Seeing is Believing? How Different Multimodal AI Models Shape User Trust in AI-Generated Diagnoses

Pallavi Sharma*
University of Wisconsin - Madison
Wisconsin, USA
pallavisharm@wisc.edu

Bhaskar Nikhil Sunkara*
University of Wisconsin - Madison
Wisconsin, USA
bsunkara3@wisc.edu

Abstract

Keywords

.

ACM Reference Format:
Pallavi Sharma* and Bhaskar Nikhil Sunkara*. 2018. Seeing is Believing? How Different Multimodal AI Models Shape User Trust in AI-Generated Diagnoses. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 7 pages. https://doi.org/XXXXXXX.XXXXXXX

1 Introduction

Artificial intelligence (AI) is increasingly transforming medical diagnostics by providing automated decision support through advanced language and vision-language language models.Systems such as Gemini 1.5 Pro and BLIP-2 leverage multimodal inputs, including textual descriptions and medical images, to generate detailed explanations for clinical decision-making. Meanwhile, domain-specific models like BioGPT provide specialized medical insights based on biomedical literature. These models hold significant potential in assisting patients and healthcare professionals by offering interpretable insights. However, a fundamental challenge persists: how do users trust and interpret AI-generated textual explanations, particularly when different models provide conflicting diagnoses or express uncertainty in varying ways?

Existing research in explainable AI (XAI) [26] and uncertainty aware AI [1] highlights that trust in artificial intelligence is not solely determined by accuracy. It is shaped by explanation style, confidence calibration [39], and cognitive biases[6]. While previous studies have examined the impact of AI-generated explanations on trust [8], few have directly compared different AI models in terms of their influence on user decision-making and reliance, particularly in high-stakes medical contexts. Research suggests that users may exhibit automation bias, leading to over-reliance on AI-generated outputs, or anchoring bias, causing them to fixate on the first explanation they receive. Additionally, the way AI models communicate uncertainty may affect user confidence and willingness to integrate AI recommendations into their decision-making processes.

This study aims to address this knowledge gap by investigating how users perceive, interpret, and trust textual diagnostic explanations generated by Gemini and BioGPT, with BLIP-2 as a backbone which is used for medical image interpretation. Specifically, this research explores three key questions:

- RQ1: How do users resolve discrepancies when two AI models provide conflicting diagnoses?
- RQ2: How does the confidence level expressed by AI influence trust in medical decision-making?
- RQ3: How do differences in explanation clarity and reasoning structure affect user trust in AI-generated medical text?

To answer these questions, this study conducts a comparative user evaluation of trust dynamics between Gemini 1.5 Pro and BioGPT in a medical diagnostic setting. By analyzing how users navigate conflicting AI outputs, assess explanation reliability, and interpret uncertainty, this research contributes to both human-computer interaction (HCI) and AI ethics. The findings aim to inform the development of more transparent and trustworthy AI-driven decision-support tools, ensuring that AI-generated explanations align with user expectations and support effective clinical decision-making.

2 Related Work

Advancements in multimodal AI models have introduced new paradigms in human-AI interaction, particularly in decision-making. Despite their increasing adoption, how users trust and interpret AI explanations—especially when AI models provide conflicting outputs—remains an open question. Trust in AI is not merely a function of accuracy but also of explainability, uncertainty communication, and cognitive biases. Existing research on explainable AI (XAI) and uncertainty-aware AI provides insights into trust dynamics, but few studies have explored how different multimodal AI models compare in influencing user trust and decision-making.

This section reviews vision-language models (VLMs), trust in AI explanations, cognitive biases in AI-assisted decision-making, and human-AI collaboration, situating our study within ongoing HCI and AI research.

2.1 Vision-Language Models (VLMs) in Medical AI

The emergence of vision-language models (VLMs) marks a major shift in AI-driven decision-making, particularly in domains requiring multimodal reasoning. Unlike unimodal AI models that process either text or images, VLMs integrate textual and visual data to enhance context understanding, generate explanations, and support decision-making [15][31]. These capabilities have fueled interest in

Unpublished working draft. Not for distribution.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
Conference acronym 'XX, Woodstock, NY
© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

Unpublished working draft. Not for distribution.

their potential applications in medical AI, where synthesizing text-based clinical records with medical images is crucial for accurate, interpretable insights.

Recent studies highlight that VLMs outperform unimodal AI models in complex reasoning tasks. For instance, GPT-4V demonstrated higher diagnostic accuracy (80.6%) when integrating image-text data, compared to text-only (66.7%) or image-only (45.2%) models [3]. Similarly, BioViL-T and Med-PaLM 2 have been developed to process structured and unstructured medical data, enhancing their ability to answer domain-specific queries and detect inconsistencies across modalities [14][37]. However, despite these improvements, concerns remain regarding trustworthiness, reliability, and explanation consistency, especially when different VLMs generate conflicting outputs for the same input [27].

