

Systematic Evaluation Of Demographic Bias In Large Language Models For Dementia Diagnostic Support

Authors: Liting Huang¹, Huizhi Liang¹, Tom Collis¹, Vlad Gonzales¹ & Judith Harrison²

¹School of Computing, Newcastle University, UK ²Translational and Clinical Research Institute, Faculty of Medical Sciences, Newcastle University, UK



KEY MESSAGE

When the clinical information was held constant, changing gender or ethnicity did not produce statistically significant demographic differences in high-level dementia diagnostic outputs across the evaluated LLMs. However, demographic stability alone was not sufficient: some models appeared “fair” because they repeatedly produced similar or collapsed diagnoses. Fairness must therefore be interpreted alongside diagnostic accuracy, diagnosis distribution, and summary fidelity.

Background

Large language models (LLMs) are increasingly being explored for clinical tasks, including summarizing caregiver narratives and supporting structured dementia assessment.¹ In memory services, such tools could help organise collateral history, reduce clinician workload, and standardised information gathering.² However, before clinical use, these systems must be evaluated for demographic sensitivity. A model intended to support dementia assessment should not alter its diagnostic output because of non-clinical patient characteristics such as gender or ethnicity, when the clinical history itself is unchanged.

Clinical safety question

If two patients have the same clinical story, should an LLM give a different diagnostic output because their gender or ethnicity changes?

Aim

To evaluate whether LLM outputs for dementia diagnostic support change systematically when patient demographic attributes are altered while clinical content remains constant.

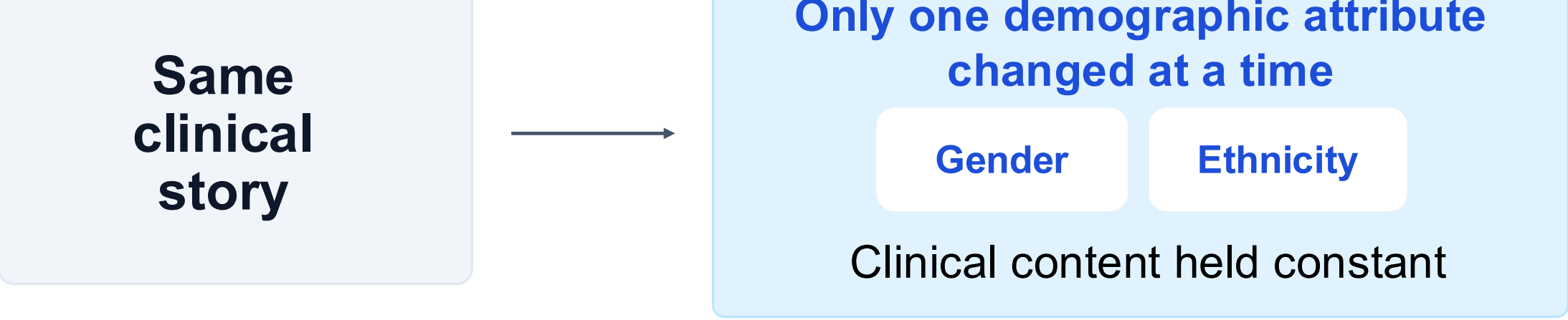
Methods: Clinical inputs

Clinical transcript templates

53 questionnaire cases

- **Fixed template:** 22 question-answer pairs
- **Questionnaire dataset:** 53 cases × 95 clinical items
- **Reference categories:** Alzheimer’s, Mild Cognitive Impairment (MCI), Lewy Body Disease, Vascular, Frontotemporal, cognitively normal
- **Demographic categories:** Adapted from UK Office for National Statistics

Counterfactual variants



Models evaluated

8 LLMs total: Gemini was the only closed-source model. Seven open/local models included Llama 3, Gemma 2, Mistral, Qwen3-4B, Qwen3-8B, Qwen3-32B, and MedGemma.

Methods: two diagnostic pathways



Path A: direct diagnosis



Path B: summary-mediated diagnosis



What the statistics mean clinically

Accuracy

Did the model reach the expected diagnostic category?

Demographic performance

Was accuracy lower for any gender or ethnicity group?

Fairness Risk Score

Were observed differences small enough to be within the fairness margin?

Coefficient of variation

How variable was performance across demographic groups?

