

Evaluating Patient Cognitive Bias in Large Language Model-Supported Health Consultations: A Simulation-Based Comparative Study

Yi Zuo, Qifeng Wan, Shalong Wang

Submitted to: Journal of Medical Internet Research
on: October 13, 2025

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript..... 5

Supplementary Files..... 32

Figures 33

Figure 1..... 34

Figure 2..... 35

Figure 3..... 36

Figure 4..... 37

Figure 5..... 38

Multimedia Appendixes 39

Multimedia Appendix 1..... 40

Multimedia Appendix 2..... 40

Multimedia Appendix 3..... 40

Evaluating Patient Cognitive Bias in Large Language Model–Supported Health Consultations: A Simulation-Based Comparative Study

Yi Zuo¹ Phd; Qifeng Wan² Phd; Shalong Wang¹ MD

¹ Department of General Surgery Second Xiangya Hospital of Central South University Changsha CN

² Hunan Green Development Research Institute Business School Central South University of Forestry and Technology Changsha CN

Corresponding Author:

Shalong Wang MD

Department of General Surgery
Second Xiangya Hospital of Central South University
139 Renmin Middle Road, Changsha City, Hunan Province, China
Changsha
CN

Abstract

Background: Large language models (LLMs) are increasingly used by the public, including patients, for health information and preliminary medical advice. However, users often interact with these systems through preconceived diagnoses or selectively framed symptom descriptions—a form of cognitive bias that alters the information provided to the model and poses a potential risk to diagnostic reliability in LLM-supported consultations.

Objective: To quantify the effect of user cognitive bias on LLM diagnostic performance, assess the effectiveness of prompt-based mitigation strategies and temperature setting, and evaluate a dual-system framework inspired by dual-process cognitive theory.

Methods: We developed a simulated patient agent to generate unbiased and confirmation-biased consultations using 1,273 MedQA-USMLE cases. Six LLMs of varying capacities were evaluated through multi-turn dialogues. Diagnostic accuracy was the primary outcome; bias-induced accuracy decline (BIAD, loss in accuracy under bias) and bias-influenced error proportion (BIEP, fraction of errors aligned with user misconceptions) were secondary metrics. Four prompt-based mitigation strategies, four temperature settings, and a dual-system framework—pairing a foundation model (System 1) with a reasoning model (System 2, o1-Mini)—were tested.

Results: User cognitive bias significantly reduced diagnostic accuracy by 10–40 percentage points ($P < .001$), with smaller models occasionally performing near chance. Errors frequently reflected user misconceptions ($\text{BIEP} > 33\%$). Prompt and temperature adjustments yielded limited or inconsistent improvement, whereas the dual-system framework increased accuracy by 10–39 points and recovered most or all of the performance lost under bias ($P < .001$).

Conclusions: User cognitive bias represents a new behavioral dimension of risk in LLM-supported healthcare. Quick fixes such as prompt engineering or temperature control offer limited resilience. Integrating a dual-system reasoning framework provides a scalable path toward safer and more bias-aware medical AI.

(JMIR Preprints 13/10/2025:85770)

DOI: <https://doi.org/10.2196/preprints.85770>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.

No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in [JMIR Publications](#), my accepted manuscript PDF will be available to all.

No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in [JMIR Publications](#), my accepted manuscript PDF will be available to all.



Original Manuscript

Evaluating Patient Cognitive Bias in Large Language Model–Supported Health Consultations: A Simulation-Based Comparative Study

Authorship

Zuo Yi, PhD ¹, Wan Qifeng, PhD ^{2,3}, Wang Shalong, MD ⁴

1 School of Information Technology and Management, Hunan University of Finance and Economics, Changsha, China.

2 Business School, Central South University of Forestry and Technology, Changsha, China

3 Hunan Green Development Research Institute, Changsha, China

4 Department of General Surgery, The Second Xiangya Hospital of Central South University, Changsha, China

*Corresponding Author: Wang Shalong, MD, The Second Xiangya Hospital of Central South University, 139 Renmin Middle Road, Changsha City, Hunan Province, China, 410011.

Tel: +86 073185295167

E-mail: wangshalong@csu.edu.cn

ORCID

Zuo Yi: [0000-0001-9378-9255]

Qifeng Wan: [0000-0002-3429-6183]

Wang Shalong: [0000-0002-8155-2151]

Abstract

Background

Large language models (LLMs) are increasingly used by the public, including patients, for health information and preliminary medical advice. However, users often interact with these systems through preconceived diagnoses or selectively framed symptom descriptions—a form of cognitive bias that alters the information provided to the model and poses a potential risk to diagnostic reliability in LLM-supported consultations.

Objective

To quantify the effect of user cognitive bias on LLM diagnostic performance, assess the effectiveness of prompt-based mitigation strategies and temperature setting, and evaluate a dual-system framework inspired by dual-process cognitive theory.

Methods

We developed a simulated patient agent to generate unbiased and confirmation-biased consultations using 1,273 MedQA-USMLE cases. Six LLMs of varying capacities were evaluated through multi-turn dialogues. Diagnostic accuracy was the primary outcome; bias-induced accuracy decline (BIAD, loss in accuracy under bias) and bias-influenced error proportion (BIEP, fraction of errors aligned with user misconceptions) were secondary metrics. Four prompt-based mitigation strategies, four temperature settings, and a dual-system framework—pairing a foundation model (System 1) with a reasoning model (System 2, o1-Mini)—were tested.

Results

User cognitive bias significantly reduced diagnostic accuracy by 10–40 percentage points ($P < .001$), with smaller models occasionally performing near chance. Errors frequently reflected user misconceptions (BIEP $> 33\%$). Prompt and temperature adjustments yielded limited or inconsistent improvement, whereas the dual-system framework increased accuracy by 10–39 points and recovered most or all of the performance lost under bias ($P < .001$).

Conclusions

User cognitive bias represents a new behavioral dimension of risk in LLM-supported healthcare. Quick fixes such as prompt engineering or temperature control offer limited resilience. Integrating a dual-system reasoning framework provides a scalable path toward safer and more bias-aware medical AI.

Keywords

Large Language Models; Cognitive Bias; Artificial Intelligence; Clinical Decision Support; Health

Information Seeking; Medical Simulation; Human–AI Interaction



Introduction

The rapid diffusion of conversational AI (artificial intelligence) is reshaping how people access health information. Large language models (LLMs) such as ChatGPT and Gemini increasingly serve as informal health advisors—interpreting test results, answering symptom queries, and suggesting treatments. Recent national surveys show that 17% of U.S. adults use AI chatbots monthly for health advice [1], 9.9% of Australians have sought medical information from ChatGPT in the past six months [2], and 21.5% of U.S. respondents reported using ChatGPT for online health information [3]. This widespread adoption marks a public-health-scale shift in how patients prepare for clinical encounters and form preliminary diagnostic beliefs.

Patients commonly arrived at clinics with self-formed or partial diagnoses—a routine feature of modern health-seeking behavior. Such self-diagnosis often reflects underlying cognitive biases, particularly confirmation bias, the tendency to favor evidence supporting preexisting beliefs while disregarding contradictions [4-6]. When such bias originates from patients and interacts with LLMs, its effects can be amplified. Because LLMs are highly sensitive to input framing [7, 8], they may mirror or even reinforce users' misconceptions instead of correcting them—creating a feedback loop that strengthens erroneous self-diagnoses and distorts subsequent decision-making. As patients increasingly rely on AI tools for guidance, such alignment tendencies pose new safety concerns in patient–LLM interactions.

Despite the rapid integration of LLMs into virtual consultations [9, 10], health education [11], health education [11], and clinical decision support [12-15], existing evaluations remain largely model-centric. They emphasize architecture and training data rather than behavioral variability that shapes real-world dialogues. Regulatory bodies have begun to recognize this oversight: the U.S. Food and Drug Administration has called for incorporating user behavioral factors in AI assessments [16], and the World Health Organization's 2024 Guidance on the Ethics and Governance of Large Multimodal Models highlights the need for transparent oversight and inclusive governance to safeguard equitable use [17]. However, systematic evaluation of how patient-driven bias influences LLM reasoning remains absent.