In this study, we focus on three models with distinct architectures and capabilities. Gemini 1.5 Pro, a general-purpose multimodal model developed by Google, supports image and text inputs and is capable of general reasoning across domains[33]. BioGPT, a domain-specific language model trained on biomedical literature, offers specialized text-based explanations for medical prompts[23]. Finally, BLIP-2, a vision-language model, is used to interpret medical images and extract clinically relevant visual findings[21]. While each model is capable of generating explanations in clinical contexts, their differences in training objectives and reasoning mechanisms may shape user trust and perception in fundamentally different ways.

Current research lacks direct comparisons of how different VLMs impact user trust and decision confidence—particularly when they produce different outputs or express uncertainty in different ways. Our study fills this gap by systematically examining how users perceive, interpret, and trust explanations provided by Gemini 1.5 Pro and BioGPT, using BLIP-2 as a visual context provider.

2.2 AI Explanations, Trust, and Uncertainty in Decision-Making

User trust in AI explanations has been extensively studied in HCI and explainable AI (XAI). However, its implications for multimodal AI models remain underexplored. Prior research suggests that explainability alone does not guarantee trust; rather, trust is shaped by how AI explanations align with users' mental models, expectations, and confidence levels [7][38].

Studies in XAI show that overly detailed explanations can reduce trust, as users may perceive them as justifications rather than transparent reasoning [36]. Conversely, minimalistic or vague explanations can fail to provide enough information for meaningful decision support [29]. These challenges are further amplified in multimodal AI, where text-based and visual explanations must be synthesized coherently.

Another critical challenge arises when different AI models provide conflicting yet equally plausible explanations. Research in uncertainty-aware AI systems suggests that users are more likely to trust AI recommendations when uncertainty is clearly communicated, but trust declines if uncertainty is perceived as incompetence [16]. Some studies advocate for confidence-calibrated explanations, where AI presents its level of certainty alongside reasoning [17]. However, in multimodal AI, where textual and visual reasoning must align, misalignment between confidence scores and explanation quality can negatively impact trust [13].

In the case of Gemini and BioGPT, these concerns become even more pronounced. If both models are presented with the same image-text input but generate different diagnoses or explanations, users must determine which AI to trust and why. Our research investigates how users resolve discrepancies, evaluate confidence indicators, and assess explanation reliability across different AI models.

2.3 Cognitive Biases in AI-Assisted Decision-Making

Cognitive biases significantly shape how users engage with AI-generated explanations. Among the most well-documented biases in AI trust research are automation bias (over-reliance on AI outputs) and confirmation bias (favoring AI responses that align with prior beliefs) [2][11]. Studies show that users are more likely to trust AI-generated recommendations when framed with high confidence—even when incorrect [28]. This raises concerns in multimodal AI decision-making, where confidence scores in text and image explanations may not always align.

A key challenge in multimodal AI is whether users perceive one modality as more authoritative than the other. Research on visual decision support tools suggests that when AI explanations include both text and images, users tend to prioritize the visually dominant representation, even if textual justifications provide richer reasoning [18]. Similarly, users may exhibit anchoring bias, where they fixate on the first explanation they receive rather than critically evaluating alternative AI-generated outputs [24].

These biases become particularly relevant when users must resolve discrepancies between Gemini and BioGPT. If one model expresses uncertainty more explicitly, while the other presents a confident but incorrect response, how do users decide which AI to trust? Our study evaluates how AI explanations interact with known cognitive biases and whether specific explanation styles lead to over-reliance or skepticism.

2.4 Human-AI Collaboration and Explainability in Multimodal AI

While AI aims to support human decision-making, its usability is dictated by how well explanations integrate into user workflows. In multimodal AI interactions, explanation coherence is crucial: Do AI-generated responses seamlessly combine visual and textual insights, or do they introduce cognitive friction? Studies in XAI interfaces indicate that users prefer explanations that allow interactive probing, enabling them to challenge AI reasoning rather than passively accept it [19, 30]. However, a critical issue in AI explanations, particularly in medical settings, is whether the reasoning provided remains consistent across different prompts and contexts. Studies have shown that trust in AI diminishes when explanations lack coherence or appear contradictory across different queries [2, 10, 20].

One of the biggest challenges with VLM-generated explanations is consistency across different queries. Research shows that users trust AI more when its reasoning remains stable across multiple questions, rather than appearing to contradict itself when prompted

Unpublished working draft.
Not for distribution.

differently [27, 35]. This issue is particularly relevant for Gemini and BioGPT, where responses may vary based on input phrasing, contextual relevance, or fine-tuned biases.