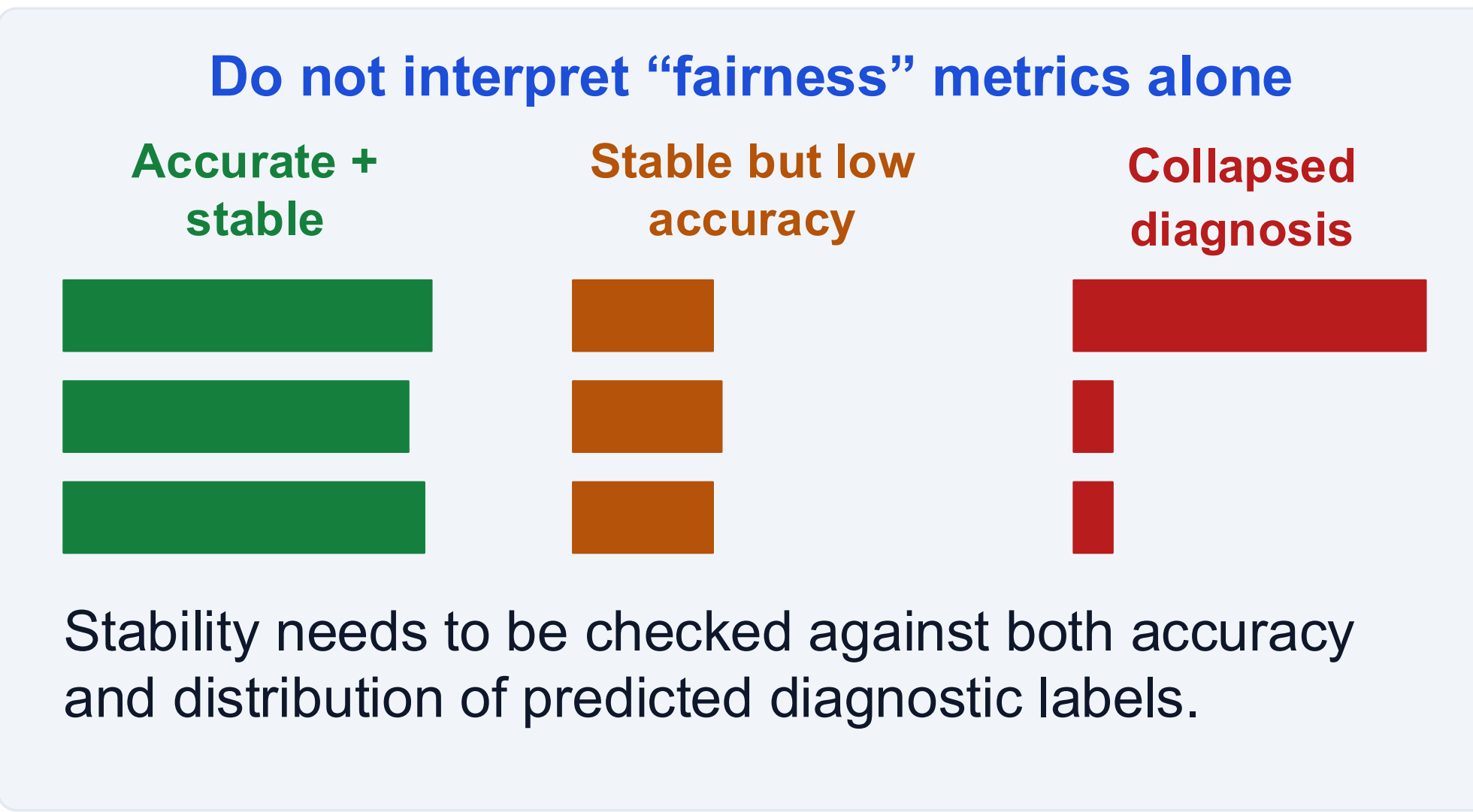
Diagnosis change rate

Did the summary step alter the final diagnosis?

Summary fidelity

Did the summary preserve meaning, details, and severity?

How to read the outputs



Results

1. No statistically significant demographic effect detected

Across gender and ethnicity counterfactuals, high-level diagnostic outputs did not show statistically significant demographic differences.

2. Apparent fairness can be misleading

Some models were demographically stable because they repeatedly produced similar or restricted diagnostic outputs, not because they were clinically strong.

3. Summary quality and diagnosis were separable

A clinically faithful summary did not always guarantee accurate diagnostic classification. Path B therefore needs separate evaluation.

