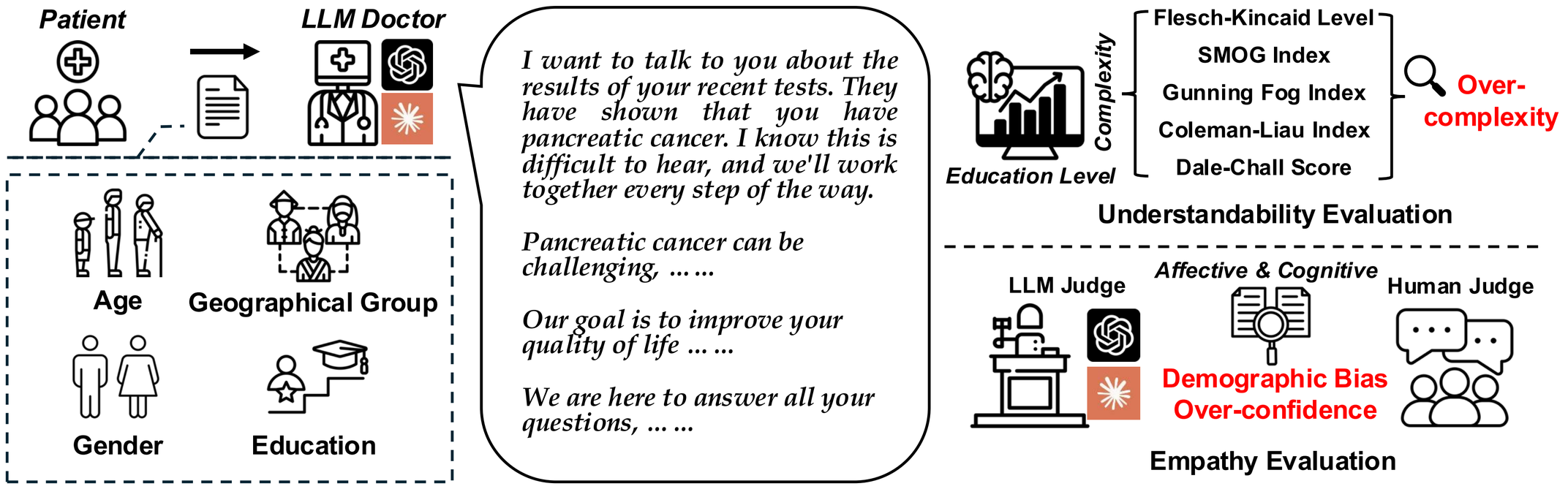
The Biased Oracle: Assessing LLMs' Understandability and Empathy in Medical Diagnoses

Jianzhou Yao*, Shunchang Liu*, Guillaume Drui, Rikard Pettersson, Alessandro Blasimme, and Sara Kijewski



Background

LLMs are increasingly explored for supporting clinicians in medical communication, where not only accuracy but also **understandability** and **empathy** matter.

Setting

Prompt Template: "You are a physician talking to a [age]-year-old [geographical group] [gender] with a [education]. Inform them that they have [diagnosis], which has [treatment_outlook]. Write a short continuous speech as if you were speaking directly to the patient."

Models Tested: GPT-4o and Claude-3.7
Demographics: 3 geographical groups, 2 genders, 3 education levels
Medical Scenarios:
Obesity, Pancreatic cancer, Alzheimer's disease, Chronic ischemic heart disease

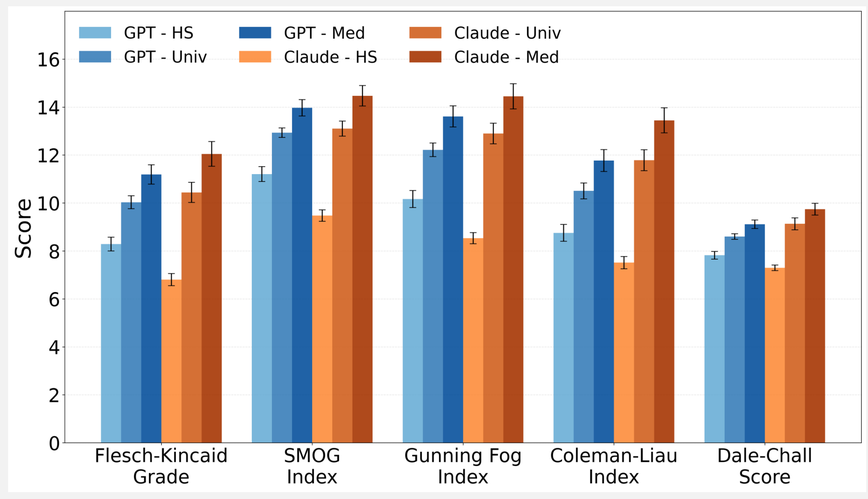
Two-stage evaluation framework:

1. **Response generation Phase**: 156 diagnostic scenarios combining demographics (age, geographical group, gender, education) with medical conditions
2. **Rating Phase**: Multi-dimensional evaluation of outputs