Our research evaluates how users perceive explanation consistency in these models and whether inconsistencies reduce trust over time. Additionally, we explore whether users prefer detailed justifications or minimalist summaries, and whether such preferences shift based on uncertainty levels in AI-generated responses. Prior studies in AI explainability indicate that detailed explanations can enhance trust by providing transparency, but they can also lead to cognitive overload, making users skeptical about overly complex justifications [9, 29, 36]. Conversely, minimalist explanations may be preferred for efficiency but risk appearing overly simplistic or vague [35]. By examining how users engage with Gemini's and BioGPT's textual explanations, this study contributes to ongoing discussions in explainable AI (XAI) and human-AI collaboration by shedding light on the mechanisms that influence trust and decision-making in AI-assisted medical diagnostics.

3 Methodology

This section outlines the study design, which aims to examine how users trust and interpret AI-generated medical diagnoses. The study will compare two AI models including BioGPT, a specialized model trained on medical data, and Gemini 1.5 Pro, a general-purpose model, by evaluating their responses to medical queries. Participants will assess these responses through structured ratings and qualitative feedback. Through a controlled laboratory experiment [5], participants will assess AI-generated responses based on trust, interpretability, and consistency. The study follows a mixed-method design, incorporating both quantitative measures, such as Likert scale ratings, and qualitative insights, obtained through open-ended feedback.

The experimental design includes both within-subjects conditions, where participants encounter consistent and conflicting diagnoses and between-subjects conditions, where the order of model presentation is counterbalanced. A total of 20 participants will be recruited using convenience sampling, allowing for a diverse sample across age, gender, and familiarity with AI. While some participants completed the study in person, the experimental tasks were presented via a web-based interface on a local device. This hybrid setup allowed for real-time observation and assistance and it helped us to understand how explanation quality, confidence language, and model disagreement affect user trust and decision-making.

3.1 Hypotheses

Building upon prior research in AI trust and explainability, we have formulated the following hypotheses:

- H1 (Trust Hypothesis): Users will exhibit higher trust in diagnoses provided by BioGPT compared to Gemini, given its specialized medical training and domain-specific knowledge. Prior research suggests that AI models trained on domain-specific data are more likely to be perceived as authoritative and trustworthy, especially in high-stakes fields such as healthcare [31][9][10]. However, general-purpose models like Gemini may sometimes provide responses that are more coherent or interpretable, despite lacking domain-specific fine-tuning.
- H2 (Conflict + Trust Hypothesis): User trust will be lower when Gemini and BioGPT provide conflicting diagnoses than when they provide consistent diagnoses, particularly in cases involving serious medical conditions. Studies on explainable AI (XAI) show that when AI models give different answers [16][13][12], people find it harder to trust them because they are unsure which one is correct. Clear and well-structured text responses help users understand and trust the information better.
- H3 (Anchoring Bias): Users will be more likely to stick to the first diagnosis they see when the two AI models provide conflicting diagnoses [24][34][25] under serious medical conditions. Inconsistencies in AI-generated outputs have been shown to negatively affect user confidence, as individuals may struggle to determine which response is more reliable.
- H4 (Confidence Hypothesis): Users will report higher confidence in their final diagnosis decision when both AI models provide consistent diagnoses, compared to conflicting diagnoses under serious medical conditions. Inconsistencies in AI-generated outputs have been shown to negatively affect user confidence, as individuals may struggle to determine which response is more reliable [22].

3.2 Participants

For this study, we have collected data from 13 participants (out of planned 20 participants), recruited from personal networks, including friends and family, to evaluate AI-generated medical diagnoses. Participants were selected based on their varying levels of familiarity with AI-based diagnostic tools. The study includes individuals across different age groups, genders, and backgrounds to ensure diverse perspectives on AI trust and reliability.

The target population for this study includes individuals who may use AI tools to assist with medical decision-making, such as general users without medical training, patients managing chronic conditions, caregivers, and healthcare students or professionals who might rely on AI support for second opinions. Participants were drawn from a convenience sample including university students, general volunteers, and friends working in the field of medicine or professionals if available. Both medically trained (e.g. medical students) and non-medical participants (e.g. regular users with no medical background) have been included to capture diverse trust profiles.

Participants were required to be at least 18 years old, fluent in English, and able to provide informed consent. Given the exploratory nature of the study and its focus on user trust in early-stage AI explanations, convenience sampling was appropriate. All participants provided informed consent prior to participating. All participants have completed an initial screening survey to collect demographic data and assess their prior knowledge of medical AI systems to ensure that trust evaluations are not significantly influenced by expertise differences.