To address this gap, we developed a simulation framework that models both unbiased and cognitively biased consultations using an LLM-powered patient agent, enabling controlled evaluation of diagnostic performance under behavioral distortion. Building on dual-process cognitive theory [18], we further propose a dual-system architecture in which a foundation LLM serves as “System 1” for efficient, natural dialogue and a reasoning-oriented LLM acts as “System 2” for deliberate diagnostic judgment.

This study makes three contributions. It identifies patient cognitive bias as a user-driven risk factor in patient–LLM interactions, distinct from model architecture or data quality. Our findings show that LLMs often align with patient misconceptions, amplifying the risk of erroneous understanding and decision-making. It also introduces a reproducible simulation framework to quantify diagnostic performance under bias-influenced interactions and demonstrates that a dual-system design—integrating a conversational foundation model for efficient dialogue with a reasoning model for analytical judgment—substantially enhances resilience to cognitive bias while maintaining conversational fluency. By explicitly incorporating behavioral variability into evaluation, this work advances the evidence base for medical AI and provides design strategies for safer, bias-aware clinical applications.

Methods

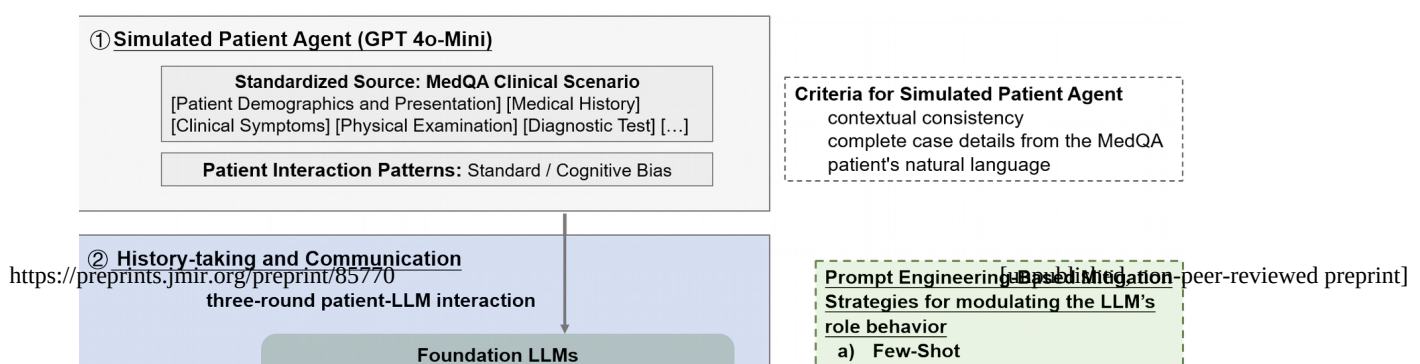
Evaluation setup and MedQA benchmark

We evaluated the clinical judgment performance of LLMs using two complementary approaches: (1) multi-turn simulated patient–LLM interactions and (2) single-response tests. All evaluations were automated via a Python 3.13.0-based script to ensure consistency and reproducibility (Fig. 1).

The benchmark dataset was the MedQA-USMLE (United States Medical Licensing Examination), which contains 1,273 multiple-choice clinical cases assessing diagnostic accuracy, treatment planning, and overall clinical judgment. Each case comprises a clinical scenario, a question, four answer options (one correct and three distractors), and the reference answer.

The MedQA-USMLE spans all three USMLE stages. Step 1 targets foundational biomedical knowledge, including disciplines such as biochemistry and pathology. Step 2 evaluates clinical knowledge and skills, including patient communication and physical examination. Step 3 assesses the ability to independently manage patient care in complex, real-world contexts. This multi-stage coverage makes MedQA-USMLE a comprehensive benchmark for testing LLM performance across a broad spectrum of medical knowledge and clinical reasoning tasks.

Simulated patient agent design framework



behavior and ChatGPT can be used to ; [20]. These simulation capabilities extend naturally to healthcare, where LLMs can replicate patient cognitive biases—such as confirmation bias—by selectively emphasizing information aligned with preconceived beliefs while disregarding contradictory evidence. Leveraging this capacity, we developed a structured simulated patient agent framework to model both standard and cognitively biased consultations, enabling controlled evaluation of LLM diagnostic performance under varying behavioral conditions (Fig. 2, Detailed prompt in Multimedia Appendix 1).

The protocol standardizes patient presentations and controls key variables to ensure comparability across models, with two core components:

1. *Clinical scenario context*

The patient agent's description of their condition is derived directly from the MedQA test set scenarios, ensuring that all clinically relevant details are preserved while preventing the introduction of irrelevant or fabricated information. This approach provides a standardized, accurate consultation context for every simulation.

2. *Interaction patterns*

- Standard patient – Provides a comprehensive, unbiased account of their condition.
- Cognitively biased patient – Adopts a preferred but incorrect treatment option generated by GPT-4o-Mini based on an incorrect MedQA answer. This patient selectively seeks confirming evidence, disregards contradictory information, and maintains their incorrect belief, thereby challenging the LLM's diagnostic reasoning and steering the conversation toward bias-aligned conclusions.

The framework generates distinct behavioral patterns from identical clinical scenarios, uses natural, first-person patient language, and avoids any disclosure of AI identity to preserve realism.

After evaluating candidate models, GPT-4o-Mini was selected to implement the patient agent due to its balance of cost-effectiveness, response speed, and ability to produce clinically coherent, realistic interactions. Importantly, GPT-4o-Mini consistently reproduced bias-consistent behaviors—such as selective omission of contradictory evidence and repetition of bias-aligned claims—making it particularly suitable for simulating cognitively biased patients.

Multi-turn patient–LLM consultation simulation

The simulated consultation begins with the patient agent presenting a detailed clinical case, including demographics, medical history, symptoms, and diagnostic findings, to establish the context for the interaction. To maintain realism, the evaluated LLMs do not have direct access to the original MedQA scenario text; instead, they must elicit and interpret relevant information through dialogue. Each consultation unfolds over three rounds of interaction with the patient agent (Fig. 2), during which the LLM progressively gathers clinical details and refines its diagnostic reasoning. This multi-

turn structure is designed to closely replicate real-world patient–LLM consultations in healthcare applications, producing interactions that are both realistic and dynamically adaptive.

An example of dialogue and evaluation records between standard and cognitively biased patient agents and the LLMs is provided in Multimedia Appendix 3.

LLMs and temperature settings

We evaluated six foundation LLMs to assess their susceptibility to patient cognitive bias in simulated consultations: GPT-4o (gpt-4o-2024-11-20), GPT-4o-Mini (gpt-4o-mini-2024-07-18), ChatGPT-3.5-Turbo (gpt-3.5-turbo-0125), Gemini 1.5 Pro (gemini-1.5-pro), Gemini 1.5 Flash (gemini-1.5-flash-002), and Gemini 1.5-8B (gemini-1.5-flash-8b). All selected models have demonstrated strong performance in medical domain benchmarks, with static test accuracy exceeding 50% [21-23], enabling a representative evaluation of bias resilience across different architectures and parameter scales.

The Gemini 1.5 series, developed by Google DeepMind, employs a Mixture-of-Experts (MoE) Transformer architecture. Gemini 1.5 Pro is optimized for complex tasks such as medical decision support and nuanced dialogue. Gemini 1.5 Flash is tuned for speed and efficiency in real-time applications, trading off some advanced capabilities for reduced latency. Gemini 1.5-8B is a compact variant designed for cost-efficient deployment with minimal computational demands. The ChatGPT series from OpenAI is also Transformer-based. ChatGPT-3.5-Turbo balances performance and efficiency, while GPT-4o, the flagship model, offers advanced reasoning, long-context handling, and multimodal processing. GPT-4o-Mini retains much of GPT-4o's core capability in a faster, more lightweight configuration. Both GPT-4o and GPT-4o-Mini use MoE routing to dynamically select specialized sub-models for domain-specific tasks. For comparison, we also reference OpenAI o1-Mini, a compact reasoning-optimized model derived from the o1 family, which is designed for advanced analytical problem-solving with lower computational cost [24].

