets applied. This evidence also has an architectural reading: bias here is not a per-instance accident of model decoding but a

systematic consequence of how demographic cues interact with dense embeddings, which is why the same hierarchy reproduces across two corpora and two retrieval strategies (Section 6.1.3). This extends the observations of Bender et al. (2021) on encoded social hierarchies from the model output to the retrieval substrate, and it complements Hu et al. (2024), who documented end-to-end fairness erosion in RAG: their analysis tells us that RAG can be unfair, ours tells us where the unfairness is generated. The RRS is the concrete instrument that makes this localization actionable, and its empirical operating points (Section 5.4) define a monitoring band rather than a single decision boundary.

7.1.2 The Retriever as a Faulty Software Component

In classical software engineering, component-level testing rests on a straightforward principle: each module must be verified in isolation before integration. This paper applies that principle to the RAG retriever, yet it differs from prior work, which treated RAG pipelines as black boxes, evaluating only the final generated output.

The diagnostic reading is that, before any generation occurs, the upstream component has already produced a fairness defect that downstream stages cannot undo. This instantiates defect propagation in component-based systems: an unaddressed fault at an early stage contaminates every downstream stage. The contribution here is not the existence of the defect but its formalization at the component level for an attribute (demographic invariance) that the RAG literature had not previously isolated. The propagation taxonomy reported in Section 5.2 (Llama as passthrough, Mistral as buffer, Nemotron as amplifier) sharpens this further: the same upstream defect can yield three qualitatively different system-level behaviors depending on the generator, which is the SE definition of an integration concern rather than a per-component one.

7.1.3 Metamorphic Testing as a Test Oracle Substitute

A challenge in testing AI systems is the oracle problem: there is no automatically verifiable correct output to compare against. Metamorphic testing resolves this by checking whether the answer changes when it should not, rather than checking what the system answers.

The Set Equivalence MRT implemented in the paper is a system invariant: given that demographic information is irrelevant to a sentiment-classification task, the output must be invariant to demographic perturbations. When that invariant is violated, a fault is detected without requiring any prelabelled ground truth.

This reasoning has a direct parallel in SE practice. In security testing, one does not know the correct response to an SQL injection attempt, but one knows the system must not behave differently with and without the injection. The logical structure is identical. Metamorphic testing extends this pattern to fairness, treating demographic neutrality as a testable invariant.

7.1.4 Mitigation Assessment as a Fairness Regression Test

The two-stage structure of the study replicates a standard SE cycle: Bias Diagnosis localizes the fault, and Mitigation Assessment regression-tests the architectural change proposed as a fix. The interpretation that matters here is what the negative outcome (graph-based reranking failing to resolve the diagnosed defect) means for the RAG ecosystem rather than the magnitude of the residual bias. It replicates a well-understood SE failure mode: optimizing along one quality dimension (relevance) without instrumenting the test suite for orthogonal quality attributes (fairness). A patch that resolves a performance ticket while regressing a correctness invariant would not be considered a valid fix; graph-enhanced retrieval should be evaluated under the same standard, and currently is not.

7.1.5 The Demographic Hierarchy and the Embedding Stage

From a software architecture perspective, one finding is that the hierarchy race $>$ sexual orientation $>$ gender $>$ age persists across two different tasks, two different corpora, and two different retrieval strategies. In SE, when a behaviour persists independently of execution context, it is suggestive of a property of the component itself rather than a configuration artifact.

We read this as evidence that the fairness sensitivity is more strongly associated with the embedding-driven retrieval stage than with the downstream stages we evaluated. This has a practical implication for where mitigation effort is likely to pay off: because the signal originates upstream of generation, the reranking change tested here did not address it, and approaches that act on the embedding stage itself, such as contrastive fine-tuning, demographic-invariant encoding, or query preprocessing that strips demographic cues before encoding, are a more promising direction. We frame this as a direction rather than a demonstrated fix, since we did not experimentally manipulate the embedding model.

This principle is established in SE as the separation of concerns: a fairness monitoring layer and a reranking component must operate independently, and fixing one does not substitute for fixing the other.

7.1.6 Model-size Interaction as an Undocumented Integration Test Case

The finding that Llama-3.2-3B is the only model to exhibit degradation under GoPs ($p = 0.0005$) introduces a concept central to integration testing: emergent properties. The degradation is not predictable by inspecting the retriever or the model in isolation; it emerges from the interaction between the retrieval variation introduced by graph reranking and the smaller model's reduced capacity to absorb that variation.

This reflects an established principle in integration testing: the behaviour of the integrated system is not necessarily the sum of its components' individual behaviours. The practical implication is direct: fairness tests must be executed per specific model–pipeline combination, not per component in iso-

lation. Aggregate metrics would have masked the degradation entirely.

The model vulnerability ranking in Mitigation Assessment (Nemotron > Mistral > Llama in absolute ASR) differs from Bias Diagnosis (Nemotron > Llama > Mistral), suggesting that task type interacts with model architecture in determining fairness vulnerability. This reinforces the need for task-specific fairness evaluation, a direct analogy to the SE requirement that regression test suites cover multiple execution environments.

7.2 Comparison with Existing Fairness Testing Approaches

7.2.1 Metamorphic Testing Effectiveness

The application of the Set Equivalence MRT to a hybrid RAG pipeline carries a methodological claim about MT itself: it scales beyond the standalone-model setting where it has been most studied. Hyun et al. (2024) and Wang et al. (2023) both apply MT to single-component AI systems (a language model in isolation, a content-moderation classifier) and report that MT detects fairness-relevant violations that conventional metrics miss. The present work shows that the same testing principle survives the addition of a retrieval stage: the MR is still violated when it should be, and the violations remain interpretable at the component level. This matters because the empirical fairness research on RAG had so far reused QA-style metrics that conflate retrieval and generation (Shrestha et al., 2024); under such metrics the location of the defect is unrecoverable, even when its presence is.

7.2.2 Limitations of Current Fairness Metrics

Existing fairness evaluation frameworks focus on either individual fairness (equal treatment of similar individuals) or group fairness (statistical parity across demographic groups), but few provide integrated assessment across system components. Our RRS metric addresses this gap by quantifying fairness violations at the retrieval stage, enabling per-component diagnosis for complex AI system debugging.

Chen et al. (2024) identifies metamorphic testing as a promising approach but notes limited application to hybrid AI architectures. Our work provides evidence that metamorphic relations can be adapted to multi-component systems, providing detection capabilities and diagnostic insights about where failures occur in the pipeline.

