UNDERSTANDING USER TRUST IN MULTIMODAL MEDICAL DIAGNOSES

Evaluating User Trust in General vs Domain-Specific Medical AI

AUTHORS: BHASKAR NIKHIL SUNKARA, PALLAVI SHARMA

MOTIVATION

Can a general-purpose LLM (like Gemini) be more trusted than a domain-specific model (like BioGPT) in critical contexts like medical diagnosis?

- Generalist models are optimized for natural language fluency and narrative reasoning.
- Domain-specific models are tuned for accuracy and terminology but may lack clarity.
- For non-expert users, explanation quality may be more important than clinical depth.

Why medicine?

Healthcare combines high cognitive load, emotional sensitivity, and urgency — making it a compelling domain for studying explanation clarity, trust calibration, and human-AI interaction.

RESEARCH QUESTIONS

Q1 HOW DO USERS RESOLVE CONFLICTING DIAGNOSES FROM DIFFERENT AI MODELS?

Q2 HOW DOES AI-EXPRESSED CONFIDENCE AFFECT TRUST CALIBRATION?

Q3 HOW DOES EXPLANATION CLARITY AND REASONING STYLE IMPACT TRUST?

RESEARCH DESIGN

A mixed-methods, between-subjects experimental design to evaluate how users trust, interpret, and prefer AI-generated medical explanations from two distinct multimodal models.

Design Overview:

- 2 × 2 Factorial Design
  → Input Type: Text vs. Image + Text
  → Diagnosis Type: Consistent vs. Conflicting outputs
- Each participant saw one scenario with both model outputs (Gemini & BioGPT)
- Model order was randomized to control for anchoring effects

Quantitative

- Trust and confidence ratings
- Model preference
- Bias trait surveys (anchoring, automation, confirmation)

Qualitative

- Open-ended explanations
- Sentiment analysis using RoBERTa
- Thematic analysis of reasoning style

Why this design?
To capture both measurable trust behavior and explanatory nuance, allowing us to understand not just what users chose, but why they trusted it.

STUDY DESIGN

PARTICIPANT PROFILE

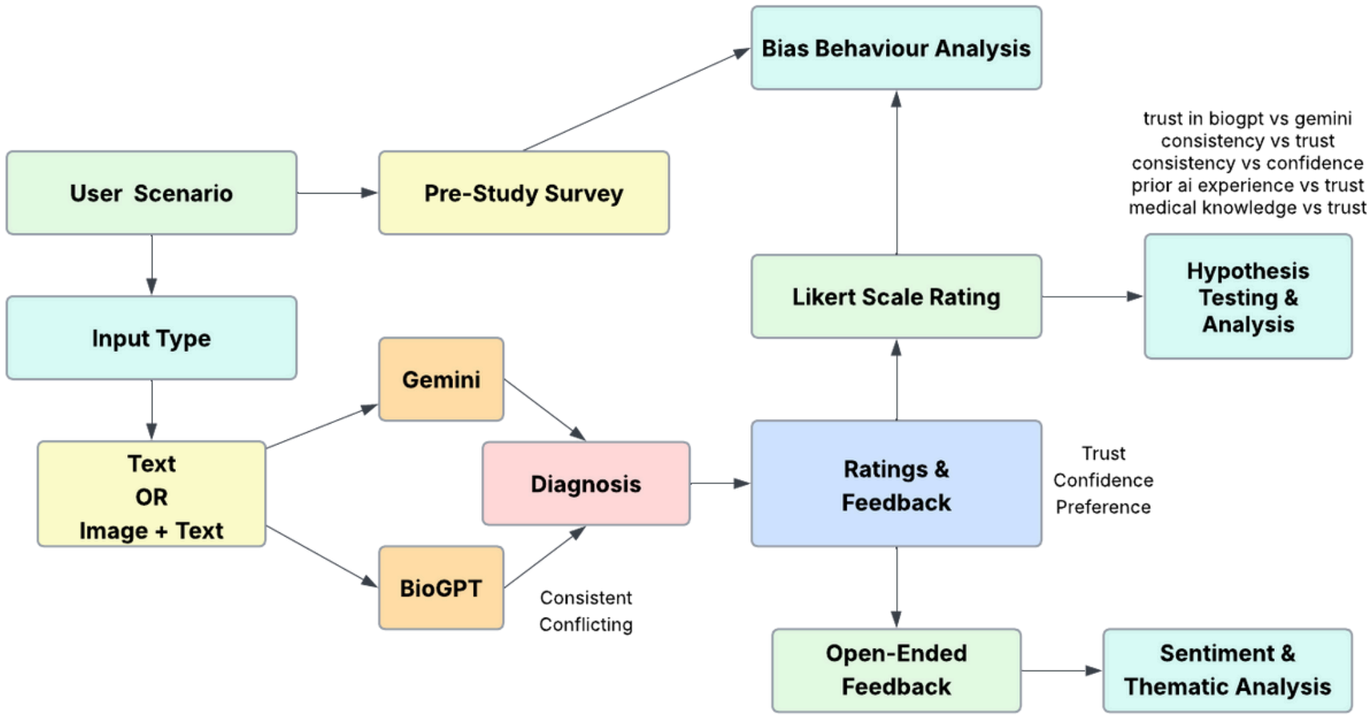
- Mean age: 27.3 years (range: 18–51)
- AI familiarity: High (mean = 4.15 / 5)
- Medical background: 81% had no formal training
- Gender split: 58% male, 42% female