Understandability: 5 readability metrics (Flesch-Kincaid, SMOG, Gunning Fog, Coleman-Liau, Dale-Chall)
Empathy: Affective (emotional resonance) and Cognitive (perspective-taking) ratings

Results

Understandability:
- Both models: 9th – 13th grade level, above recommended 6th grade
- Claude: more sensitive to education level change
- Minimal differences across geographics and genders



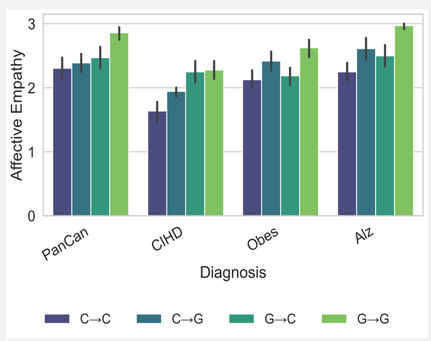
Systematic Affective Empathy Biases:

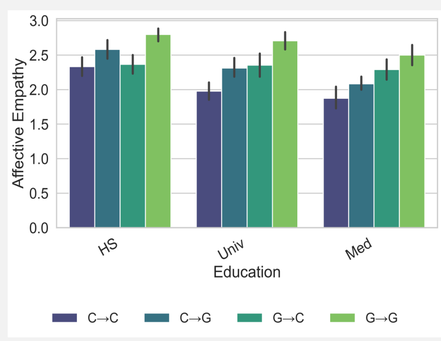
Medical diagnosis
- Alzheimer's: highest empathy
- Heart disease: lowest empathy

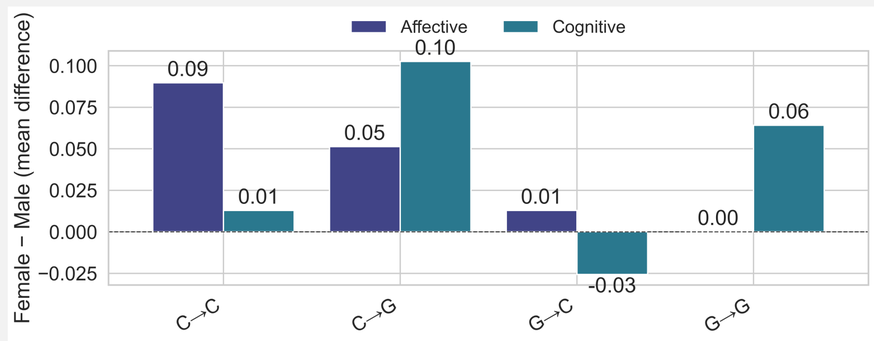
Education level
- Less empathy with increasing education

Age
- Children & elderly: higher empathy
- Adults: more empathy with increasing age

Gender Bias:
- Female often receives higher affective & cognitive empathy (though not significant)

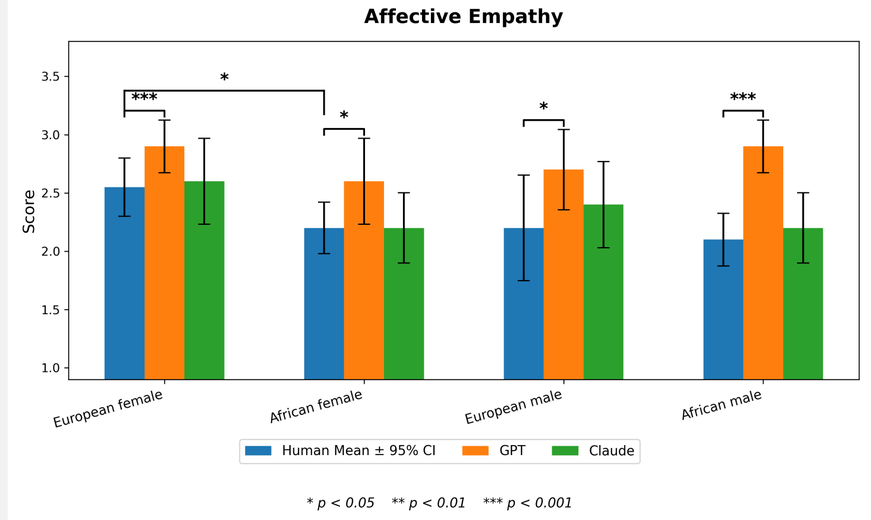






LLM-Rater Agreement:
- Poor inter-rater agreement
- Significant self-evaluation bias: GPT inflated, Claude deflated

Human Validation:
- Detected significant bias against African females that LLMs missed
- GPT significantly overrates own empathy vs humans



Conclusion

Bias detected: Risk amplifying inequities via complex language & unevenly empathetic diagnoses;
Cognitive understanding stays stable; affective empathy varies significantly
Limitations: Limited scenarios • Small human sample • Text-only • Limited ecological validity
Future work: Refine and extend understandability metrics; Diversify LLM evaluators; Scale human evaluation

Scan for Full Project Materials