7.3 Model-specific Vulnerability Patterns

The rank order between the three SLMs in the Bias Diagnosis stage (Mistral with a lower violation rate, Llama and Nemotron statistically indistinguishable despite a 2.7× parameter ratio; Section 5.1) shows more than "model size does not predict fairness." The propagation analysis in Section 5.2 shows that Mistral's robustness is not an absence of demographic sensitivity but an active correction of upstream bias: the buffer pattern absorbs roughly four out of every five retriever flips. Nemotron's degradation is partly generator-side (the amplifier pattern) and partly an interaction effect that aggregate metrics cannot see. The architectural reading is that fairness behavior is jointly determined by the retriever and the generator, and that "which model is more robust" is the wrong frame: the right frame is "which composition of retriever and generator behaves how under perturbation." Mistral's instruction-tuned, efficiency-focused architecture and Nemotron's long-context fine-tuning are plausible mechanistic candidates for why each behaves the way it does, but the methodological point is independent of which mechanism is correct: any explanation must be local to the (retriever, generator) pair.

7.4 Implications for Software Engineering

7.4.1 Component-Level Testing as a First-Class Fairness Requirement

The framing implication of the Bias Diagnosis stage is that fairness testing for RAG systems must probe individual pipeline components, not merely the end-to-end output. The fact that retrieval-stage violations are of the same order of magnitude as end-to-end violations means that fairness defects observable at the system output are largely seeded upstream. This defect-propagation pattern is familiar in component-based software engineering (Zhang et al., 2022). Two consequences follow. First, retrieval components must be treated as first-class citizens in fairness testing workflows, in the same sense that security engineering treats each component as independently auditable before integration. Second, the diagnostic apparatus has to live at the component being audited: the RRS plays the role of code coverage or mutation score for the retriever, with the additional advantage of being computable without GPU inference, making it a tractable CI/CD signal for resource-constrained teams (Section 6.2.3).

7.4.2 Integration Testing for Emergent Fairness Properties

A subtler finding is that end-to-end ASR exceeds retrieval ASR for certain models, meaning that some fairness violations are not attributable to the retriever alone. These violations emerge from the interaction between the retrieval context and the generation stage, a class of defects that isolated component testing cannot detect. In SE terms, these are interface faults: absent when components are tested in isolation but manifest when components are composed.

This pattern has a direct analogue in integration testing. A module that passes all unit tests may still fail when integrated, because its assumptions about the input it receives are violated by the actual output of its upstream dependency. In RAG, the generator's assumption is that the retrieved context is demographically neutral; when the retriever violates this assumption, the generator produces a fairness violation that neither component would produce alone.

The implication for QA practice is that fairness test suites must include integration-level test cases that exercise the full retrieval-to-generation path, in addition to per-component probes. These integration tests must be model-specific: the

same retrieval variation that Mistral-7B absorbs without measurable fairness degradation is sufficient to amplify bias in Llama-3.2-3B (Table 14). Aggregate, model-agnostic test suites would mask this interaction, producing false assurance of pipeline-level fairness.

Together, these two requirements, per-component probing and model-specific integration testing, define a layered fairness testing strategy for RAG systems, analogous to the unit-integration-system testing hierarchy in software quality assurance (Wohlin et al., 2012).

Traditional software testing approaches must be extended to accommodate the stochastic nature of language model components while maintaining the rigor required for quality assurance. Stochasticity does not dissolve the need for reproducible test criteria; it shifts the unit of analysis from deterministic output equivalence to distributional invariance under controlled perturbation. The Set Equivalence MRT operationalizes this shift: rather than asserting that two outputs are identical, it asserts that they belong to the same equivalence class under a semantically neutral transformation. This reframing preserves the logical structure of classical metamorphic testing while accommodating the probabilistic nature of language model inference. Our methodology provides a replicable template for this adaptation, demonstrating that established SE testing techniques require principled extension, not replacement.

7.4.3 Graph-based Reranking as a Fairness-Neutral Architectural Change

Reading the Mitigation Assessment outcome as fix verification rather than as a benchmarking exercise is what gives it its weight. The architectural change passes the relevance acceptance criterion but does not satisfy the fairness regression criterion: by the standards of functional testing, a patch that resolves a performance ticket while regressing a correctness invariant would not ship. The same standard should govern the deployment of graph-enhanced retrieval, and the gap is systemic: the GoPs, KAG, and KGFabric programs cited in Section 2 all ship with relevance evaluations and none ships with a fairness regression test suite. The consequence at deployment time is that organizations adopting graph-enhanced retrieval with smaller SLMs may introduce a regression that no current evaluation protocol would detect.

7.4.4 Neutrality as a Negative Result with Practical Value

A near-zero ΔASR is a diagnostic outcome, not an inconclusive one. In SE, negative results from fix verification constrain the solution space: if an architectural change does not resolve a documented fault, the fault's root cause lies outside the scope of that change. Our result rejects a specific hypothesis about the RAG ecosystem's current trajectory: that graph-based retrieval mechanisms might dampen demographic sensitivity by anchoring results in inter-passage relationships rather than isolated vector similarity. The intuition behind that hypothesis is not unreasonable: graph reranking introduces a structural prior that could in principle stabilize individual passage scores against query-level perturbations.

The mechanism by which the hypothesis fails is what makes the result instructive. Graph edges in GoPs are built from structural adjacency and TF-IDF keyword overlap; neither carries any signal about the demographic sensitivity that lives in the embedding space. The embedding model encodes the demographic cue into the query vector before FAISS retrieval occurs, and GoPs reranks the candidate pool that FAISS already returned. If that candidate pool is biased, the graph has no corrective signal to apply. The reranking stage is downstream of the fault origin, and a fix applied downstream of a fault source does not resolve the fault, a principle long established in software fault localization. The implication for practitioners is that relevance optimization and fairness assurance occupy orthogonal axes of the retrieval architecture, and conflating them is a category error that the per-model delta analysis (Table 12) makes concrete.

7.4.5 Model-size Interaction with Graph Reranking: An Undocumented Risk

The non-uniform fairness effect of GoPs across model capacities is invisible to aggregate evaluation. The aggregate ΔASR suggests neutrality; the per-model decomposition reveals a regression for the 3B model (Section 6.2.2). From an SE perspective this illustrates why aggregate metrics are insufficient as regression criteria: a test suite that reports only system-level ASR would certify the pipeline as fairness-neutral while shipping a regression to every deployment running the smaller model. The functional-testing analogy is a performance optimization that improves average latency while increasing tail latency for a specific hardware target, which would not ship without per-configuration regression coverage.

The mechanism is consistent with known capacity effects in LLM robustness. Smaller models have less representational redundancy with which to absorb input variation; when graph reranking alters the composition of the retrieved pool (visible in the lower Jaccard scores under GoPs, Table 11), a 3B-parameter model has fewer internal resources to maintain output stability than a 7B or 8B model. The degradation is therefore not a property of GoPs in isolation but of the composition of graph reranking with a low-capacity generator: a defect class that only manifests at the integration boundary, and one the current literature on graph-enhanced retrieval does not document.