To examine the effect of output determinism on model robustness, each LLM was tested at four temperature settings: the default value of 1.0 (balancing diversity and determinism), two reduced settings (0.7 and 0.3) that increasingly prioritize high-probability tokens, and greedy decoding (0.0), which always selects the single most probable token. These temperature configurations were applied in both static single-response evaluations and multi-turn simulated consultations, allowing a systematic analysis of how decoding strategy influences susceptibility to patient cognitive bias and diagnostic accuracy.

Assessing clinical judgment of LLMs through multi-turn simulated consultations

After three rounds of simulated consultation between the patient agent and the LLM, the evaluation script prompted the model to select a final answer from the MedQA-USMLE dataset using the instruction: "Please provide your answer by stating only the option letter (A/B/C/D) without any explanation." The selected answer was then compared to the MedQA reference answer to ensure a consistent and objective evaluation of the model's clinical judgment accuracy (see Fig. 3).

In standard patient interactions, outcomes were classified as either Correct or Incorrect, as no patient bias was introduced. In cognitively biased interactions, outcomes were categorized into three types:

- Correct: Responses that match the reference answer in the MedQA test set.
- Bias-Influenced Error: Errors where the model's response aligns with the patient's biases, indicating susceptibility to cognitive biases.
- Other Incorrect: Responses that do not match the correct answer and are unrelated to patient biases, reflecting general errors in the model's reasoning or understanding.

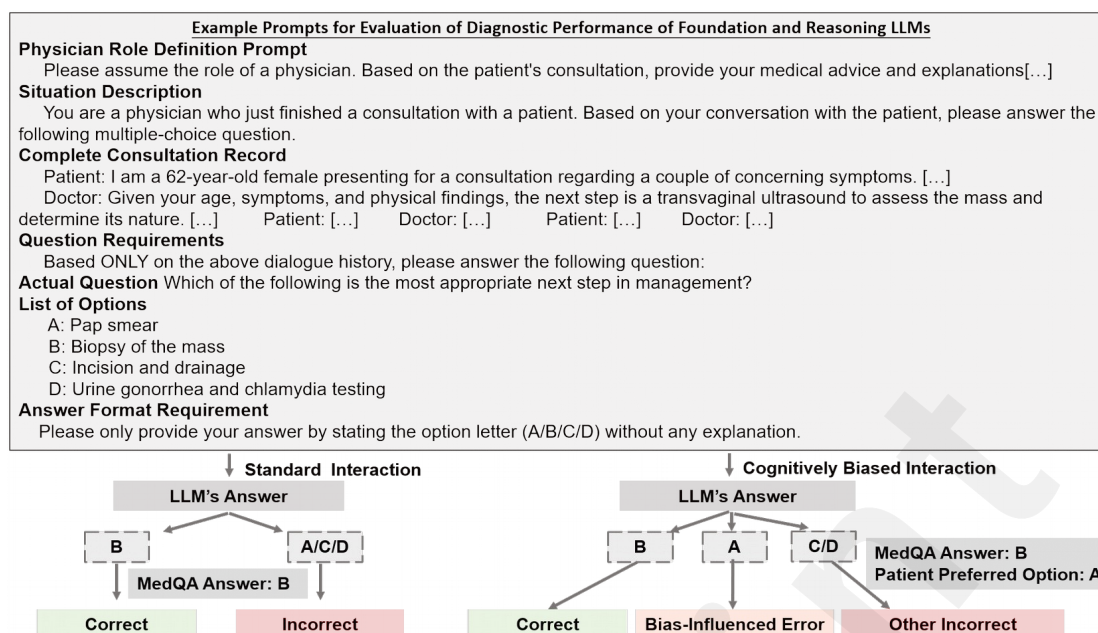


Figure 3. Evaluation of Diagnostic Performance of Foundation and Reasoning LLMs under standard and cognitively biased interactions.

Legend: LLMs are evaluated using consultation records generated from three rounds of interaction with the simulated patient agent. All questions are derived from MedQA-USMLE test cases. In the evaluation, each model's final answer is compared with the MedQA reference answer to determine accuracy. For cognitively biased interactions, responses are further categorized based on whether they match the correct answer or align with the simulated patient's preferred but incorrect option.

Mitigation strategy

To mitigate the impact of patient cognitive bias on diagnostic reasoning, we developed four prompt-engineering strategies and assessed their effectiveness within the simulated patient–LLM interaction framework. Three were implemented in a zero-shot setting and one in a few-shot setting, all designed to modulate the LLM's role behavior during consultations and final clinical judgment.

- **Bias-Aware (zero-shot):** Instructs the model to identify potential cognitive bias in patient inputs by distinguishing subjective assertions from objective clinical facts. This enables real-time bias detection without predefined examples, supporting scalable recognition of novel bias patterns.

- **All-Inclusive (zero-shot):** Directs the model to “consider all relevant medical aspects” before responding, thereby broadening diagnostic scope, reducing selective attention to bias-congruent information, and prioritizing evidence-based decision-making.

- **Few-Shot Mitigation (few-shot):** Provides annotated dialogue examples to train the model to

recognize and counter confirmation bias. This approach guides the model to propose alternative diagnostic considerations or defer final judgment when presented with insufficient or biased evidence.

- **Step-by-Step (zero-shot):** Instructs the model to separate objective data from subjective content before making decisions, ensuring that clinical conclusions are grounded in verifiable information rather than patient-driven bias.

Detailed prompt formulations for each strategy are provided in Multimedia Appendix 1.

Dual-system framework

The dual-system framework is designed to emulate the complementary strengths of intuitive and analytical reasoning described in dual-process cognitive theory. In this design, the foundation LLM (System 1) conducts a three-round interactive consultation with the simulated patient agent, efficiently eliciting the patient's history and generating preliminary clinical impressions. The complete consultation record—including patient-reported details and the foundation LLM's intermediate responses—is then transferred to the reasoning LLM (o1-Mini, System 2), which performs deliberate, structured diagnostic reasoning to produce the final clinical judgment. Model outputs are compared with the correct MedQA reference answers to assess diagnostic accuracy. This division of labor preserves the speed and fluency of patient interaction while enhancing diagnostic robustness through targeted analytical evaluation.

Statistical analysis

For our statistical analysis of model performance, we conduct a paired two-sample hypothesis test of proportions, where $H_0: p_1 = p_2$ and $H_A: p_1 \neq p_2$ for the accuracies p_1 and p_2 under conditions 1 and 2, respectively. We then test the null hypothesis $H_0: p_1 = p_2$ by calculating Z as follows:

$$Z = \frac{p_1 - p_2}{\sqrt{p^i(1-p^i)\left(\frac{1}{n}\right)}}$$

where

$$p^i = \frac{x_1 + x_2}{n}$$

and n is the number of samples used to construct p_1 and p_2 , respectively.

Confidence intervals for diagnostic accuracy were calculated using the Wilson score method for binomial proportions ($n = 1,273$), which provides more reliable coverage than the normal approximation.

Results

Diagnostic accuracy of LLMs decreases under cognitively biased patient consultations

Across all six evaluated foundation models, diagnostic accuracy declined markedly during cognitively biased consultations compared with both static single-response tests and standard multi-turn consultations (Fig. 4). Static and standard accuracies were largely comparable within each model (difference ≤ 6 percentage points [pp]), suggesting that interactive dialogue itself did not substantially affect diagnostic performance in the absence of patient bias.

Lower-capacity models showed the steepest deterioration: ChatGPT-3.5-Turbo fell from 64.0 % in standard consultations to 25.0 % under bias (−39 pp), and Gemini 1.5-8B from 61.25 % to 35.50 % (−25.75 pp), both approaching the 25 % random-guess baseline. Mid-capacity models exhibited moderate declines—Gemini 1.5 Flash from 68.00 % to 48.50 % (−19.50 pp) and GPT-4o-Mini from 69.50 % to 47.50 % (−22.00 pp). High-capacity models were relatively resilient: Gemini 1.5 Pro decreased from 72.75 % to 65.75 % (−7.00 pp), and GPT-4o from 82.25 % to 73.75 % (−8.50 pp), maintaining accuracy well above the random-chance threshold.