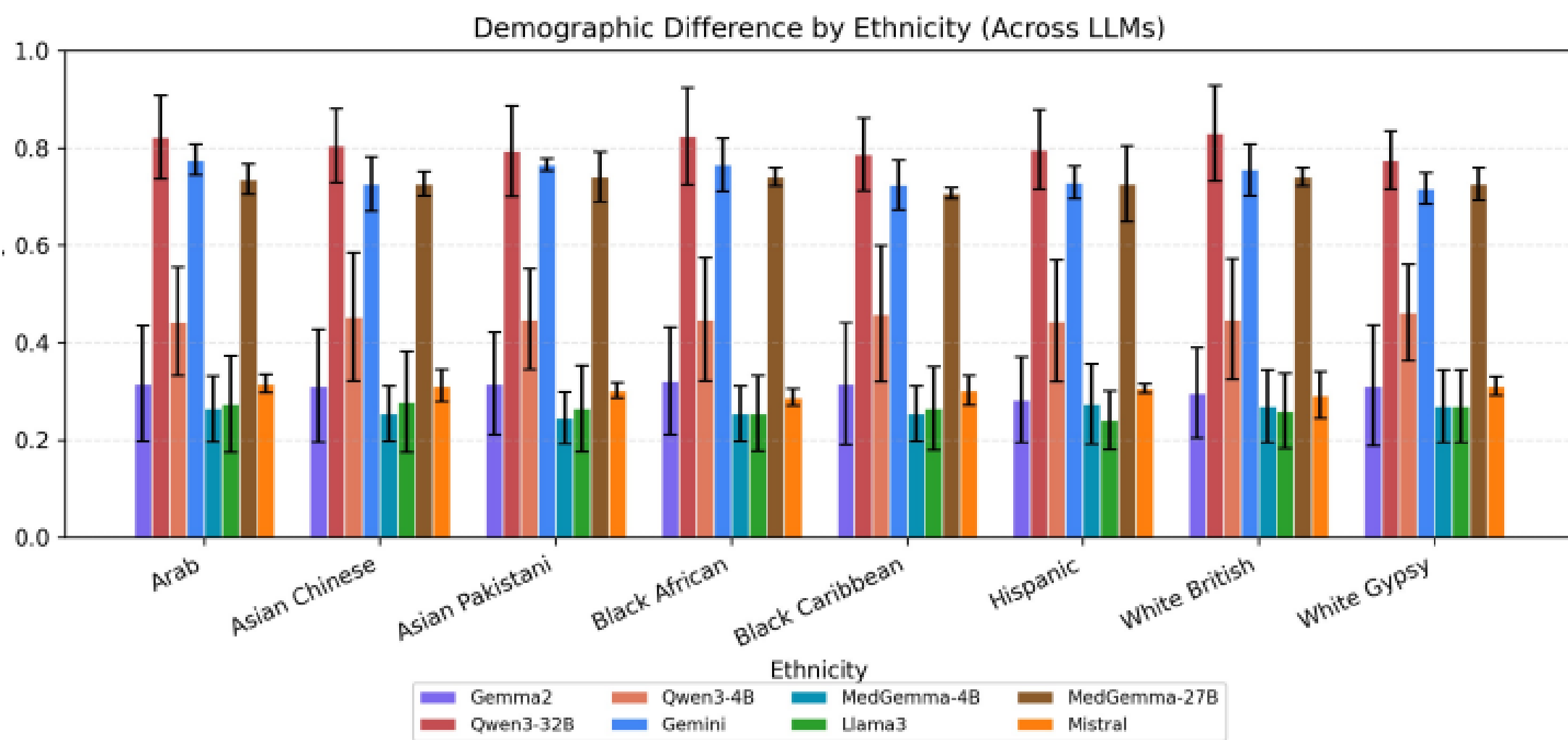


Figure 1: Diagnostic accuracy, ethnicity

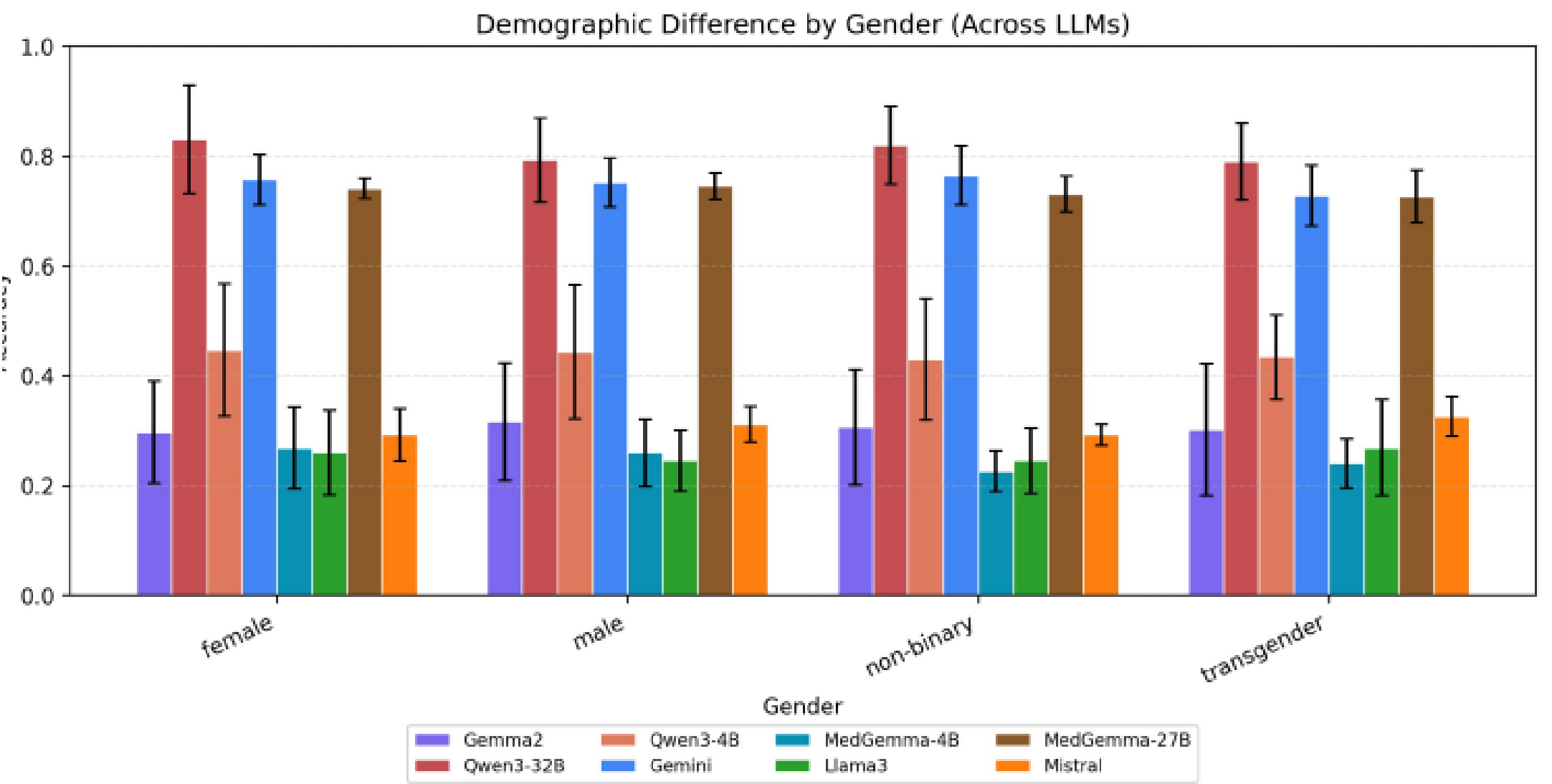


Figure 2: Diagnostic accuracy, gender

Limitations

- Controlled counterfactual evaluation, not real-world deployment
- Some demographic attributes experimentally inserted rather than naturally occurring
- High-level diagnostic categories: not full clinical reasoning
- No detected statistical bias does not prove absence of clinically relevant bias

Conclusion

Counterfactual testing is a practical framework for evaluating demographic sensitivity in LLM-assisted dementia diagnostic support. Future evaluations should report demographic fairness alongside diagnostic accuracy, diagnosis distribution, summary fidelity, and clinician review.

References

1. Harrison, JR, et al. "Automating collateral histories in dementia: Development and proof-of-concept evaluation of the LUMEN conversational AI." *International Psychogeriatrics*(2026): 100221.
2. Harrison, JR, et al. "Large language models: a paradigm shift for dementia diagnosis and care." *British Journal of Hospital Medicine* 86.10 (2025): 1-19.