7.5 Practical Deployment Considerations for Flat RAG-based Systems

For organizations deploying RAG-based systems in production, our findings suggest several recommendations:

- **Risk-Based Testing Strategy:** Prioritize testing with race-related perturbations, as these are associated with higher rates of fairness violations across the evaluated models and system components;
- **Retrieval System Auditing:** Implement continuous monitoring of retrieval behavior using metrics like RRS to detect fairness regressions before they propagate to end users;

- **Model Selection Criteria:** Consider fairness robustness alongside traditional performance metrics when selecting SLMs for deployment. Model architecture and training methodology may be more predictive of fairness behavior than model size;
- **Component-Level Mitigation:** Develop bias mitigation strategies that target both retrieval and generation stages, recognizing that end-to-end approaches may be insufficient for addressing compound effects of multi-stage bias propagation.

These recommendations provide guidance for software engineering practitioners working with RAG systems, translating research findings into deployment strategies that can improve system fairness without sacrificing functionality.

7.6 Integrated Implications for Practitioners

Combining the findings from both studies, we extend the recommendations from Section 7.5 with guidance specific to graph-enhanced RAG:

- **Graph Reranking Is Not a Fairness Strategy:** Organizations should not expect graph-based retrieval enhancements to address demographic bias. GoPs, and by extension similar approaches, optimize for relevance and coherence, not demographic invariance. Fairness testing remains necessary after adopting graph-enhanced pipelines;
- **Mitigate at the Embedding Layer:** Since the root cause of bias lies in the embedding model, mitigation strategies should target this layer directly, for instance through demographic-invariant fine-tuning, contrastive debiasing, or query preprocessing that strips demographic cues before encoding;
- **Extra Caution with Small Models:** Graph reranking can amplify fairness vulnerabilities in smaller models, as documented for the 3B-parameter generator in Section 6.1.4. Organizations using SLMs under 8B parameters should perform additional fairness testing when introducing graph-based retrieval components, rather than assuming that aggregate-level fairness neutrality transfers to their specific deployment;
- **Monitor Retrieval Stability, Not Just Accuracy:** The Jaccard similarity metric provides a model-independent indicator of retrieval fairness. A Jaccard below 0.30 under demographic perturbation signals high retrieval instability and should trigger further investigation, regardless of whether downstream generation metrics appear acceptable.

8 Threats to Validity

Following established guidelines for empirical software engineering research (Wohlin et al., 2012), we identify and discuss potential threats to the validity of our findings across the four standard dimensions, with particular attention to threats raised by reviewers of the conference version. For each threat we describe the underlying mechanism, the mitigation steps embedded in our experimental design, and how the principal claims should be read in light of the threat.

8.1 Construct Validity

Measurement of Fairness via Set Equivalence MRT. The Set Equivalence MRT operationalizes fairness as invariance of system output under demographic perturbation, an instance of individual fairness in the metamorphic-testing tradition. The mechanism by which this could distort findings is that several other fairness conceptions (group fairness, equality of opportunity, intersectional bias, calibration parity) are not measured. *Mitigation:* the study is framed as a fault-localization exercise rather than a fairness audit, and the SE concept invoked throughout the paper, defect propagation under a metamorphic relation, does not require a particular fairness conception, only a violated invariant. *Reading:* the ASRs we report are lower bounds on demographic sensitivity, not estimates of total fairness harm, and the localization claim (the retriever as the principal site of the defect) holds independently of which fairness conception is privileged downstream.

Assessed NLP Tasks. The mechanism of concern is task-specificity: a fairness fault observed in sentiment analysis and open-domain Q&A might not transfer to summarization, structured information extraction, or generation tasks where demographic context is itself part of the prompt. *Mitigation:* the two-stage design (Bias Diagnosis on Toxic Conversations, Mitigation Assessment on Wikipedia Q&A) was deliberately structured as a cross-environment regression test (Section 4.3), so that the consistency of findings across two corpora and two task types provides empirical evidence against task-specificity. *Reading:* the demographic hierarchy and the embedding-layer locus generalize across the two evaluated tasks, but extrapolation to different task families requires re-evaluation under appropriate MRs.

Demographic Perturbation Coverage. The 21 perturbation terms across four categories (race, sexual orientation, gender, age) do not cover all relevant demographic axes. Socioeconomic status, disability status, religion, nationality, and immigration status are omissions. The mechanism by which this could mislead is selection effects: a hierarchy derived from four categories might invert under additional axes. *Mitigation:* term selection follows the METAL framework (Hyun et al., 2024), which itself derives from prior fairness-testing literature, providing terminological consistency rather than convenience-sampling. *Reading:* the rank order race $>$ orientation $>$ gender $>$ age holds within the evaluated coverage, and additional axes should be expected to integrate into this ordering rather than overturn it; verifying that integration is empirical work this paper does not perform.

STS Threshold Calibration. The violation threshold $\tau_{\text{STS}} = 0.85$ was calibrated on Llama-3.2-3B using a 20-query subset prior to the full Mitigation Assessment run. The mechanism of concern is overfitting of the threshold to one model, which could distort absolute ASR magnitudes for the other two models. *Mitigation:* the principal claim of Mitigation Assessment is comparative (ΔASR between flat FAISS and GoPs), and both pipelines are evaluated under the same

threshold for the same model, so any threshold-induced bias affects both arms of the comparison symmetrically. *Reading:* absolute ASR levels in the Mitigation Assessment results table should be interpreted with the calibration provenance in mind, but the per-model regression evidence (Section 6.1.4) is robust to this threat.

RRS Transferability Across Models. The aggregate $d_{th} = 1.3089$ threshold for the Retriever Robustness Score is calibrated on the union of the three models' flip distributions and operates as a population-level monitoring cutoff. The Mann-Whitney and ROC analyses in Section 5.4 show that the metric carries genuine discriminative signal for Llama-3.2-3B (AUC $= 0.557$) and Mistral-7B (AUC $= 0.565$), but is inverted for Nemotron-8B (AUC $= 0.444, r_{rb} = -0.112$): for Nemotron, pairs that flip have lower median RRS than pairs that do not. *Mechanism:* the aggregate threshold is therefore not a model-agnostic decision boundary, and applying it as such to Nemotron would invert the operational decision. *Mitigation:* the propagation analysis in Section 5.2 provides the explanatory mechanism (Nemotron's flips include a generator-side amplifier component decoupled from retrieval-level semantic drift, which a retrieval-side metric cannot fully discriminate). *Reading:* RRS-based monitoring should be calibrated per model, and the categorical bands in Table 4 should be interpreted as conservative upper-quartile risk flags rather than calibrated single-pair decision boundaries.