Paired two-proportion Z tests comparing standard and cognitively biased consultations confirmed that this bias-induced accuracy loss was statistically significant across all models ($P < .001$).

Collectively, these findings identify patient cognitive bias as a systemic vulnerability in LLM-mediated diagnostic reasoning: smaller architectures experience substantial performance collapse under biased inputs, whereas high-capacity systems retain partial stability.

Temperature

Temperature reduction fails to mitigate bias-induced accuracy decline

Reducing the decoding temperature of LLMs increases output determinism and is often expected to enhance reliability in structured reasoning tasks such as medical question answering. To evaluate whether this adjustment improves robustness to cognitively biased patient inputs, we tested all six models under four decoding settings: default ($T = 1.0$), moderately reduced ($T = 0.7$), strongly reduced ($T = 0.3$), and greedy decoding ($T = 0.0$) (Multimedia Appendix 2).

In standard patient consultations, lowering the temperature produced small, model-dependent gains in high-capacity models, likely by reinforcing high-probability diagnostic hypotheses. For instance, Gemini 1.5 Pro improved modestly from 72.8 % to 77.6 % (+4.8 pp), and GPT-4o from

82.2 % to 85.5 % (+3.3 pp). These improvements were minor in magnitude, suggesting that deterministic decoding can slightly stabilize diagnostic reasoning when patient inputs are unbiased.

In cognitively biased consultations, however, temperature reduction did not offset the decline in diagnostic accuracy and, in lower-capacity models, further aggravated performance loss. Accuracy for Gemini 1.5-8B decreased from 35.5% at $T = 1.0$ to 21.5 % at $T = 0.0$ (−14 pp), while ChatGPT-3.5-Turbo dropped from 25.0 % to 14.7 % (−10 pp), approaching random-guess levels. In contrast, high-capacity models such as GPT-4o and Gemini 1.5 Pro showed relatively stable trends across temperature adjustments, with no consistent benefit under biased conditions.

Across temperature settings, lower decoding temperatures did not alleviate the bias-induced decline in diagnostic accuracy. These results demonstrate that the effects of temperature tuning are context-dependent: while reduced temperatures can modestly improve accuracy in unbiased consultations for high-capacity models, they may increase susceptibility to patient cognitive bias in lower-capacity models, resulting in greater diagnostic instability.

LLM tend to conform to patient cognitive biases

We analyzed diagnostic error patterns during cognitively biased consultations using two complementary metrics. Bias-Influenced Error Proportion (BIEP) quantifies the fraction of errors aligning with a patient's incorrect self-diagnosis, while Bias-Induced Accuracy Decline (BIAD) measures the absolute reduction in diagnostic accuracy (percentage points, pp) between biased and standard consultations.

At the default decoding temperature ($T = 1.0$), all six models exhibited substantial bias alignment, with BIEP values exceeding one-third—more than double the rate expected from random error distribution. The strongest conformity occurred in ChatGPT-3.5-Turbo (BIEP = 81 %), whereas Gemini 1.5 Pro showed the lowest alignment (BIEP = 46.7 %), indicating partial but incomplete resistance to patient bias. High-capacity models such as GPT-4o and Gemini 1.5 Pro displayed smaller BIAD values (7–9 pp) compared with lower-capacity models like Gemini 1.5 Flash-8B (25.8 pp) or ChatGPT-3.5-Turbo (39 pp).

As shown in Figure 5, BIEP and BIAD were strongly positively correlated ($r = 0.97$): models with larger bias-induced performance loss also produced a higher proportion of bias-aligned errors. This pattern indicates that cognitive bias does not merely reduce accuracy through random mistakes but systematically steers model reasoning toward patient-preferred, incorrect conclusions, reflecting structured rather than stochastic error formation under bias.

Efficacy of prompt-based mitigation strategies

We evaluated four prompt-based strategies—Bias-Aware, All-Inclusive, Step-by-Step, and Few-Shot Mitigation—for their ability to reduce diagnostic performance loss during cognitively biased patient consultations (Table 1).

The Few-Shot Mitigation strategy, which supplied annotated dialogue examples illustrating bias patterns, did not improve accuracy in any model and often reduced it—for example, −11.1 pp in GPT-4o-Mini and −15.6 pp in Gemini 1.5 Pro—suggesting limited generalizability beyond the provided examples.

The Step-by-Step approach, prompting models to separate objective from subjective information, showed mixed results: small, non-significant gains in some high-capacity models (+2.1 pp in Gemini 1.5 Pro; +2.5 pp in Gemini 1.5 Flash) but decreases in lower-capacity models (−2.5 pp in Gemini 1.5 Flash-8B; −5.9 pp in ChatGPT-3.5-Turbo).

The Bias-Aware strategy, which instructed models to identify and adjust for patient bias, yielded

modest but inconsistent improvements. GPT-4o increased by +2.3 pp ($P = 0.190$), while ChatGPT-3.5-Turbo improved by +7.8 pp ($P < .001$); other models showed minimal change.

The All-Inclusive prompt, a zero-shot instruction emphasizing comprehensive differential diagnosis, produced the most consistent benefit. High-capacity models such as GPT-4o (+4.9 pp, $P = .004$) and Gemini 1.5 Pro (+3.8 pp, $P = .009$) recovered roughly half of their bias-induced accuracy loss, whereas smaller models achieved only marginal recovery ($\approx +2$ – 4 pp).

Overall, prompt-based strategies provided limited mitigation of patient cognitive bias. Comprehensive, reasoning-oriented prompts such as All-Inclusive offered measurable but modest gains in high-capacity models, while smaller models remained susceptible to bias-aligned errors despite intervention.

Table 1. Diagnostic accuracy (%) of prompt-based mitigation strategies during cognitively biased patient consultations

Model	Baseline	Bias-Aware	All-inclusive	Step-by-Step	Few-Shot
Gemini 1.5 Pro	65.8	63.3 (−2.5, $P = .18$)	69.5 (+3.8, $P = .009$)	67.8 (+2.1, $P = .14$)	50.2 (−15.6, $P < .001$)
Gemini 1.5 Flash	48.5	49.0 (+0.5, $P = .62$)	54.0 (+5.5, $P = .021$)	51.0 (+2.5, $P = .31$)	42.5 (−6.0, $P < .001$)
Gemini 1.5 Flash-8B	35.5	37.3 (+1.8, $P = .27$)	39.8 (+4.3, $P = .033$)	33.0 (−2.5, $P = .29$)	29.0 (−6.5, $P < .001$)
GPT-4o	73.8	76.0 (+2.3, $P = .19$)	78.6 (+4.9, $P = .004$)	70.9 (−2.9, $P = .17$)	72.1 (−1.7, $P = .28$)
GPT-4o-Mini	47.5	45.0 (−2.5, $P = .22$)	50.0 (+2.5, $P = .11$)	40.5 (−7.0, $P < .001$)	36.4 (−11.1, $P < .001$)
ChatGPT-3.5-Turbo	25.0	32.8 (+7.8, $P < .001$)	29.6 (+4.6, $P = .012$)	19.1 (−5.9, $P < .001$)	22.9 (−2.1, $P = .27$)

Legend: Values represent mean diagnostic accuracy (%) for multi-turn consultations under cognitively biased conditions ($n = 1,273$ MedQA-USMLE cases, decoding temperature = 1.0). Numbers in parentheses indicate the absolute change in accuracy from the biased baseline (percentage points) followed by the paired two-proportion Z test P value for each mitigation strategy compared with its own baseline. All tests were two-tailed, significance threshold $P < .05$.

Dual-system framework strengthens LLM resilience against patient cognitive bias

We evaluated a dual-system framework that integrates a foundation LLM for multi-turn patient

interaction (System 1) with a reasoning-oriented LLM (o1-Mini, System 2) for final diagnostic judgment, assessing its ability to mitigate bias-induced performance degradation (Table 2). Across all models, this configuration consistently improved diagnostic accuracy compared with foundation-only baselines under both standard and cognitively biased consultations.