The demographic composition of the sample is as follows:

- Age range: 20-55 years (Mean: 30.4, SD = 10.6)
- Gender distribution: 8 Male, 5 Female (for 13 participants)

Unpublished working draft. Not for distribution.

- Education or Occupation: Software Engineers, Healthcare Professionals, Students.
- Familiarity with AI: 10 out of 13 participants (7̃7%) had strong prior experience with AI chatbots or diagnostic tools

To maintain consistency in data collection, all participants completed the study in a controlled setting (in-person or via zoom) using a standardized web-based interface, allowing consistent task delivery.

3.3 Study Design

This study employs a mixed-model experimental design [35] to investigate user trust in AI-generated medical diagnoses, incorporating both within-subjects and between-subjects factors. This mixed approach allows the study to compare how the same participant reacts to consistent and conflicting diagnosis, while also comparing how different participant groups (with different AI order and medical backgrounds) make decisions. This combination helps capture both the personal experience of each participant and the broader differences between types of participants, giving a more complete view of how people trust and use AI in serious medical situations.

3.3.1 Between – Participants Design: To measure potential anchoring bias, we implemented a between-participants manipulation based on the order in which AI responses were displayed[24]. Participants were randomly assigned to see either Gemini's response first or BioGPT's response first. By randomizing presentation order across participants, we aimed to observe whether early exposure to one model influenced subsequent trust or decision-making, particularly in conflicting scenarios. This mirrors the anchoring bias hypothesis examined in H3.

3.3.2 Independent Variables. The study incorporated two primary independent variables. The first was the AI model type, either Gemini 1.5 Pro or BioGPT, which determined the source of the diagnostic explanation presented to the user. The second independent variable was conflict presence, referring to whether both models provided consistent (matching) diagnoses or conflicting (non-matching) ones. These variables were systematically manipulated to investigate their effect on user trust, confidence, and decision preference.

3.3.3 Dependent Variables. Several dependent variables were collected through the interface. User trust was measured on a 5-point Likert scale after reviewing each set of AI responses, capturing perceptions of reliability and credibility. Participants also selected their preferred model after each trial, allowing us to track model preference trends. Anchoring behavior was inferred based on whether participants favored the model presented first, especially in conflicting trials. Finally, participants rated how confident they felt in their final decision after viewing both responses, which provided insight into the role of consensus and explanation clarity in shaping diagnostic confidence.

3.3.4 Control Variables. To reduce potential confounding effects, we controlled for several participant and task-level variables. Participant medical knowledge was assessed through a pre-study survey to account for variations in domain expertise. We also included a cognitive bias assessment prior to the taskt to measure predispositions influencing AI trust . In addition, all medical cases used in the study were pre-curated and standardized in complexity to ensure that case difficulty did not independently affect participant responses.

This structured study design ensures that any differences in trust, preference, and confidence are directly influenced by the AI-generated diagnoses rather than external factors. By carefully controlling these variables, we aim to provide clear and reliable insights into how users navigate AI-assisted medical decision-making.

3.4 Study Task

In this study, participants took on the role of everyday users trying to make sense of medical information with help from two AI systems – BioGPT, LLM trained on biomedical literature, and Gemini 1.5 Pro, a more general VLM. They went through different types of medical situations, like describing symptoms, uploading pictures of skin issues or wounds, or looking at medical images such as X-rays or CT scans. For each case, BLIP-2 was used to generate initial image-based interpretations, which were then followed by Gemini and BioGPT explanations based on those visual cues.

Sometimes, the two AIs agreed [32], giving the same diagnosis, and other times they disagreed (mostly when dealing with image input), showing completely different opinions. Participants will need to look at both, decide which AI they trust more, and share how confident they feel about their final choice. To make the experience feel more natural, and a bit more like real-life decision-making, there may also be short, unrelated tasks mixed in to create small distractions, just like the interruptions people face in everyday life. In some cases, participants might even be told that both AIs are equally reliable, even though the goal is to see how they handle unexpected disagreement [20]. At the end of the study, participants will complete a post-task questionnaire, where they will reflect on their experiences, trust perceptions, and overall confidence in AI-assisted medical decision-making.

All of this helps us understand not just which AI people prefer, but also how they build (or lose) trust when the answers aren't so clear-cut.

3.5 Study Procedure

The procedure is divided into three phases:

3.5.1 Pre – Study Setup: Participants began by reading a brief introduction explaining the study's purpose and their role in comparing two AI models for medical diagnoses. After reviewing the information, they signed a consent form. Following this, they completed a background survey to provide demographic information, educational background, medical knowledge (if any), and prior experience with AI health tools. To familiarize themselves with the interface and evaluation process, participants then proceeded through a practice session in which they reviewed a sample case, observed diagnostic outputs from both AI models, and practiced selecting the model they trusted more while rating their confidence in that decision.