Metric Change Across Experiments. Bias Diagnosis uses discrete label equivalence, whereas Mitigation Assessment uses continuous STS similarity. The mechanism of concern is that the operational definition of "violation" differs between stages, which complicates direct cross-stage comparison of magnitudes. *Mitigation:* we do not compare ASRs across stages as if they were the same measurement; the cross-study claim is qualitative (the demographic hierarchy persists, race-related perturbations remain dominant) rather than quantitative. *Reading:* the persistence of the qualitative pattern under two different operationalizations strengthens, rather than weakens, the structural-property interpretation of the embedding hierarchy.

8.2 Internal Validity

Confounding Variables in Model Comparison. The three SLMs differ along several axes simultaneously: parameter count (3B/7B/8B), architecture family (Llama, Mistral, Nemotron), training corpus, and instruction-tuning regime. The mechanism is the standard observational-study confound: an effect attributed to "the model" could in principle be attributed to any of these axes. *Mitigation:* we draw no causal claims about which architectural attribute is responsible for which behavior; the study reports descriptive patterns (passthrough/buffer/amplifier) and uses them to argue an SE point about heterogeneity that does not require ascribing causes. *Reading:* claims of the form "Mistral is more robust because of architecture X" are explicitly out of scope; claims of the form "the (retriever, generator) composition exhibits heterogeneous propagation patterns whose ranking is not predicted by parameter count" are supported.

Graph Construction Approximation. GoPs as proposed by Li et al. (2025) extracts keywords by prompting an LLM, which is computationally infeasible for a 129,389-chunk corpus at our scale. We substitute TF-IDF cosine similarity for keyword overlap (Section 4.3). The mechanism of concern is that TF-IDF keywords are surface-lexical while LLM-extracted keywords are semantic, producing topologically different graphs. *Mitigation:* the substitution is documented and the calibration of $\tau_{\text{edge}} = 0.15$ aimed at preserving the connectivity profile (average degree ≈ 315, density $\approx 0.24\%$) reported in the GoPs literature, so that the structural prior introduced by the graph remains comparable in magnitude. *Reading:* our results characterize TF-IDF GoPs and the negative aggregate result is interpreted as evidence about the class of unsupervised, lexically-grounded graph reranking strategies, not necessarily about LLM-keyword variants.

Single Graph Configuration. Mitigation Assessment evaluates one GoPs configuration ($\alpha = 0.5$, $\tau_{\text{edge}} = 0.15$, $n = 20$ candidates). The mechanism of concern is that a different configuration could change the magnitude or even the sign of the fairness effect. *Mitigation:* the configuration corresponds to the defaults reported by the GoPs authors (Li et al., 2025), ensuring that we test the published recommendation rather than an arbitrarily picked operating point. *Reading:* claims about graph-based reranking in this paper are claims about its default-configured form; tuning α or n specifically for fairness-aware retrieval is a separate research question that our results do not answer.

Duplicated Question. The METAL pool of 100 questions contained one inadvertent duplicate. Mitigation Assessment uses the 99 unique questions, which yield the 2,079 (query, term) pairs per model–pipeline cell and 12,474 total pairs reported throughout. The mechanism of concern is whether the near-duplicate, had it been retained, would have contaminated the paired statistical tests through sample replication. *Mitigation:* de-duplication was applied before the full run, and recomputation that retains the duplicate changes ASR estimates by less than 0.4%. *Reading:* the impact on every reported statistic is negligible relative to the effect sizes claimed.

8.3 External Validity

Scope Limitations. We state the main limitations on the generality of our findings directly. First, the study evaluates a single embedding model (all-MiniLM-L6-v2) and does not compare multiple embedding models, so we cannot claim that the observed behavior is representative of sentence encoders in general. Second, the behavior we report may differ under other dense retrievers, and we did not test alternative retrieval backbones. Third, our graph results rely on a TF-IDF cosine-similarity proxy for edge construction, which may not capture the full semantics of the LLM-extracted keyword graph in the original GoPs formulation; a different

edge-construction strategy could behave differently. Fourth, all models evaluated are open-weight SLMs in the 3B–8B range, and the findings do not necessarily generalize to proprietary frontier LLMs. The remainder of this subsection examines several of these threats in more detail.

Choice of SLMs Rather than Frontier-Scale LLMs. We use three SLMs in the 3B–8B parameter range. This limits external validity in two ways. The construct concern is that some of the absolute ASR magnitudes could be a capacity effect rather than a property of RAG retrieval, since smaller models have less representational redundancy to absorb input variation. The external concern is that the SE community deploying RAG includes both frontier proprietary models (GPT-4, Claude, Gemini) and open-weight SLMs, and a study restricted to the latter cannot speak to the former.

The choice was deliberate for three reasons.

- *Operational relevance.* The SLM regime is where the cost-fairness tradeoff is sharpest. SMEs and resource-constrained teams adopt 3B–8B open-weight models for RAG because frontier APIs are economically out of reach. These deployments are the least likely to receive fairness audit infrastructure.
- *Localization is separable from generator scale.* The association we observe, that the fairness sensitivity is more strongly linked to the embedding-driven retrieval stage than to the generator, does not depend on generator size. The embedding model (all-MiniLM-L6-v2) is identical across the three SLMs, and the retrieval-stage measurements that support this reading are taken before any generation occurs (Section 5.3). The retrieval-stage ASR and the demographic hierarchy are properties of the embedding component, and we would expect them to transfer to pipelines that pair the same embedding family with frontier LLMs, though we did not test this.
- *Heterogeneity already visible.* Within the 3B–8B range, the propagation taxonomy of Section 5.2 (Llama passthrough, Mistral buffer, Nemotron amplifier) and the inverted ROC for Nemotron show that "larger model is more robust" is false even within the SLM band. There is no theoretical basis to expect frontier LLMs to exhibit a single uniform behavior on the same MR; what should be expected is a broader taxonomy of generator-side patterns, not the disappearance of the retriever-side fault.

The reading of the findings under this threat: (i) the association between the fairness signal and the embedding-driven retrieval stage may extend to RAG pipelines that pair the same embedding family with frontier LLMs, but we did not test this; (ii) absolute end-to-end ASR magnitudes are estimates for SLM-deployed pipelines and may be upper bounds for frontier-LLM pipelines, while the demographic hierarchy and the retriever-side defect-propagation pattern appear less dependent on the magnitudes; (iii) the SLM-specific finding under graph reranking is sensitive to model scale, and replication on frontier LLMs is a natural extension. That replication is computationally out of scope here and is flagged as a priority direction in Section 9.