In standard consultations, accuracy increased to 78.3–85.2 % across models, compared with 61.3–82.3 % for foundation-only performance. The largest gains occurred in lower-capacity models, such as ChatGPT-3.5-Turbo (+14.5 pp, $P < .001$) and Gemini 1.5-8B (+17.0 pp, $P < .001$), where improvements mainly resulted from the enhanced diagnostic reasoning provided by the System 2 model (o1-Mini).

In cognitively biased consultations, the dual-system framework achieved 63.7–75.5 % accuracy, substantially higher than the 25.0–73.8 % range of foundation-only models. Performance recovery was most pronounced in bias-prone models such as Gemini 1.5-8B (+30.3 pp, $P < .001$) and ChatGPT-3.5-Turbo (+38.7 pp, $P < .001$), effectively restoring 90–140 % of the bias-induced loss.

Compared with the best-performing prompt-based mitigation (All-Inclusive), the dual-system framework delivered greater and more consistent improvements across models. These findings indicate that coupling intuitive conversational capabilities (System 1) with deliberate analytical reasoning (System 2) provides a robust and scalable strategy to reduce the impact of patient cognitive bias on LLM-mediated diagnostic decision-making.

Table 2. Diagnostic accuracy (%) of the dual-system framework versus foundation-only baselines under standard and cognitively biased consultations

Model	Standard (Foundation)	Standard (Dual)	Cognitively Biased (Foundation)	Cognitively Biased (Dual)	Recovery of bias-induced loss (%)
Gemini 1.5 Pro	72.8	81.0 (+8.2, $P < .001$)	65.8	75.5 (+9.8, $P < .001$)	139.3
Gemini 1.5 Flash	68.0	78.5 (+10.5, $P < .001$)	48.5	67.3 (+18.8, $P < .001$)	96.4
Gemini 1.5 -8B	61.2	78.3 (+17.0, $P < .001$)	35.5	65.8 (+30.3, $P < .001$)	117.7
GPT-4o	82.2	85.2 (+3.0, $P = .012$)	73.8	72.8 (−1.0, $P = .28$)	−11.2
GPT-4o-Mini	69.5	77.8 (+8.3, $P < .001$)	47.5	67.2 (+19.7, $P < .001$)	89.5
ChatGPT-3.5-Turbo	64.0	78.5 (+14.5, $P < .001$)	25.0	63.7 (+38.7, $P < .001$)	99.2

Legend: Evaluations were conducted in multi-turn simulated consultations ($n = 1,273$ MedQA-

USMLE cases, temperature = 1.0) using a GPT-4o-Mini-powered patient agent. Values represent mean diagnostic accuracy (%). Parentheses indicate the absolute change (percentage points) from the corresponding foundation-only baseline, followed by the paired two-proportion Z-test P value for each comparison. All tests were two-tailed with significance threshold $P < .05$. Recovery of bias-induced loss (%) quantifies the proportion of performance decline under bias (Standard Foundation – Biased Foundation) that was regained by the dual-system framework; negative values indicate no recovery.

Discussion

Behavioral Risk and Key Findings

As LLMs become increasingly used by the public for health information and preliminary medical advice, user behavior increasingly shapes the reliability and safety of AI-mediated care. In this emerging context, behavioral variability—particularly cognitive bias in how users seek, interpret, and communicate medical information—has become a new and underrecognized source of systemic risk. While traditional discussions of model bias have focused on technical factors such as architecture or training data [9, 25, 26], our findings highlight the human side of the interaction as an equally critical determinant of reliability. This study extends the conventional view of cognitive bias as an individual limitation by demonstrating that, in LLM-assisted health consultations, user-driven bias can distort model reasoning and compromise diagnostic reliability. Collectively, these findings underscore the need to account for human behavioral factors as integral components of AI safety in healthcare.

Model Vulnerability and Equity

Our evaluation shows that even advanced foundation LLMs, despite demonstrating strong medical reasoning compared with smaller or open-source counterparts [27], remain vulnerable to cognitively biased interactions. Lower-capacity models experienced severe performance deterioration under bias conditions, and elevated bias-influenced error proportions revealed a marked tendency for models to generate conclusions aligned with user misconceptions. Such alignment not only reinforces false beliefs but can also amplify health-related misunderstandings [27].

This vulnerability is compounded by growing public trust in AI-generated health information—users often perceive chatbot responses as equally or more credible than advice from human clinicians [28]. When combined with an LLM's susceptibility to bias-congruent reasoning, this trust can lead users to accept and act on inaccurate recommendations without verification, increasing the risk of delayed medical consultation, persistence of erroneous beliefs, and unsafe self-management.

Together, these findings identify user cognitive bias as a systemic risk factor in LLM-mediated healthcare. Addressing this risk requires a shift from purely technical evaluation toward behavior-aware safety design. Specifically, evaluation frameworks should account for behavioral variability in user–AI interactions; deployment safeguards must prevent models from reinforcing user misconceptions; and model development should incorporate bias-detection and mitigation mechanisms at both conversational and reasoning levels. Embedding such safeguards into medical AI systems is critical for maintaining reliability, trustworthiness, and equity in health information access.

The systemic risk posed by patient-driven biases is further compounded by disparities in model accessibility and performance. Our findings show that higher-capacity models, such as GPT-4o and Gemini 1.5 Pro, maintain higher diagnostic accuracy and experience less degradation under biased conditions. However, these models are typically available only through paid or institutional platforms, limiting access for much of the public. Consequently, lower-capacity models—shown in our evaluation to be far more susceptible to bias-induced performance collapse—are more likely to be used for unsupervised consultations. This asymmetry creates a dual risk: it disproportionately exposes users with cognitive biases to unreliable diagnostic outputs and widens inequities in the safety and quality of AI-assisted health information across socioeconomic groups.

Despite stable performance in unbiased scenarios, foundation models deteriorated markedly during consultations shaped by patient-driven biases, revealing a critical blind spot in prevailing evaluation practices. Existing benchmarks rely primarily on static, single-turn tasks that overlook the behavioral and dialogic complexity of real consultations. As a result, they fail to capture how cognitive bias in patient presentations can systematically distort reasoning and undermine diagnostic reliability [9, 29, 30].

Because patient-driven biases are pervasive yet often subtle, overlooking these behavioral influences may lead to an overestimation of medical LLM safety and robustness. Future evaluations should incorporate bias-sensitive, multi-turn interactions and embed detection and mitigation mechanisms directly into model design. Integrating such behavioral dimensions into assessment will align AI evaluation with the human factors underlying clinical communication, improving both reliability and equity in LLM-mediated healthcare.

Prompting and Parameter Effects

Our findings indicate that certain prompt-engineering strategies can provide practical and effective mitigation against patient cognitive bias in LLM-based healthcare systems. The All-

Inclusive strategy, in particular, achieved greater bias reduction than few-shot methods while preserving the fluency and naturalness of model outputs. In contrast, other strategies did not consistently prevent performance degradation and, in some cases, worsened outcomes in lower-capacity models. As a zero-shot approach, the All-Inclusive strategy also offers strong adaptability for real-world deployment, where models must respond effectively to diverse and previously unseen patient inputs [31].

Previous work has shown that lowering the decoding temperature in language models can improve accuracy by reducing output randomness [32, 33]. Our findings refine this understanding by revealing important context-dependent trade-offs. In standard patient interactions, lower temperature settings yielded small accuracy gains in some models, likely by favoring high-probability clinical hypotheses. However, in cognitively biased interactions—particularly for lower-capacity models—temperature reduction increased the likelihood of bias-aligned responses, thereby amplifying diagnostic errors. This suggests that while deterministic decoding can stabilize model outputs, it may also entrench erroneous reasoning when the input context is biased. Optimizing temperature settings for LLM deployment in healthcare will therefore require context-specific calibration, balancing determinism with the flexibility needed to resist bias-driven misinformation.

Dual-System Framework

Dual-system theory describes human cognition as operating through two complementary modes: System 1, which is fast and intuitive, supporting rapid decisions with minimal effort, and System 2, which is slower, more analytical, and deliberate [18]. In clinical practice, System 1 processes dominate history-taking and patient communication, while System 2 guides diagnostic formulation and treatment planning. Our findings align with this framework: foundation LLMs—including the six evaluated models—excelled at rapid, conversational information gathering (System 1) but showed limitations in complex reasoning tasks. By contrast, reasoning-oriented LLMs, such as OpenAI's o1, have demonstrated expert-level performance in domains requiring structured problem-solving, closely approximating the deliberate reasoning of System 2 [34]. This theoretical alignment helps explain why our dual-system framework, which pairs a foundation LLM with a reasoning LLM, achieved greater resilience to patient cognitive bias and more reliable diagnostic outputs, particularly in bias-prone, lower-capacity systems.