INTERFACE

- Diagnoses varied in consistency and confidence style
- In some trials, a confidence score was explicitly shown for both models to examine the effect of perceived certainty on trust

COLLECTED MEASURES

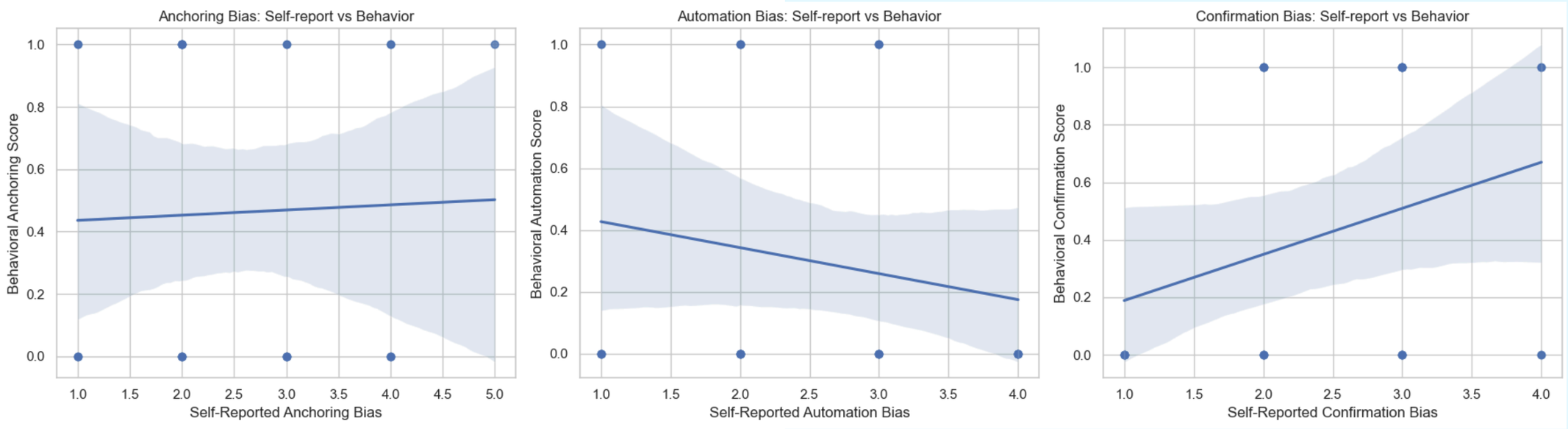
- Ratings: Trust, confidence, preference (Likert scale)
- Bias traits: Anchoring, automation, confirmation (self-report)
- Open-ended reasoning: Justification for model preference



EVALUATION AND FINDINGS

DO USERS RECOGNIZE THEIR OWN BIASES?

Evaluating whether participants' self-awareness of cognitive biases aligns with their actual behavior during AI diagnosis evaluation.