Language and Cultural Context. The study is limited to English-language content with demographic categories rooted in North American discourse. The mechanism of concern is that demographic salience is culturally specific: the hierarchy that emerges in English embeddings may not map onto embeddings trained on other languages or onto categories that carry different sociopolitical weight in non-English contexts. *Mitigation:* scope is restricted explicitly rather than implicitly, and METAL's term taxonomy is documented for replicability in other languages. *Reading:* the structural claim (the embedding layer carries a demographic hierarchy) is expected to generalize; the specific rank order is an empirical finding for English embeddings and should be re-derived for other languages.

Generalizability of Graph-based Findings. Our findings about graph reranking should be interpreted as evidence about lightweight, unsupervised graph construction (TF-IDF edges, default α, no learned weighting). The mechanism of concern is that supervised, learned, or LLM-extracted graph variants might exhibit qualitatively different fairness behavior. *Mitigation:* Section 6.1.1 is explicit that the result generalizes to the class of graph reranking strategies that share the property of being downstream of the embedding layer, not necessarily to the broader graph-RAG paradigm. *Reading:* a learned graph reranker that incorporates fairness signals in its training objective is a different system, and its evaluation is a separate empirical question.

8.4 Conclusion Validity

Statistical Power and Sample Size. Bias Diagnosis evaluates 2,100 pairs per model and Mitigation Assessment evaluates 12,474 pairs total. The mechanism of concern is that a finding could be a false positive driven by undetected dependencies among observations (multiple terms per category, multiple categories per question) or a false negative driven by inadequate sample size. *Mitigation:* the formal tests we report (Cochran's Q with McNemar follow-ups, Wilson confidence intervals, Cohen's h) are paired or marginal tests appropriate for within-subject designs; sample sizes are large enough that effects of practical interest yield p-values orders of magnitude below conventional thresholds (e.g., $p \approx 10^{-32}$ for the omnibus model-difference test in Section 5.1), reducing the false-negative concern. *Reading:* the statistical claims of the paper are calibrated to the within-pair, within-model design, and any extrapolation to a between-pair or between-model meta-analysis would require additional inference.

Statistical Test Granularity. The Wilcoxon signed-rank tests in Section 6.1.4 operate on continuous STS scores rather than on the binary ASR violation flag. The mechanism of concern is that the test could be sensitive to STS shifts that fall short of crossing the violation threshold and therefore do not affect ASR. *Mitigation:* the complementary ΔASR analysis (Table 12) is computed on the binary metric and provides an effect-size view aligned with the practitioner-facing definition of a violation. *Reading:* when the Wilcoxon and the

ΔASR analyses agree (as they do for the 3B model, where both indicate degradation under GoPs), the conclusion is robust; when they would disagree, the binary ΔASR is the operationally meaningful metric.

9 Conclusion and Future Work

Summary of Contributions. This paper advances fairness testing for AI-enabled systems by treating RAG pipelines as a software testing problem rather than a model evaluation problem. The shift in framing determines which components are tested, which metrics are defined, and which architectural conclusions follow from the data.

Our two-stage empirical study comprises Bias Diagnosis followed by Mitigation Assessment. Bias Diagnosis applies 21 demographic metamorphic relations across four identity categories to three SLMs, localizing the retriever as a fault source (retrieval ASR $= 29.14\%$) and introducing the RRS as a component-level fairness probe. Mitigation Assessment regression-tests graph-based reranking via GoPs and returns a negative verification: GoPs is fairness-neutral at the aggregate level ($\Delta\text{ASR} = +0.0008$), fails to resolve any of the three documented failure modes, and introduces a fairness regression in the 3B model ($p = 0.0005$, Llama-3.2-3B).

What is new relative to the conference version. This manuscript extends Oliveira et al. (2025) along three dimensions, summarized in Table 1. The conference version contributed the metamorphic-testing framework for RAG fairness, the RRS metric, and a descriptive Bias Diagnosis on the Toxic Conversations corpus. The present version adds:

(a) A new empirical stage (Mitigation Assessment) that performs fairness regression testing of graph-based reranking with GoPs against flat FAISS over a Wikipedia corpus, scaling the test design from 2,100 to 14,574 evaluated pairs.
(b) A statistical layer over the Bias Diagnosis stage (Wilson 95% CIs, Cochran's Q with pairwise McNemar and Cohen's h, retriever-to-generator ΔASR with paired McNemar, χ^2 category homogeneity, pairwise z-tests, Mann-Whitney/ROC validation of the RRS, per-term ranking, and a pair-level failure-mode taxonomy) that replaces the conference version's descriptive treatment.
(c) A software-engineering reframing in which Bias Diagnosis is component-level testing and Mitigation Assessment is regression testing, accompanied by the three-layer architectural pattern (retrieval / fairness monitoring / embedding), a Jaccard-based CI/CD quality gate, and per-model deployment recommendations.

The propagation analysis (Llama as passthrough, Mistral as buffer, Nemotron as amplifier) and the negative fix-verification result for graph-enhanced retrieval are also original to this version.

Findings and Their Engineering Significance. Four findings carry direct implications for SE practice:

(i) *Retriever-level bias is a first-class fault.* Bias propagation begins before generation. The retrieval ASR (29.14%) approaches the end-to-end ASR (17.95%–33.00%); fairness defects seeded at the retrieval stage account for the majority of system-level violations. Quality assurance frameworks that test only final outputs will miss the fault origin.
(ii) *The demographic hierarchy is a structural property of the embedding space.* The ordering race $>$ sexual orientation $>$ gender $>$ age persists across two tasks, two corpora, and two retrieval strategies. Its stability indicates that the root cause resides in the embedding model, not in dataset properties or retrieval configuration. No downstream architectural change can substitute for embedding-level mitigation.
(iii) *Graph-enhanced retrieval is not a fairness fix.* GoPs, KAG, and KGFabric are entering enterprise production with documented relevance gains and no fairness regression test suites. Relevance optimization and fairness assurance operate on orthogonal dimensions of the retrieval architecture. Conflating them is a category error with measurable consequences.
(iv) *Model capacity modulates graph-reranking risk.* The degradation of Llama-3.2-3B under GoPs is invisible to aggregate evaluation and undetected by existing protocols for graph-enhanced retrieval. It constitutes an undocumented risk for organizations that pair resource-constrained SLMs with graph-based pipelines.

Practical Artifacts for Practitioners. Beyond findings, this work delivers three artifacts that practitioners can adopt directly. The RRS provides a GPU-free, component-level fairness probe integrable into CI/CD pipelines as a quality gate. The Jaccard-based threshold scheme (green/yellow/red) operationalizes retrieval stability monitoring without requiring end-to-end model inference. The three-layer architectural pattern (retrieval layer for relevance, fairness monitoring layer for demographic invariance, embedding layer for root-cause mitigation) provides a separation-of-concerns blueprint for RAG systems that must satisfy both accuracy and fairness requirements.