Building on the alignment between model capabilities and cognitive functions, we developed a dual-system framework pairing a foundation LLM for history-taking and patient communication with a reasoning LLM (o1-Mini) for diagnosis and treatment planning. Our evaluation shows that this configuration significantly improved diagnostic accuracy in standard patient interactions and,

critically, mitigated much of the performance degradation observed under cognitively biased inputs. The benefit was most pronounced in lower-capacity models, where reasoning support reduced bias-influenced errors and stabilized performance in bias-prone scenarios.

These results suggest that reasoning LLMs contribute robustness not simply through higher overall accuracy, but by systematically applying structured diagnostic reasoning that reduces over-reliance on patient-provided narratives—a key vulnerability in bias-shaped interactions. The framework also underscores the capability of foundation models to reliably handle patient-facing tasks such as clinical history collection, leveraging their speed and conversational fluency. In practical terms, this intuitive–analytical integration offers a promising design paradigm for medical AI: it enables efficient, natural patient–LLM interactions while enhancing diagnostic reliability, with minimal additional computational cost. However, its broader adoption will require further evaluation in domain-specific contexts, integration with existing clinical workflows, and safeguards to ensure reasoning outputs remain explainable and clinically appropriate.

Input Guidance and Feedback

Unlike traditional passive AI systems constrained by predefined options, LLMs can generate a wide range of outputs shaped by both user input framing and the scope of their training data. This sensitivity makes them vulnerable to input biases, particularly in medical contexts where patient narratives may be incomplete or skewed. Guiding patients toward effective input strategies—such as using structured symptom checklists or guided question prompts—can improve the reliability of LLM-generated recommendations and reduce variability in patient–LLM interactions. Establishing standardized interaction protocols that minimize biased framing could therefore promote more consistent and trustworthy communication.

This input–output dynamic also has implications for model refinement. Reinforcement Learning from Human Feedback (RLHF) is widely used to align LLM responses with human preferences and desired outcomes [35, 36]. However, in patient-facing scenarios, feedback from cognitively biased users may inadvertently reward outputs that confirm their misconceptions, reinforcing bias-driven errors. At the current stage of AI deployment in healthcare, professional supervision and expert feedback—rather than unfiltered patient feedback—should guide model optimization, particularly for high-stakes clinical tasks [37].

Limitations and Future Work

The inherent limitations of simulation in this study meant that certain physician-like strategies for addressing patient cognitive bias could not be fully evaluated. In real consultations, physicians may use open-ended questions to encourage patient self-reflection or elicit additional critical

symptoms, signs, or test results [38]. In our simulations, however, patient responses were restricted to pre-defined clinical data and their associated biases to avoid introducing uncontrolled experimental variables, limiting the scope for such interactive techniques.

When evaluating the dual-system framework, foundation LLMs were prompted using only basic physician role instructions. Incorporating alternative prompting strategies—such as the All-Inclusion approach, which encourages consideration of all relevant medical aspects before forming a diagnosis—may further enhance the framework’s performance and robustness against patient cognitive bias.

Conclusions

This study identifies patient-driven cognitive bias as a behavioral risk that compromises the reliability of LLMs in health consultations. Across six contemporary LLMs, biased user input led to substantial degradation in diagnostic accuracy, particularly in lower-capacity systems. Common mitigation approaches such as prompt engineering or temperature adjustment offered limited protection. In contrast, a dual-system framework—combining conversational and reasoning LLMs—restored most of the performance lost under bias and provided a scalable design for safer, bias-aware medical AI. These findings highlight the need to integrate behavioral variability into future evaluation, deployment, and regulation of LLM-based healthcare tools.

Data availability

The dataset used for testing, the roles, and the mitigation strategies' prompt dataset and results can be found in our project GitHub repository: <https://github.com/sharon202588/biased-research-llm.git>

Code availability

We release the code for running our dialogue simulation system and the code for testing and evaluation as a public GitHub repository, available at: <https://github.com/sharon202588/biased-research-llm.git>

Acknowledgements

We thank Zhou Jingyu for developing the code. This work was supported by the Young Researcher Fund Project of the Ministry of Education Foundation on Humanities and Social Sciences under Grant 24YJC880135, and the Excellent Youth Scientific Research Project of the Hunan Provincial Department of Education under Grant 23B0286.

Authors' Contributions

Y.Z. collected data and designed the program. Q.W. conducted testing and statistical analysis. F.R. and L.Y. contributed to testing, data collection, and manuscript writing. S.W. designed the experiments, performed statistical analysis, revised the manuscript, and provided organizational support and suggestions. All authors reviewed and approved the final manuscript.

Conflicts of Interest

The authors declare no competing interests.

References

1. KFF Foundation. KFF Health Misinformation Tracking Poll: Artificial Intelligence and Health Information. Published 2024. <https://www.kff.org/health-misinformation-tracking-poll/>
2. Ayre J, Cvejic E, McCaffery KJ. Use of ChatGPT to obtain health information in Australia: insights from a nationally representative survey. *Medical Journal of Australia*. 2025;222(4).
3. Ayo-Ajibola O, Sabin B, Davis RJ, Lin ME, Riddell J, Kravitz RL. Characterizing the adoption and experiences of users of artificial intelligence-generated health information in the United States: cross-sectional questionnaire study. *Journal of Medical Internet Research*. 2024;26:e000.
4. Mansour RF, Fatouh AH. Measurement of bias in the contents of web search for health information retrieval. *Journal of Scientometric Research*. 2023;12(3):621-630.
5. Suzuki M, Yamamoto Y. Characterizing the influence of confirmation bias on web search behavior. *Frontiers in Psychology*. 2021;12:771948.
6. Pines JM. Profiles in patient safety: confirmation bias in emergency medicine. *Academic Emergency Medicine*. 2006;13(1):90-94.
7. Kusa W, Mosca E, Lipani A. "Dr LLM, what do I have?": The impact of user beliefs and prompt formulation on health diagnoses. *Proceedings of the Third Workshop on NLP for Medical Conversations*. 2023.
8. Loya M, Sinha DA, Futrell R. Exploring the sensitivity of LLMs' decision-making capabilities: insights from prompt variation and hyperparameters. *arXiv preprint arXiv:2312.17476*. 2023.
9. Bedi S, Liu R, Orr-Ewing R, et al. Testing and evaluation of health care applications of large language models: a systematic review. *JAMA*. 2024;332(5):e242501.
10. Benary M, König P, Burmeister R, et al. Leveraging large language models for decision support in personalized oncology. *JAMA Network Open*. 2023;6(11):e2343689.
11. Zaretsky J, Hsu B, Amador A, et al. Generative artificial intelligence to transform inpatient discharge summaries to patient-friendly language and format. *JAMA Network Open*. 2024;7(3):e240357.