Unpublished working draft. Not for distribution.

3.5.2 *AI evaluation phase:*

Participants have worked through a series of medical cases involving either entering symptoms, uploading images of visible conditions (such as wounds or rashes), or reviewing medical imagery such as X-rays, CT scans, or pathology slides. For each case, when an image was provided as input, diagnostic explanations were first generated by BLIP-2, which extracted relevant visual information. These outputs were then interpreted by Gemini and BioGPT to produce textual diagnostic responses. In some cases, the models provided consistent diagnoses, while in others, they offered conflicting outputs that differed in either conclusion or reasoning. After reviewing both responses, participants selected the model they trusted more and rated their confidence in their chosen diagnosis.

To simulate real-world decision-making environments, short unrelated tasks were occasionally introduced between cases to add cognitive load, mimicking real-life interruptions. In select cases, participants were told that both AI models were equally reliable, even when their outputs disagreed, to study how users responded to conflicting yet ostensibly credible advice.

3.5.3 *Post – Study Reflection:*

After completing all diagnostic cases, participants filled out a post-task questionnaire to reflect on their experience. They were asked to indicate which AI model they trusted more overall, how confident they felt in their final decisions, and what factors influenced their choices, such as explanation clarity, reasoning quality, or presentation order. Following the questionnaire, participants were debriefed about the study’s purpose and any experimental conditions that were intentionally varied. They were then thanked for their participation and dismissed.

3.6 Measures and Analyses

To evaluate user trust, decision-making, and confidence in AI-generated medical diagnoses, this study incorporates quantitative and qualitative measures.

3.6.1 *Quantitative Measurement:*

Participants’ trust in AI-generated diagnoses was measured using a 5-point Likert scale, ranging from "Not at all trustworthy" to "Extremely trustworthy." This allowed us to assess how reliable and credible participants found each AI response to be. Along with this, they also indicated which AI model they relied on more for their final decision. This preference provided insight into whether people favored a specialized medical AI like BioGPT or a more general-purpose model like Gemini when making medical judgments.

To investigate anchoring bias,participants were asked to rate their confidence after reading the first diagnosis before being shown the second. They then rated their final confidence level again after reviewing both AI responses. If confidence ratings remained higher for the first diagnosis, it suggested that participants were sticking to their initial impression rather than fairly considering both options. Finally, participants rated how confident they felt in their final decision. This helped us determine whether agreement between AI models makes people more certain or if conflicting responses leave them feeling uncertain.

Additionally, to assess how AI-expressed confidence affects user trust, the system randomly varied whether each model’s confidence estimate was shown. When shown, confidence will be conveyed using a one-line summary (e.g., “confidence score: 8/10”). This randomized presentation allows us to determine whether the presence of a confidence estimate, rather than just its content, influences perceived trustworthiness. Even when an explicit score is not shown, participants were asked to rate how confident the AI appeared to be in its own explanation. These ratings will help us explore the relationship between perceived AI confidence and human trust.

3.6.2 *Qualitative Measurement:*

Participants also provided open-ended explanations about why they trusted one AI model over the other, particularly in cases where the diagnoses conflicted. These qualitative reflections offer deeper insight into the cognitive factors shaping trust, such as explanation clarity, medical reasoning, or model communication style. For instance, participants were prompted with questions such as: *“Which AI explanation did you find clearer and easier to understand, and why?”* and *“Did you find the AI’s reasoning logically structured, and were elements like step-by-step explanations or medical evidence important in your decision?”*

Participants also described how they perceived the tone and style of each AI model by responding to questions like *“How would you characterize the communication style of each AI (e.g., authoritative, cautious, professional), and did that influence your trust?”* Their responses will be analyzed thematically to identify common traits that impact users’ perceptions, including clarity, logic, tone, and the presence of uncertainty.

Participants also provided open-ended explanations about why they trusted one AI model over the other, particularly in cases where the diagnoses conflicted. These qualitative reflections helped us understand the factors influencing their trust decisions, such as the clarity and comprehensibility of AI-generated explanations. For instance, participants were asked *"Which AI explanation did they find clearer and easier to understand, and why?"* to directly assess the clarity of the information provided. Participants also reflected on the reasoning structure employed by each AI, responding to questions like *"Did they find the AI’s reasoning logically structured, and were aspects such as step-by-step explanations or explicit medical evidence important in your decision-making?"* Additionally, they described their perceptions of each AI model’s communication style by answering questions such as *"How would they characterize the communication style of each AI model (for example, authoritative, cautious, friendly, or professional), and did this style affect their trust?"* We carefully analyzed these responses by identifying common themes related to clarity of explanation, coherence and logical structure of the reasoning, and perceived AI personality. This thematic analysis will provide deeper insights into how these specific characteristics shape users’ trust and their final diagnostic decisions.