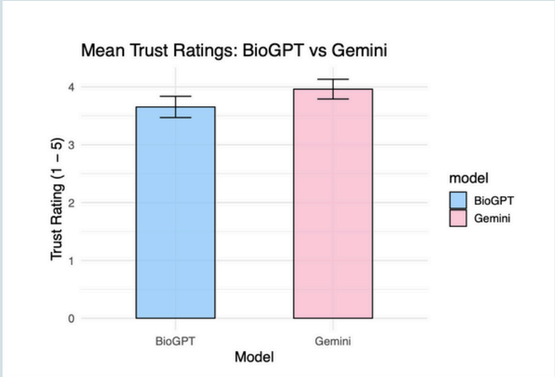


METHOD

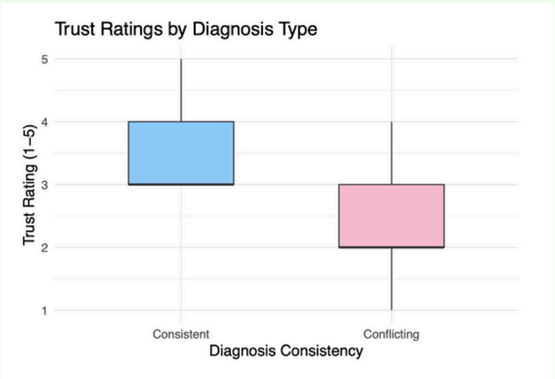
- Measured self-reported scores for:
  a. Anchoring bias
  b. Automation bias
  c. Confirmation bias
- Compared with behavioral indicators (based on model order, agreement following, etc.)
- Used Pearson correlation to assess alignment

- **Anchoring Bias:** No meaningful correlation (r = 0.037, p = 0.857) — users unaware of anchoring effects
- **Automation Bias:** Weak negative correlation (r = −0.190, p = 0.352) — suggests partial mismatch between belief and behavior
- **Confirmation Bias:** Moderate alignment trend (r = 0.315, p = 0.117) — users somewhat aware of confirmatory behavior, but not significantly

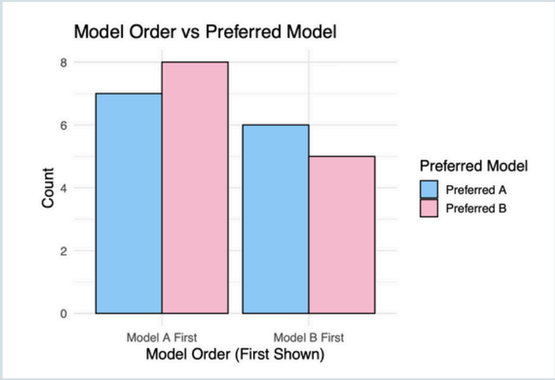
HYPOTHESIS TESTING



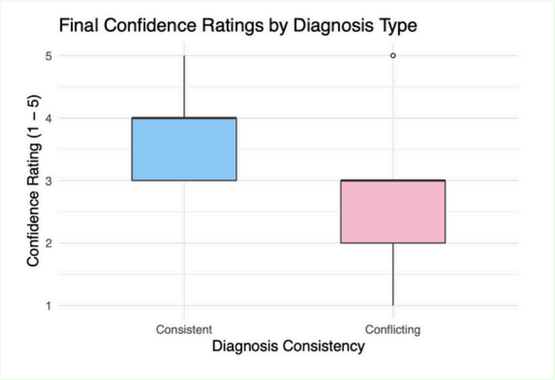
Users will exhibit higher trust in BioGPT due to its medical domain specialization.
→ Mann-Whitney U Test
p-value: **0.268**
Conclusion: No statistically significant difference in trust between BioGPT and Gemini.



Trust ratings will be lower when the diagnoses are conflicting versus consistent.
→ Mann-Whitney U Test
p-value: **0.0019** (significant)
Effect size: 0.614 (large)
Conclusion: Users trusted AI significantly less when the models disagreed.

