Trustworthiness of LLMs in Grading and Demographic Fairness in Medical RAG

Large Language Models (LLMs) have quickly become essential in healthcare, but even top models trained on vast biomedical datasets can generate factually incorrect or biased responses, with notable implications for medicine. Prior work shows that retrieval-augmented generation (RAG) can improve accuracy (Alkhalaf et al., 2024, Xiong et al., 2024, Li et al., 2025), raising the question of whether retrieval mitigates, exacerbates, or leaves bias unaffected, or if it introduces new forms of bias. To study this, we investigate patient-level bias through gender and ethnicity perturbations and physician-level bias through persona prompting, where models grade answers of a "female doctor" or "male doctor". For the patient-level bias experiment, we compare a baseline LLM (GPT-4o Mini) to a MedRAG-enhanced version with two retrieval strategies(MedCPT and BM25) and two external corpora(Textbooks and Statpearls), while we extend evaluation to multiple models (GPT-4o Mini, Meta LLaMA 3.1 8B, Gemini, and Claude 3.5 Haiku) for the physician-level bias experiment, measuring endorsement and rejection accuracy to detect conservativeness or sycophancy.

We evaluate on the *DiversityMedQA* dataset, which perturbed *MedQA* questions along gender and ethnicity, totaling to 1,040 gender items and 1,068 ethnicity items where the clinically correct answer remains invariant under demographic edits. To reduce confounding, we modified only references to the primary patient, creating parallel male, female, and genderless versions, and used the released ethnicity subset as it was released by the dataset authors. We selected 300 gender pairs and 300 ethnicity pairs for analysis and utilized evaluation metrics including first-index accuracy, Majority@5, and total proportion, and applied two-proportion Z-tests to measure statistically significant disparities across demographic perturbations.

Altering the patient's gender or ethnicity in the query had a negligible effect on base model performance: for both perturbations, first-answer accuracy, within the range of 69–71% for the original questions showed similar pattern for the perturbed pairs, and statistical tests confirmed no significant differences, indicating that diagnostic accuracy was not noticeably biased by patient demographic description. Retrieval augmentation (RAG) consistently boosted accuracy into the low-to-mid 70% range, improving performance without introducing new disparities. In the physician-persona experiments, all models showed a consistent *conservative bias*—more accurate at rejecting incorrect answers than endorsing correct ones—with highly significant gaps ranging from ~9 points (Claude) to over 40 points (GPT-4o Mini and LLaMA 3.1 8B). Persona settings (female vs. male doctor) produced very similar patterns of conservativeness with no significant gender differences, and we did not observe evidence of *sycophancy*. Instead, models systematically leaned conservative, rejecting incorrect answers more reliably than confirming correct ones.