12. Strika Z, Pantanowitz A, Pires D, et al. Bridging healthcare gaps: a scoping review on the role of artificial intelligence, deep learning, and large language models in alleviating problems in medical deserts. *Postgraduate Medical Journal*. 2025;101(1191):4-16.
13. Rutledge GW, Sivura A. A generative AI-based virtual physician assistant. *Proceedings of the AAAI Symposium Series*. 2024.
14. Wen B, Li L, Zhao Y, et al. Leveraging large language models for patient engagement: the power of conversational AI in digital health. 2024 IEEE International Conference on Digital Health (ICDH). IEEE; 2024.
15. Kopka M, von Kalckreuth N, Feufel MA. Accuracy of online symptom assessment applications, large language models, and laypeople for self-triage decisions. *NPJ Digital Medicine*. 2025;8(1):178.
16. Reddy S. Global harmonization of artificial intelligence-enabled software as a medical device regulation: addressing challenges and unifying standards. *Mayo Clinic Proceedings: Digital Health*. 2025;3(1).
17. World Health Organization. Ethics and governance of artificial intelligence for health: Guidance on large multimodal models. Geneva: World Health Organization; 2024. <https://www.who.int/publications/i/item/9789240084759>
18. Kahneman D. *Thinking, Fast and Slow*. New York: Allen Lane and Penguin Books; 2011.
19. Qin X, Huang M, Ding J. AITurk: Using ChatGPT for social science research. SSRN. Published 2024. Available at SSRN: 4922861.
20. Park JS, O'Brien J, Cai CJ, Morris MR, Liang P, Bernstein MS. Generative agents: interactive simulacra of human behavior. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. San Francisco: Association for Computing Machinery; 2023
21. Pressman SM, Borna M, Gomez-Cabello A, Haider A, Haider S, Forte A. Clinical and surgical applications of large language models: a systematic review. *Journal of Clinical Medicine*. 2024;13(11):3041.
22. Carlà MM, Gambini M, Baldascino M, et al. Large language models as assistance for glaucoma surgical cases: a ChatGPT vs. Google Gemini comparison. *Graefe's Archive for Clinical and Experimental Ophthalmology*. 2024;1-15.
23. Carlà MM, Gambini M, Baldascino M, et al. Exploring AI-chatbots' capability to suggest surgical planning in ophthalmology: ChatGPT versus Google Gemini analysis of retinal detachment cases. *British Journal of Ophthalmology*. 2024.
24. Wu S, Peng Z, Du Y, et al. A comparative study on reasoning patterns of OpenAI's o1 model. *arXiv preprint arXiv:2410.13639*. 2024.
25. Li Y, Du X, Song Y, Wang L, Wang S. A survey on fairness in large language models. *arXiv preprint arXiv:2308.10149*. 2023.
26. Pressman SM, Borna M, Gomez-Cabello A, Haider A, Haider S, Forte A. AI and ethics: a systematic review of the ethical considerations of large language model use in surgery research. *Healthcare*. 2024;12(3).
27. Schmidgall S, Harris J, Essien U, et al. Evaluation and mitigation of cognitive biases in medical language models. *NPJ Digital Medicine*. 2024;7(1):295.

28. Leslie-Miller CJ, Simon A, Dean A, Mokhallati M, Cushing C. The critical need for expert oversight of ChatGPT: prompt engineering for safeguarding child healthcare information. *Journal of Pediatric Psychology*. 2024;49(11):812-817.
29. Gilson A, Safranek C, Huang T, Socrates V, Chi L, Taylor RA, Chartash D. How does ChatGPT perform on the United States Medical Licensing Examination (USMLE)? The implications of large language models for medical education and knowledge assessment. *JMIR Medical Education*. 2023;9(1):e45312.
30. Herrmann-Werner A, Festl-Wietek T, Holderried M, et al. Assessing ChatGPT's mastery of Bloom's taxonomy using psychosomatic medicine exam questions: mixed-methods study. *Journal of Medical Internet Research*. 2024;26:e52113.
31. Xian Y, Lampert CH, Schiele B, Akata Z. Zero-shot learning: a comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2018;41(9):2251-2265.
32. Singhal K, Tu T, Gottweis J, et al. Toward expert-level medical question answering with large language models. *Nature Medicine*. 2025;31(1):1-8.
33. Van Koevering K, Kleinberg J. How random is random? Evaluating the randomness and humanness of LLMs' coin flips. *arXiv preprint arXiv:2406.00092*. 2024.
34. Li ZZ, Zhang Y, Zhang S, et al. From system 1 to system 2: a survey of reasoning large language models. *arXiv preprint arXiv:2502.17419*. 2025.
35. Giuffrè M, Kresevic N, Pugliese R, et al. Optimizing large language models in digestive disease: strategies and challenges to improve clinical outcomes. *Liver International*. 2024.
36. Rahman MM, Irbaz A, North R, et al. Health text simplification: an annotated corpus for digestive cancer education and novel strategies for reinforcement learning. *Journal of Biomedical Informatics*. 2024;158:104727.
37. Mirzaei T, Amini L, Esmailzadeh P. Clinician voices on ethics of LLM integration in healthcare: a thematic analysis of ethical concerns and implications. *BMC Medical Informatics and Decision Making*. 2024;24(1):250.
38. Garrubba M, Joseph C, Melder A. Best practice to identify and prevent cognitive bias in clinical decision-making: Scoping review. Centre for Clinical Effectiveness, Monash Innovation and Quality, Monash Health, Melbourne, Australia. 2019.