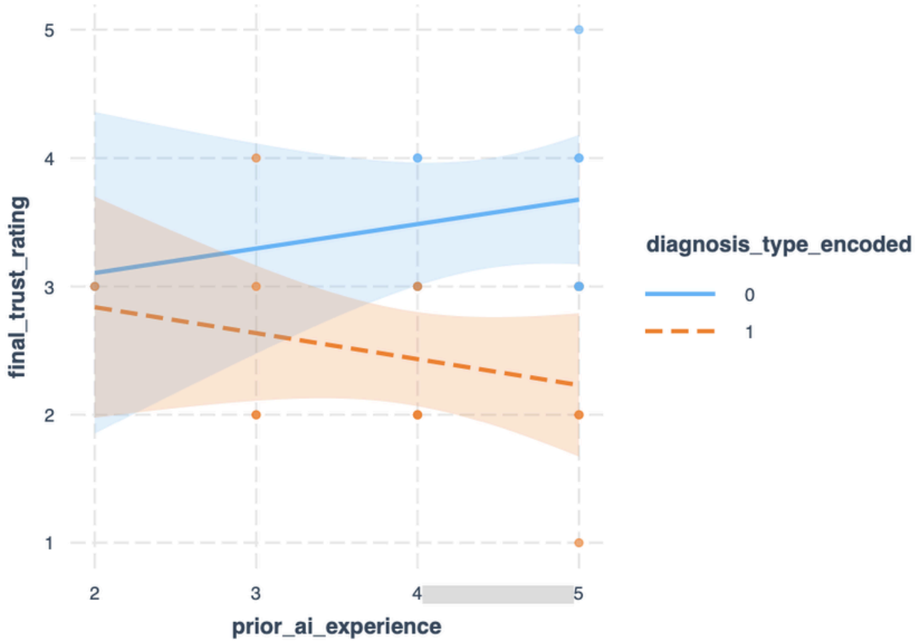


Users will be more likely to prefer the model shown first.
→ Chi-Square Test of Independence -
p-value: **0.9**
Conclusion: No evidence of anchoring — model order did not significantly influence preference.

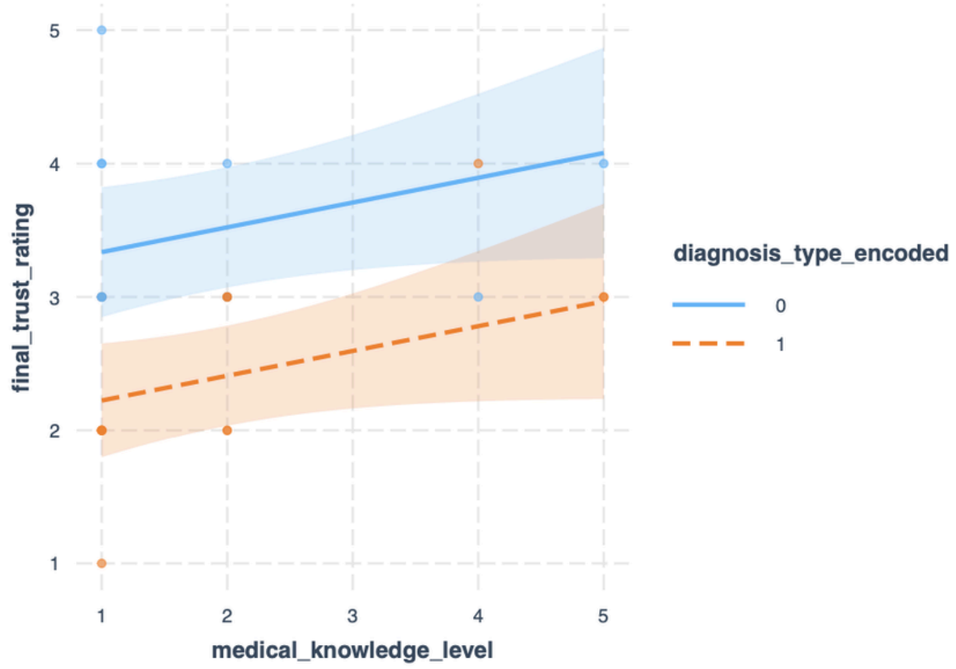


Users will report higher confidence when the diagnoses are consistent.
→ Mann-Whitney U Test-
p-value: **0.038** (significant)
Effect size: 0.411 (moderate)
Conclusion: Users were significantly more confident in their decision when the models agreed.

DO BACKGROUND TRAITS INFLUENCE AI TRUST?