3.6.3 *Data Analysis:*

The quantitative data collected from Likert-scale ratings (trust, model preference, confidence scores) will be analyzed using descriptive statistics (mean, standard deviation) to summarize user trends. To compare how trust and confidence change depending on whether the AI models give the same or different diagnoses, we will use a Repeated Measures ANOVA. Additionally, we will use t-tests to check if people’s confidence changes after seeing both AI responses, and chi-square tests to see if one

Unpublished working draft.
Not for distribution.

AI model is consistently chosen over the other when their answers don't match.

For the qualitative data, we are planning on using thematic analysis following Braun & Clarke's method [4]. We will look for common themes, such as why participants trusted one AI model more, how clear they found the explanations, and how they made their final decisions. We may also use sentiment analysis to see if participants had generally positive, neutral, or negative reactions to the AI-generated explanations.

References

[1] 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

[2] Taslima Akter, Manohar Swaminathan, and Apu Kapadia. 2024. Toward Effective Communication of AI-Based Decisions in Assistive Tools: Conveying Confidence and Doubt to People with Visual Impairments at Accelerated Speech. In *Proceedings of the 21st International Web for All Conference* (Singapore, Singapore) *(W4A '24)*. Association for Computing Machinery, New York, NY, USA, 177–189. doi:10.1145/3677846.3677862

[3] 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] https://arxiv.org/abs/2311.05591

[4] David Byrne. 2022. A worked example of Braun and Clarke's approach to reflexive thematic analysis. *Quality Quantity* 56 (06 2022). doi:10.1007/s11135-021-01182-y

[5] A. CHAPANIS. 1967. The Relevance of Laboratory Studies to Practical Situations. *Ergonomics* 10, 5 (1967), 557–577. doi:10.1080/00140136708930910 arXiv:https://doi.org/10.1080/00140136708930910 PMID: 6074338.

[6] Davide Cirillo and María José Rementeria. 2022. Chapter 3 - Bias and fairness in machine learning and artificial intelligence. In *Sex and Gender Bias in Technology and Artificial Intelligence*, Davide Cirillo, Silvina Catuara-Solarz, and Emre Guney (Eds.). Academic Press, 57–75. doi:10.1016/B978-0-12-821392-6.00006-6

[7] Yujiro Otsuka Kouhei Kawakami Naoki Koishi Ken Oba Toru Bando Masaki Matsusako Yasuyuki Kurihara Daisuke Yamada, Fumitsugu Kojima. 2024. Multimodal modeling with low-dose CT and clinical information for diagnostic artificial intelligence on mediastinal tumors: a preliminary study. *BMJ Open Respiratory Research* 11, e002249 (2024). doi:10.1136/bmjresp-2023-002249

[8] Paul Denny, Hassan Khosravi, Arto Hellas, Juho Leinonen, and Sami Sarsa. 2023. Can We Trust AI-Generated Educational Content? Comparative Analysis of Human and AI-Generated Learning Resources. arXiv:2306.10509 [cs.HC] https://arxiv.org/abs/2306.10509

[9] Ophelia Deroy, Davide Bacciu, Bahador Bahrami, Cosimo Della Santina, and Sabine Hauert. 2024. Shared Awareness Across Domain-Specific Artificial Intelligence: An Alternative to Domain-General Intelligence and Artificial Consciousness. *Advanced Intelligent Systems* 6, 10 (2024), 2300740. doi:10.1002/aisy.202300740 arXiv:https://advanced.onlinelibrary.wiley.com/doi/pdf/10.1002/aisy.202300740

[10] Murat Dikmen and Catherine Burns. 2022. The effects of domain knowledge on trust in explainable AI and task performance: A case of peer-to-peer lending. *Int. J. Hum.-Comput. Stud.* 162, C (June 2022), 11 pages. doi:10.1016/j.ijhcs.2022.102792

[11] Moran Gendler, Girish N Nadkarni, Karin Sudri, Michal Cohen-Shelly, Benjamin S Glicksberg, Orly Efros, Shelly Soffer, and Eyal Klang. 2024. Large Language Models in Cardiology: A Systematic Review. *medRxiv* (2024). doi:10.1101/2024.09.01.24312887 arXiv:https://www.medrxiv.org/content/early/2024/09/01/2024.09.01.24312887.full.pdf