Broader Implications for Software Architecture. Our findings reposition fairness as a cross-cutting architectural concern, analogous to security or observability. Evidence suggests that fairness assurance in RAG systems is seldom effective when restricted to a single component or a single metric, necessitating a multi-layered approach akin to security protocols. Architects must account for demographic invariance at the embedding layer, monitor it at the retrieval layer, and verify it through model-specific integration tests at the generation layer. This three-layer decomposition is grounded in the fault localization evidence from Bias Diagnosis and the negative fix verification from Mitigation Assessment.

Limitations and Future Work. This study has boundaries that future work should address. The evaluation is limited to English-language content and North American demographic categories; extending the framework to multilingual corpora and culturally distinct demographic taxonomies is a neces-

sary step toward generalization. The three SLMs evaluated share a parameter range (3B–8B); replication with larger models (>10B parameters) and proprietary systems would establish whether the capacity effects documented here hold at scale. The GoPs configuration tested represents one instantiation of graph-enhanced retrieval; evaluating KAG and GNN-based rerankers under the same fairness MR framework would determine whether the fairness-neutrality result generalizes to the broader family of graph-based methods.

On the methodological side, the Set Equivalence MRT operationalizes individual fairness but does not capture group fairness or intersectional bias. Complementing our approach with group-level metrics and multi-attribute perturbations would provide a more complete fairness profile. The STS threshold ($\tau = 0.85$) was calibrated on a single model; a calibration study across multiple models and task types would establish more transferable thresholds.

From an industrial adoption perspective, the engineering effort required to integrate RRS-based quality gates into existing CI/CD pipelines remains unquantified. A field study with software teams deploying RAG in production would assess feasibility and produce the cost-benefit evidence needed for adoption decisions.

This work identifies the embedding layer as the mitigation point but does not evaluate mitigation strategies. Future work should compare debiased embeddings, contrastive fine-tuning, and demographic-invariant query preprocessing under the same MR framework, closing the loop from fault localization to verified remediation.

Our work shifts the lens of fairness testing from the model to the system. The methodological contribution is a reusable testing framework grounded in SE principles; the empirical contribution is evidence that fairness defects in RAG systems can originate in individual components and are best diagnosed there. Both contributions point toward the same conclusion: fairness assurance in hybrid AI systems is a software engineering problem, and it demands software engineering answers.

Artifact Availability

All the necessary data to replicate this work is available in our public repository[12]. A `README.md` file with replication instructions is also provided.

Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brasil, Finance Code 001, by the Fundação de Apoio ao Desenvolvimento do Ensino, Ciência e Tecnologia do Estado de Mato Grosso do Sul (FUNDECT, Call No. 42/2024), and by the Federal University of Mato Grosso do Sul (UFMS).

[12] `https://github.com/Neural-Matheus/Metamorphic-SLM`

References

Amershi, S., Begel, A., Bird, C., DeLine, R., Gall, H., Kamar, E., Nagappan, N., Nushi, B., and 2019, T. Z. (2019). Software engineering for machine learning: A case study.

Baeza-Yates, R. (2018). Bias on the web.

Basili, V. R., Caldiera, G., and 1994, H. D. R. (1994). The goal question metric approach.

Bender, E. M., Gebru, T., McMillan-Major, A., and 2021, S. S. (2021). On the dangers of stochastic parrots: Can language models be too big?

Cantini, R., Orsino, A., Ruggiero, M., and Talia, D. (2025). Benchmarking adversarial robustness to bias elicitation in large language models: Scalable automated assessment with llm-as-a-judge. *Machine Learning*, 114(11):249.

Chen, J., Dong, H., Wang, X., Feng, F., Wang, M., and 2023, X. H. (2023). Bias and fairness in information retrieval systems.

Chen, T. Y., Kuo, F.-C., Liu, H., Poon, P.-L., Towey, D., Tse, T. H., and 2018, Z. Q. Z. (2018). Metamorphic testing: A review of challenges and opportunities.

Chen, Z., Zhang, J. M., Hort, M., Harman, M., and 2024, F. S. (2024). Fairness testing: A comprehensive survey and analysis of trends.

Cruz, A. F., Saleiro, P., Belém, C., Soares, C., and Bizarro, P. (2021). Promoting fairness through hyperparameter optimization. In *2021 IEEE international conference on data mining (ICDM)*, pages 1036–1041. IEEE.

Es, S., James, J., Anke, L. E., and Schockaert, S. (2024). Ragas: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th conference of the european chapter of the association for computational linguistics: system demonstrations*, pages 150–158.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., and 2023, H. W. (2023). A comprehensive survey of retrieval-augmented generation (rag): Evolution, current landscape and future directions.

Giramata, S., Srinivasan, M., Gudivada, V. N., and Kanewala, U. (2025). Efficient fairness testing in large language models: Prioritizing metamorphic relations for bias detection. In *2025 IEEE International Conference on Artificial Intelligence Testing (AITest)*, pages 191–200. IEEE.

Hu, M., Wu, H., Guan, Z., Zhu, R., Guo, D., Qi, D., and 2024, S. L. (2024). No free lunch: Retrieval-augmented generation undermines fairness in llms, even for vigilant users.

Hyun, S., Guo, M., and 2024, M. A. B. (2024). METAL: Metamorphic testing framework for analyzing large-language model qualities.

Ji, Z., Yu, T., Xu, Y., Lee, N., Ishii, E., and 2023, P. F. (2023). Towards mitigating LLM hallucination via self reflection.

Karpukhin, V., Oguz, B., Min, S., Lewis, P., Wu, L., Edunov, S., Chen, D., and tau Yih. 2020, W. (2020). Dense passage retrieval for open-domain question answering.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., tau Yih, W., Rocktäschel, T., and et al. 2020 (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks.

Li, M., Zhang, Z., Cao, T., Liu, F., Zhang, C., and 2024, K. M.

(2024). Trustworthiness in retrieval-augmented generation systems: A survey.

Li, Y., Du, M., Song, R., Wang, X., and 2023, Y. W. (2023). A survey on fairness in large language models.

Li, Z., Guo, Q., Shao, J., Song, L., Bian, J., Zhang, J., and Wang, R. (2025). Graph neural network enhanced retrieval for question answering of large language models.

Liang, L., Bo, Z., Gui, Z., Zhu, Z., Zhong, L., Zhao, P., Sun, M., Zhang, Z., Zhou, J., Chen, W., et al. (2025). Kag: Boosting llms in professional domains via knowledge augmented generation. In *Companion Proceedings of the ACM on Web Conference 2025*, pages 334–343.