Prior AI Experience → Context-Dependent Trust
In consistent diagnoses, trust rose with AI familiarity
But in conflicting diagnoses, experienced users trusted the AI less
Familiar users may be more critical when AI outputs disagree

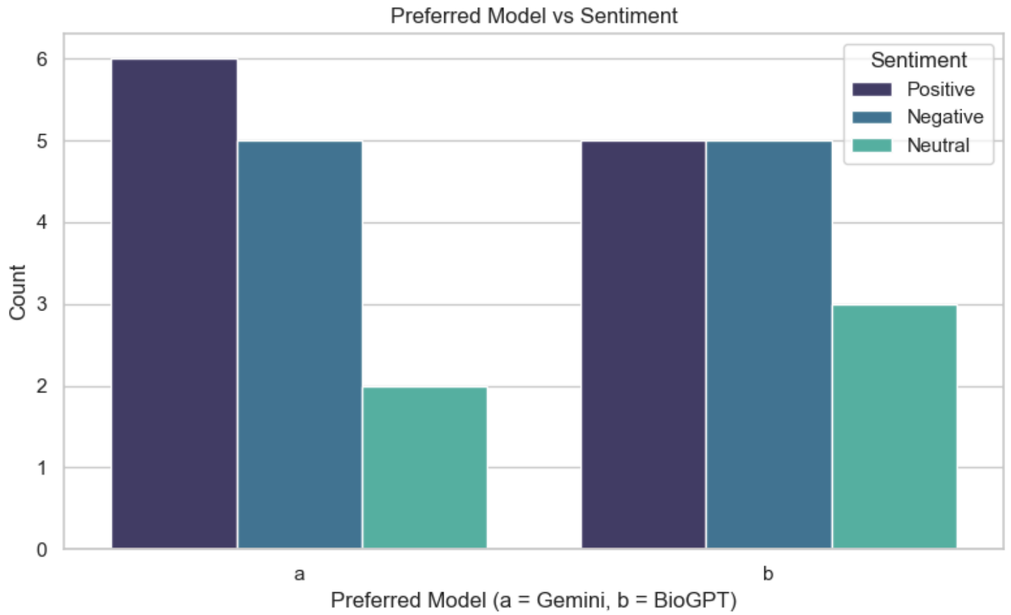


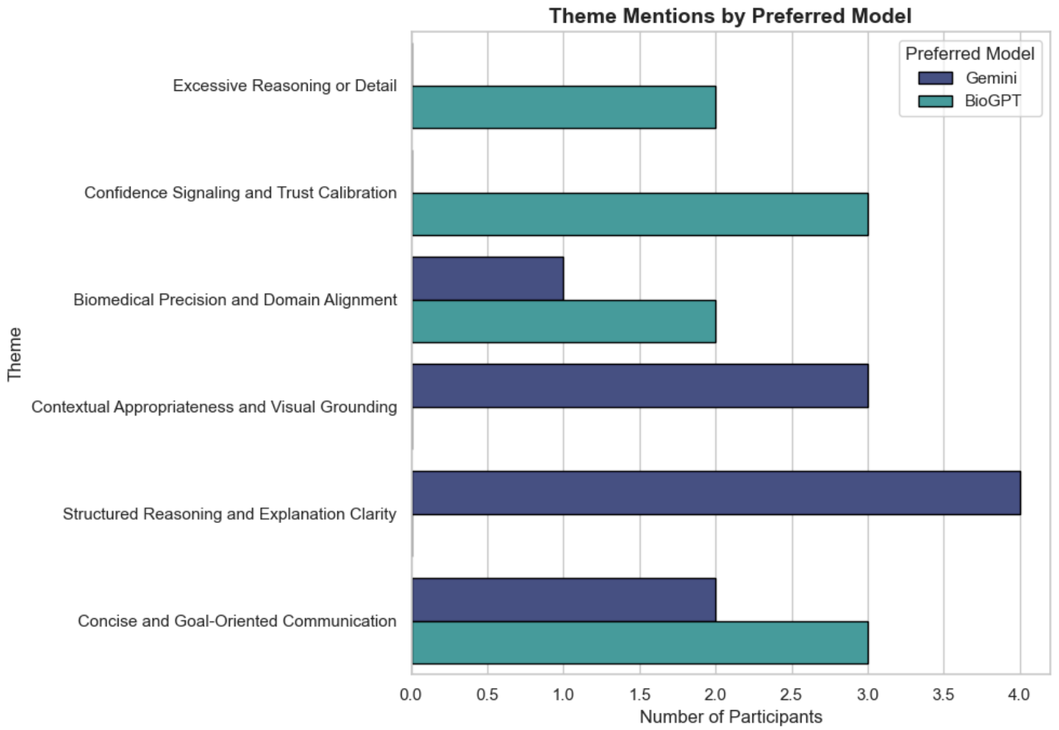
Medical Knowledge → Higher Trust
Trust increased consistently with medical expertise
Especially strong effect in conflicting diagnoses, where non-experts were less confident
Domain knowledge helps users evaluate explanations more critically