[12] Eliya Habba, Ofir Arviv, Itay Itzhak, Yotam Perlitz, Elron Bandel, Leshem Choshen, Michal Shmueli-Scheuer, and Gabriel Stanovsky. 2025. DOVE: A Large-Scale Multi-Dimensional Predictions Dataset Towards Meaningful LLM Evaluation. doi:10.48550/arXiv.2503.01622

[13] Lili He, Hailong Li, Ming Chen, Jinghua Wang, Mekibib Altaye, Jonathan R. Dillman, and Nehal A. Parikh. 2021. Deep Multimodal Learning From MRI and Clinical Data for Early Prediction of Neurodevelopmental Deficits in Very Preterm Infants. *Frontiers in Neuroscience* 15 (2021). doi:10.3389/fnins.2021.753033

[14] Jiaxing Huang and Jingyi Zhang. 2024. A Survey on Evaluation of Multimodal Large Language Models. arXiv:2408.15769 [cs.CV] https://arxiv.org/abs/2408.15769

[15] Shih-Cheng Huang, Malte Jensen, Serena Yeung-Levy, Matthew P. Lungren, Hoifung Poon, and Akshay S Chaudhari. 2024. Multimodal Foundation Models for Medical Imaging - Systematic Review and Implementation Guidelines. *medRxiv* (2024). doi:10.1101/2024.10.23.24316003

[16] Md Saiful Islam, Tariq Adnan, Jan Freyberg, Sangwu Lee, Abdelrahman Abdelkader, Meghan Pawlik, Cathe Schwartz, Karen Jaffe, Ruth B. Schneider, E Ray Dorsey, and Ehsan Hoque. 2024. Accessible, At-Home Detection of Parkinson's Disease via Multi-task Video Analysis. arXiv:2406.14856 [cs.CV] https://arxiv.org/abs/2406.14856

[17] Zifan Jiang, Salman Seyedi, Emily Griner, Ahmed Abbasi, Ali Bahrami Rad, Hyeokhyen Kwon, Robert O. Cotes, and Gari D. Clifford. 2023. Multimodal mental health assessment with remote interviews using facial, vocal, linguistic, and cardiovascular patterns. *medRxiv* (2023). doi:10.1101/2023.09.11.23295212 arXiv:https://www.medrxiv.org/content/early/2023/09/12/2023.09.11.23295212.full.pdf

[18] Julius Keyl, Philipp Keyl, Grégoire Montavon, René Hosch, Alexander Brehmer, Liliana Mochmann, Philipp Jurmeister, Gabriel Dernbach, Moon Kim, Sven Koitka, Sebastian Bauer, Nikolaos Bechrakis, Michael Forsting, Dagmar Führer-Sakel, Martin Glas, Viktor Grünwald, Boris Hadaschik, Johannes Haubold, Ken Herrmann, Stefan Kasper, Rainer Kimmig, Stephan Lang, Tienush Rassaf, Alexander Roesch, Dirk Schadendorf, Jens T. Siveke, Martin Stuschke, Ulrich Sure, Matthias Totzeck, Anja Welt, Marcel Wiesweg, Hideo A. Baba, Felix Nensa, Jan Egger, Klaus-Robert Müller, Martin Schuler, Frederick Klauschen, and Jens Kleesiek. 2023. Decoding pan-cancer treatment outcomes using multimodal real-world data and explainable artificial intelligence. *medRxiv* (2023). doi:10.1101/2023.10.12.23296873 arXiv:https://www.medrxiv.org/content/early/2023/10/19/2023.10.12.23296873.full.pdf

[19] Hana Kopecka, Jose Such, and Michael Luck. 2024. Preferences for AI Explanations Based on Cognitive Style and Socio-Cultural Factors. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 109 (April 2024), 32 pages. doi:10.1145/3637386

[20] 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

[21] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. arXiv:2301.12597 [cs.CV] https://arxiv.org/abs/2301.12597

[22] Jingshu Li, Yitian Yang, Renwen Zhang, and Yi chieh Lee. 2024. Overconfident and Unconfident AI Hinder Human-AI Collaboration. arXiv:2402.07632 [cs.AI] https://arxiv.org/abs/2402.07632

[23] 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/article-pdf/23/6/bbac409/47144271/bbac409.pdf bbac409.