Liu, Y., Yao, Y., Ton, J.-F., Zhang, X., Guo, R., Cheng, H., Klochkov, Y., Taufiq, M. F., and 2023, H. L. (2023). Trustworthy llms: A survey and guideline for evaluating large language models' alignment.

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., and Mian, A. (2025). A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, 16(5):1–72.

Oliveira, M., Silva, J., and Fontão, A. (2025). Fairness testing in retrieval-augmented generation: How small perturbations reveal bias in small language models. In *Simpósio Brasileiro de Qualidade de Software (SBQS)*, pages 595–605. SBC.

Reddy, H., Srinivasan, M., and Kanewala, U. (2025). Metamorphic testing for fairness evaluation in large language models: Identifying intersectional bias in llama and gpt. In *2025 IEEE/ACIS 23rd International Conference on Software Engineering Research, Management and Applications (SERA)*, pages 239–246. IEEE.

Shrestha, R., Zou, Y., Chen, Q., Li, Z., Xie, Y., and Deng, S. (2024). Fairrag: Fair human generation via fair retrieval augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11996–12005.

Tukey, J. W. et al. (1977). *Exploratory data analysis*, volume 2. Addison-wesley Reading, MA.

Van Nguyen, C., Shen, X., Aponte, R., Xia, Y., Basu, S., Hu, Z., Chen, J., Parmar, M., Kunapuli, S., Barrow, J., et al. (2025). A survey on small language models. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing-Natural Language Processing in the Generative AI Era*, pages 807–821.

Wang, C., Wan, Z., Kang, H., Chen, E., Xie, Z., Krishna, T., Janapa Reddi, V., and Du, Y. (2026). Slm-mux: Orchestrating small language models for reasoning. In *International Conference on Learning Representations*, volume 2026, pages 73697–73723.

Wang, W., Huang, J.-t., Wu, W., Zhang, J., Huang, Y., Li, S., He, P., and Lyu, M. R. (2023). Mttm: Metamorphic testing for textual content moderation software. In *2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE)*, pages 2387–2399. IEEE.

Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., and 2012, A. W. (2012). *Experimentation in Software Engineering*.

Xiong, L., Xiong, C., Li, Y., Tang, K.-F., Liu, J., Bennett, P., Ahmed, J., and 2020, A. O. (2020). Approximate nearest neighbor negative contrastive learning for dense text retrieval.

Yi, P., Liang, L., Zhang, D., Chen, Y., Zhu, J., Liu, X., Tang, K., Chen, J., Lin, H., Qiu, L., et al. (2024). Kgfabric: A scalable knowledge graph warehouse for enterprise data interconnection. *Proceedings of the VLDB Endowment*, 17(12):3841–3854.

Yu, Yue, Xiong, Wei, Li, Zihan, Kong, Lingpeng, Wang, Qi, Chen, Qinyuan, Wang, and Yang, W. (2024). Rankrag: Unifying context ranking with retrieval-augmented generation in llms.

Zhang, J. M., Harman, M., Ma, L., and 2022, Y. L. (2022). Machine learning testing: Survey, landscapes and horizons.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., and et al. 2023 (2023). A survey of large language models.

Zheng, Z., Ren, D., Liu, H., Chen, T. Y., and Li, T. (2025). Identifying the failure-revealing test cases in metamorphic testing: A statistical approach. *ACM Transactions on Software Engineering and Methodology*, 34(2):1–26.

A Mathematical Formulations

This appendix gathers the formal definitions of all metrics, metamorphic relations, and graph operators used throughout the paper. The main text refers to each equation by its label (e.g. Eq. 9); the prose descriptions in Sections 4.2.4, 4.3.5, 5.4, and 4.3.2 are sufficient for reading the paper, while this appendix provides the formal specification needed for replication and for verifying our claims.

A.1 Bias Diagnosis Metrics

Set Equivalence Metamorphic Relation. The relation underlying our fairness evaluation requires that the SLM produce equivalent outputs for inputs with identical task content but different demographic profile descriptions. Formally, the relation is satisfied when

$$\forall o \in O : M(\text{input}, \text{prompt}) = o, \tag{1}$$

where O denotes the set of outputs produced by the SLM M from inputs that share content but vary in demographic descriptor.

Attack Success Rate (Bias Diagnosis). The ASR is the fraction of demographic perturbations for which the relation in Eq. 1 is violated:

$$\text{ASR} = \frac{\text{Number of different outputs}}{\text{Total number of demographic variations tested}}. \tag{2}$$

The metric is applied at two stages of the pipeline: at the retriever stage (toxicity-count MR over the top-k retrieved passages, used in Section 5.3), and at the end-to-end stage (Set Equivalence MR over the SLM's final sentiment label, used in Section 5.5).

Retriever Robustness Score. The RRS is constructed in seven sequential steps over the top-k retrieved passages of the original and perturbed queries.

1. Embedding normalization to the unit sphere:

$$\hat{\mathbf{e}} = \frac{\mathbf{e}}{\|\mathbf{e}\|_2}, \qquad \mathbf{e} \in \mathbb{R}^d. \tag{3}$$

2. Distances between two normalized passage embeddings $\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2 \in \mathbb{R}^d$ (where $\hat{\mathbf{e}}_1$ is the normalized embedding of a passage retrieved for the *original* query and $\hat{\mathbf{e}}_2$ that of a passage retrieved for the *perturbed* query), computed under the Euclidean and cosine variants:

$$d_E(\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2) = \sqrt{2 - 2\,\hat{\mathbf{e}}_1^\top \hat{\mathbf{e}}_2}, \qquad d_C(\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2) = 1 - \hat{\mathbf{e}}_1^\top \hat{\mathbf{e}}_2. \tag{4}$$

3. Full $k \times k$ pairwise distance matrix between the original and perturbed retrieved-passage embeddings:

$$M_{ij} = d(\hat{\mathbf{e}}_{1,i}, \hat{\mathbf{e}}_{2,j}), \qquad i, j \in \{0, \ldots, k-1\}. \tag{5}$$

4. Selection of the k nearest pairs under a bipartite (one-to-one) matching constraint:

$$S = \arg \min_{\substack{S' \subseteq \{0,\ldots,k-1\}^2, \\ |S'|=k,\ \text{unique rows/cols}}} \sum_{(i,j)\in S'} M_{ij}. \tag{6}$$

   Here $\{0, \ldots, k-1\}^2$ denotes the set of ordered index pairs (i, j), where i indexes an original-query passage and j a perturbed-query passage; the constraint "unique rows/cols" enforces a one-to-one matching between the two retrieved sets.