THEMATIC & SENTIMENT ANALYSIS

SENTIMENT ANALYSIS

- Gemini's preference aligned with more positive sentiment, reflecting its clarity and accessibility.
- BioGPT evoked more neutral or negative tone—respected, but harder to follow.





THEMATIC ANALYSIS

- Gemini was frequently associated with clarity, structured reasoning, and visual grounding, reinforcing its perceived communicative strength.
- BioGPT was linked to domain precision and confidence signaling, but also noted for excessive detail by some users.

Clarity Wins Over Expertise
Even in a medical setting, users often preferred Gemini's explanations due to clarity and structure — a surprising result that challenges assumptions about domain-specific model dominance

Trust is Fragile in Conflict
When models disagreed, both trust and confidence significantly dropped, showing that consistency matters more than accuracy alone in user decision-making.

Bias Awareness ≠ Bias Behavior
Participants' self-reported cognitive biases did not strongly correlate with their actual behavior, suggesting limited bias self-awareness, particularly for anchoring and automation biases.

Sentiment and Style Matter
RoBERTa sentiment scores and thematic coding revealed that positive tone and narrative coherence played a key role in how users justified their trust.

Expertise Boosts Trust, Experience Can Undermine It
Trust rose with medical knowledge but dropped with AI experience during conflicts — suggesting that familiarity made users more skeptical.

KEY TAKEAWAYS AND DESIGN IMPLICATIONS

IMPLICATION:

Explanation design and delivery are as critical as correctness in health AI. Users trust what they can understand — even if it's less medically sophisticated.

FUTURE WORK

- Include medical professionals to compare expert vs non-expert trust behavior
- Test explainability aids (e.g., model summaries, highlights)
- Explore AI collaboration — have models explain or challenge each other

REFERENCES

1. Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (Barcelona, Spain) (FAT* '20). Association for Computing Machinery, New York, NY, USA, 295–305. doi:10.1145/3351095.3372852
2. Thomas Buckley, James A. Diao, Pranav Rajpurkar, Adam Rodman, and Arjun K. Manrai. 2024. Multimodal Foundation Models Exploit Text to Make Medical Image Predictions. arXiv:2311.05591 [cs.CV]
3. Chandan Agrawal, Ashish Papanai, and Jerome White. 2024. Maintaining User Trust Through Multistage Uncertainty Aware Inference. arXiv:2402.00015 [cs.AI] https://arxiv.org/abs/2402.00015
4. Helena Löfström, Karl Hammar, and Ulf Johansson. 2022. A Meta Survey of Quality Evaluation Criteria in Explanation Methods. Springer International Publishing, 55–63. doi:10.1007/978-3-031-07481-3_7
5. Yoonjoo Lee, Kihoon Son, Tae Soo Kim, Jisu Kim, John Joon Young Chung, Eytan Adar, and Juho Kim. 2024. One vs. Many: Comprehending Accurate Information from Multiple Erroneous and Inconsistent AI Generations. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (Rio de Janeiro, Brazil) (FAccT '24). Association for Computing Machinery, New York, NY, USA, 2518–2531. doi:10.1145/3630106.3662681
6. Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. BioGPT: generative pre-trained transformer for biomedical text generation and mining. Briefings in Bioinformatics 23, 6 (09 2022). doi:10.1093/bib/bbac409 arXiv:https://academic.oup.com/bib/articlepdf/23/6/bbac409/47144271/bbac409.pdf bbac409.

THANK YOU

Questions?