[24] 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

[25] J. C. Pichel M. Fernández-Pichel and D. E. Losada. 2025. Evaluating search engines and large language models for answering health questions. *npj Digital Medicine* 8 (2025), 153. doi:10.1038/s41746-025-01546-w

[26] Miquel Miró-Nicolau, Gabriel Moyà-Alcover, Antoni Jaume i Capó, Manuel González-Hidalgo, Maria Gemma Sempere Campello, and Juan Antonio Palmer Sancho. 2024. To Trust or Not to Trust: Towards a novel approach to measure trust for XAI systems. arXiv:2405.05766 [cs.CV] https://arxiv.org/abs/2405.05766

[27] Masao Noda, Hidekane Yoshimura, Takuya Okubo, Ryota Koshu, Yuki Uchiyama, Akihiro Nomura, Makoto Ito, and Yutaka Takumi. 2024. Feasibility of Multimodal Artificial Intelligence Using GPT-4 Vision for the Classification of Middle Ear Disease: Qualitative Study and Validation. *JMIR AI* 3 (31 May 2024), e58342. doi:10.2196/58342

[28] Jidong Qi. 2024. Neurophysiological and psychophysical references for trends in supervised VQA multimodal deep learning: An interdisciplinary meta-analysis. *Applied and Computational Engineering* 30 (2024), 189–201. https://doi.org/10.54254/2755-2721/30/20230096

[29] Pedro Bispo Santos and Iryna Gurevych. 2018. Multimodal prediction of the audience's impression in political debates. In *Proceedings of the 20th International Conference on Multimodal Interaction: Adjunct* (Boulder, Colorado) *(ICMI '18)*. Association for Computing Machinery, New York, NY, USA, Article 6, 6 pages. doi:10.1145/3281151.3281157

[30] Max Schemmer, Patrick Hemmer, Maximilian Nitsche, Niklas Kühl, and Michael Vössing. 2022. A Meta-Analysis of the Utility of Explainable Artificial Intelligence in Human-AI Decision-Making. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (Oxford, United Kingdom) *(AIES '22)*. Association for Computing Machinery, New York, NY, USA, 617–626. doi:10.1145/3514094.3534128

[31] Marc Cicero Schubert, Maximilian Lasotta, Felix Sahm, Wolfgang Wick, and Varun Venkataramani. 2023. Evaluating the Multimodal Capabilities of Generative AI in Complex Clinical Diagnostics. *medRxiv* (2023). doi:10.1101/2023.11.01.23297938 arXiv:https://www.medrxiv.org/content/early/2023/11/02/2023.11.01.23297938.full.pdf

[32] Thomas Ströder and Maurice Pagnucco. 2009. Realising deterministic behavior from multiple non-deterministic behaviors. In *Proceedings of the 21st International Joint Conference on Artificial Intelligence* (Pasadena, California, USA) *(IJCAI'09)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 936–941.

Unpublished working draft.
Not for distribution.

[33] Gemini Team. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv:2403.05530 [cs.CL] https://arxiv.org/abs/2403.05530

[34] Aleksandra Urman and Mykola Makhortykh. 2025. The silence of the LLMs: Cross-lingual analysis of guardrail-related political bias and false information prevalence in ChatGPT, Google Bard (Gemini), and Bing Chat. *Telemat. Inf.* 96, C (Feb. 2025), 16 pages. doi:10.1016/j.tele.2024.102211

[35] Koen van Turnhout, Arthur Bennis, Sabine Craenmehr, Robert Holwerda, Marjolein Jacobs, Ralph Niels, Lambert Zaad, Stijn Hoppenbrouwers, Dick Lenior, and René Bakker. 2014. Design patterns for mixed-method research in HCI. In *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational* (Helsinki, Finland) *(NordiCHI '14)*. Association for Computing Machinery, New York, NY, USA, 361–370. doi:10.1145/2639189.2639220

[36] Yu Xiao, Ying Lin, Junji Ma, Jiehui Qian, Zijun Ke, Liangfang Li, Yangyang Yi, Jinbo Zhang, Cam-CAN, and Zhengjia Dai. 2021. Predicting visual working memory with multimodal magnetic resonance imaging. *Human Brain Mapping* 42, 5 (2021), 1446–1462. doi:10.1002/hbm.25305 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/hbm.25305

[37] Junlin Xie, Zhihong Chen, Ruifei Zhang, Xiang Wan, and Guanbin Li. 2024. Large Multimodal Agents: A Survey. arXiv:2402.15116 [cs.CV] https://arxiv.org/abs/2402.15116

[38] Yujiao Zhang, Yunfeng Hu, Ke Li, Xiangjun Pan, Xiaoling Mo, and Hong Zhang. 2024. Exploring the influence of transformer-based multimodal modeling on clinicians' diagnosis of skin diseases: A quantitative analysis. *DIGITAL HEALTH* 10 (2024), 20552076241257087. doi:10.1177/20552076241257087 arXiv:https://doi.org/10.1177/20552076241257087

[39] 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

Unpublished working draft.
Not for distribution.