5. Mean of the k matched distances (the semantic-drift component):

$$\text{MeanDist} = \frac{1}{k} \sum_{(i,j)\in S} M_{ij}. \tag{7}$$

6. Normalized Hamming distance over the toxicity labels of the matched passage pairs (the label-drift component), where l_i is the toxicity label of the original-query passage i and l'_j that of its matched perturbed-query passage j under the matching S of Eq. 6:

$$H = \frac{1}{k} \sum_{(i,j)\in S} \mathbf{1}[\, l_i \neq l'_j \,]. \tag{8}$$

7. The final Retriever Robustness Score:

$$\text{Score} = \text{MeanDist} + H. \tag{9}$$

The components have theoretical ranges MeanDist $\in [0, 2]$ (under either distance variant), $H \in [0, 1]$, and therefore Score $\in [0, 3]$. A Score of 0 denotes a fully robust retriever (neither semantic nor label drift under the demographic perturbation), whereas values approaching the maximum of 3 denote maximal demographic sensitivity; the categorical risk bands used in practice are reported in Table 4.

RRS Decision Boundary (RQ1-2). The empirical threshold for classifying retrieval pairs as robust vs. degraded is taken as the third quartile of the MeanDist distribution restricted to the no-flip group:

$$d_{th} = Q_3(\text{MeanDist}_{\text{no-flip}}) = 1.3089. \tag{10}$$

Pairs with MeanDist $\leq d_{th}$ are predicted as robust (no polarity change), and pairs with MeanDist $> d_{th}$ as degraded (polarity inversion).

A.2 Mitigation Assessment Metrics

Jaccard Similarity (retrieval-level invariance). The overlap between the top-k passage IDs retrieved for the original and perturbed queries is measured by the Jaccard index:

$$J(q, q') = \frac{|R(q) \cap R(q')|}{|R(q) \cup R(q')|}, \tag{11}$$

where $R(\cdot)$ denotes the set of top-k retrieved passage IDs. The metric satisfies $J = 1$ for complete retrieval stability and $J = 0$ for completely disjoint retrieved sets.

Semantic Textual Similarity (generation-level invariance). Following the Distance MRT of the METAL framework (Hyun et al., 2024), let $\phi : \mathcal{S} \to \mathbb{R}^{384}$ denote the embedding produced by all-MiniLM-L6-v2, and let $a_q = \mathcal{G}(q, R(q))$ and $a_{q'} = \mathcal{G}(q', R(q'))$ denote the generator's answers for the original and perturbed queries under retrieval function $R \in \{R_{\text{base}}, R_{\text{GoPs}}\}$. The semantic textual similarity between the two answers is the cosine similarity of their embeddings:

$$\text{STS}(q, q') = \frac{\phi(a_q) \cdot \phi(a_{q'})}{\|\phi(a_q)\|\, \|\phi(a_{q'})\|}. \tag{12}$$

The score lies in $[-1, 1]$; in practice all observed values fall in $[0, 1]$ because the answers share the same topic.

Attack Success Rate (Mitigation Assessment). A metamorphic relation is violated when $\text{STS}(q, q') < \tau$, with $\tau = 0.85$ as the calibrated threshold (Section 4.3.5). The ASR is then the fraction of pairs falling below the threshold:

$$\text{ASR} = \frac{|\{(q, q') : \text{STS}(q, q') < \tau\}|}{|\text{total pairs}|}. \tag{13}$$

Delta ASR. To isolate the fairness effect of the GoPs reranking stage at the (model, demographic-category) granularity, we define

$$\Delta\text{ASR}_{m,c} = \text{ASR}^{\text{GoPs}}_{m,c} - \text{ASR}^{\text{base}}_{m,c}, \tag{14}$$

where m indexes the model and c the demographic category. Negative values indicate that GoPs reduces bias for that cell; positive values indicate amplification.

Null Hypothesis for Fairness Neutrality. Under the assumption that graph-based reranking is fairness-neutral, the per-category deltas are identically zero:

$$H_0 : \Delta\text{ASR}_{m,c} = 0 \quad \forall c \in \{\text{gender}, \text{age}, \text{race}, \text{orientation}\}. \tag{15}$$

We test H_0 for each model with a Wilcoxon signed-rank test on the paired STS scores across all (query, term) combinations.

A.3 Graph Construction and Retrieval Pipelines

Baseline Pipeline. The baseline retrieval function returns the top-k passages from the FAISS index over the corpus C:

$$R_{\text{base}}(q) = \text{top}_k(\text{FAISS}(q, C)). \tag{16}$$

Structural Edges. The structural edges of the GoPs graph connect consecutive paragraphs within the same source article:

$$E_{\text{struct}} = \big\{ (v_i, v_{i+1}) : \text{doc}(v_i) = \text{doc}(v_{i+1}) \\ \wedge\ \text{pos}(v_{i+1}) = \text{pos}(v_i) + 1 \big\}, \tag{17}$$

where $\text{doc}(v)$ and $\text{pos}(v)$ denote, respectively, the source article and the positional index of chunk v within that article.

Keyword Edges. The keyword edges connect chunks whose TF-IDF cosine similarity exceeds a calibration threshold τ_{edge}. Let $\mathbf{t}_v \in \mathbb{R}^{5,000}$ denote the TF-IDF vector of chunk v:

$$E_{\text{kw}} = \Big\{ (v_i, v_j) : \frac{\mathbf{t}_{v_i} \cdot \mathbf{t}_{v_j}}{\|\mathbf{t}_{v_i}\| \, \|\mathbf{t}_{v_j}\|} > \tau_{\text{edge}}, \\ i \neq j \Big\}. \tag{18}$$

GoPs Integrated Distance Score. For each FAISS candidate c, let $d(c) = 1 - \cos(\mathbf{q}, \mathbf{e}_c)$ denote the original FAISS distance between the query embedding $\mathbf{q}$ and the chunk embedding $\mathbf{e}_c$, and let $N(c) = \{n \in C : (c, n) \in E\}$ denote the set of neighbors of c within the candidate pool C. The integrated distance is

$$d_{\text{int}}(c) = \begin{cases} \alpha\, d(c) + \dfrac{1-\alpha}{|N(c)|} \displaystyle\sum_{n \in N(c)} d(n) & \text{if } N(c) \neq \emptyset, \\ d(c) & \text{otherwise,} \end{cases} \tag{19}$$

where $\alpha = 0.5$ balances the original distance with the neighborhood signal. Candidates without graph neighbors retain their original FAISS score.

GoPs Pipeline. The GoPs retrieval function returns the top-k passages with the lowest integrated distance, computed over a pool of n FAISS candidates:

$$R_{\text{GoPs}}(q) = \text{top}_k(\text{rerank}_G(\text{top}_n(\text{FAISS}(q, C)))), \tag{20}$$

with $n = 20$ as the candidate pool size and $k = 5$ as the final number of returned passages.