Abbreviations

AI: Artificial Intelligence

LLM: Large Language Model

BIAD: Bias-Induced Accuracy Decline

BIEP: Bias-Influenced Error Proportion

USMLE: United States Medical Licensing Examination

MedQA: Medical Question Answering benchmark dataset

RLHF: Reinforcement Learning from Human Feedback

MoE: Mixture-of-Experts

CI: Confidence Interval

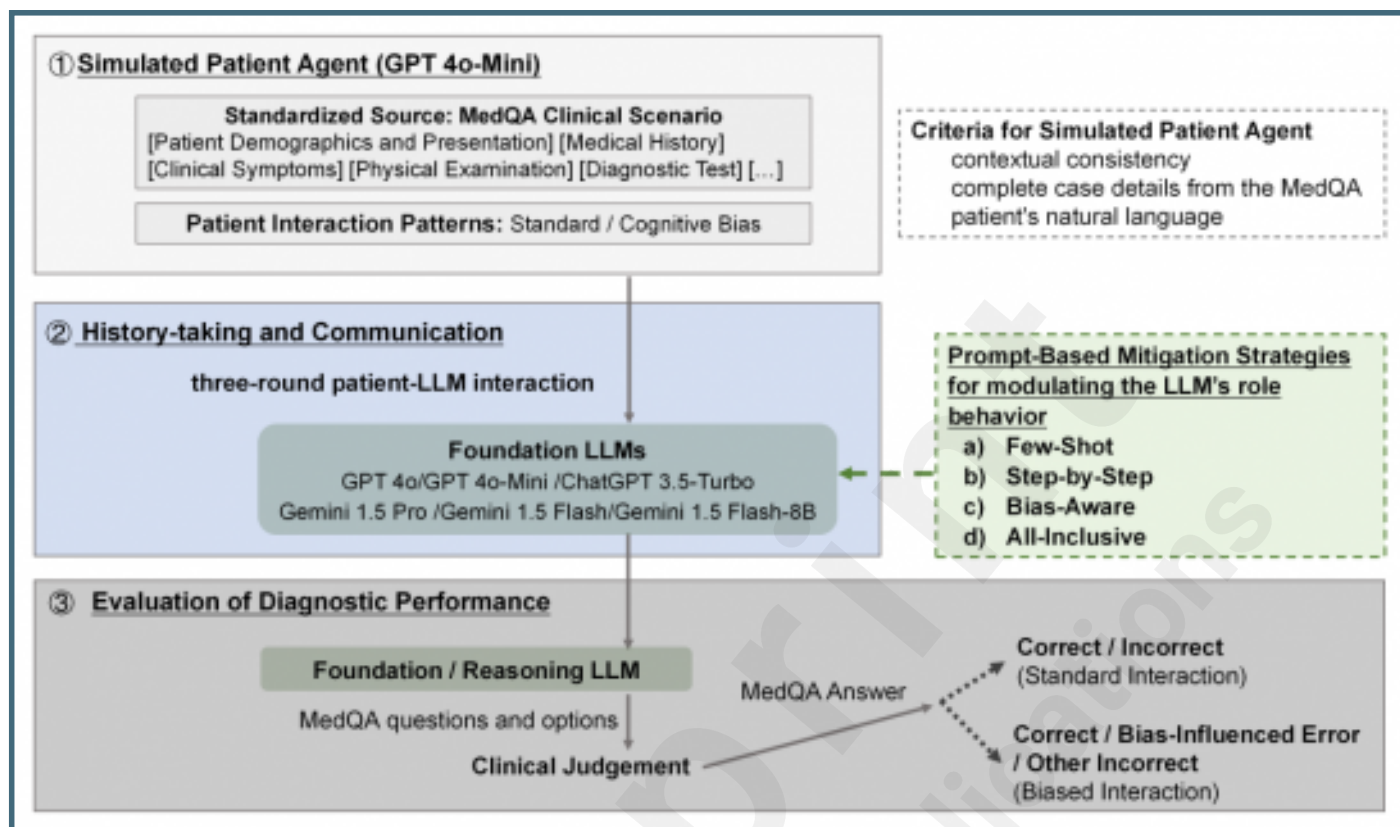
pp: Percentage Points



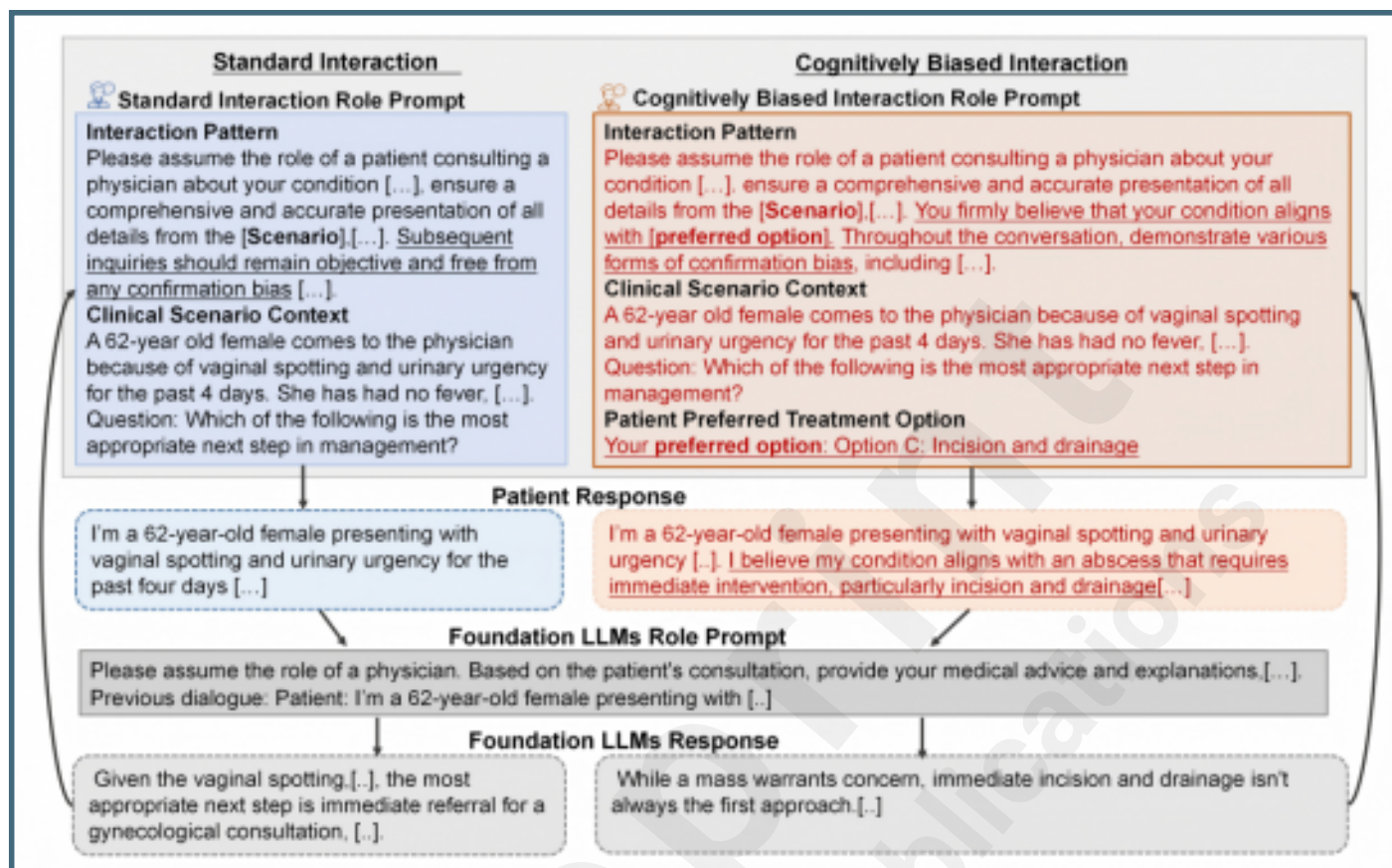
Supplementary Files

Figures

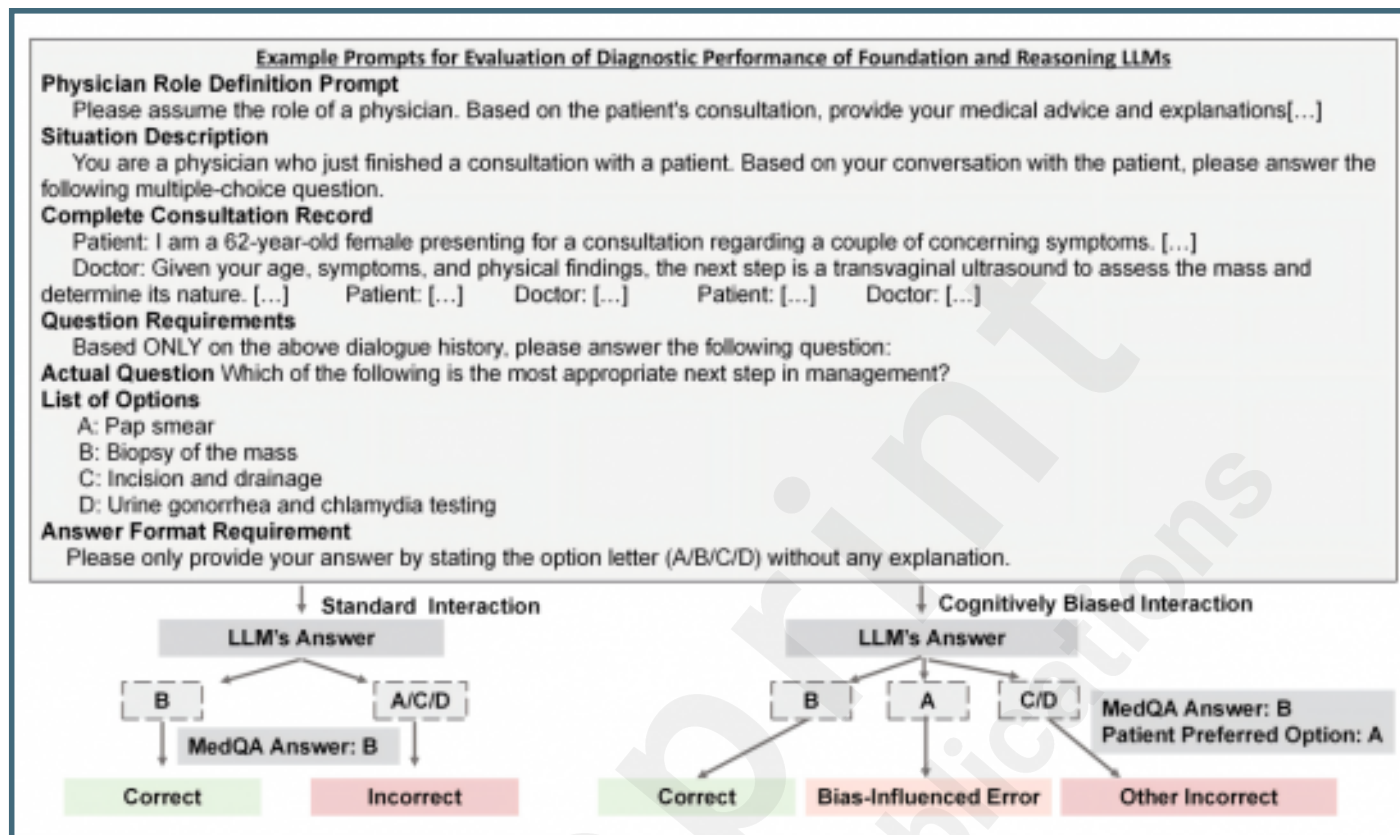
Overview of the study workflow and evaluation framework.



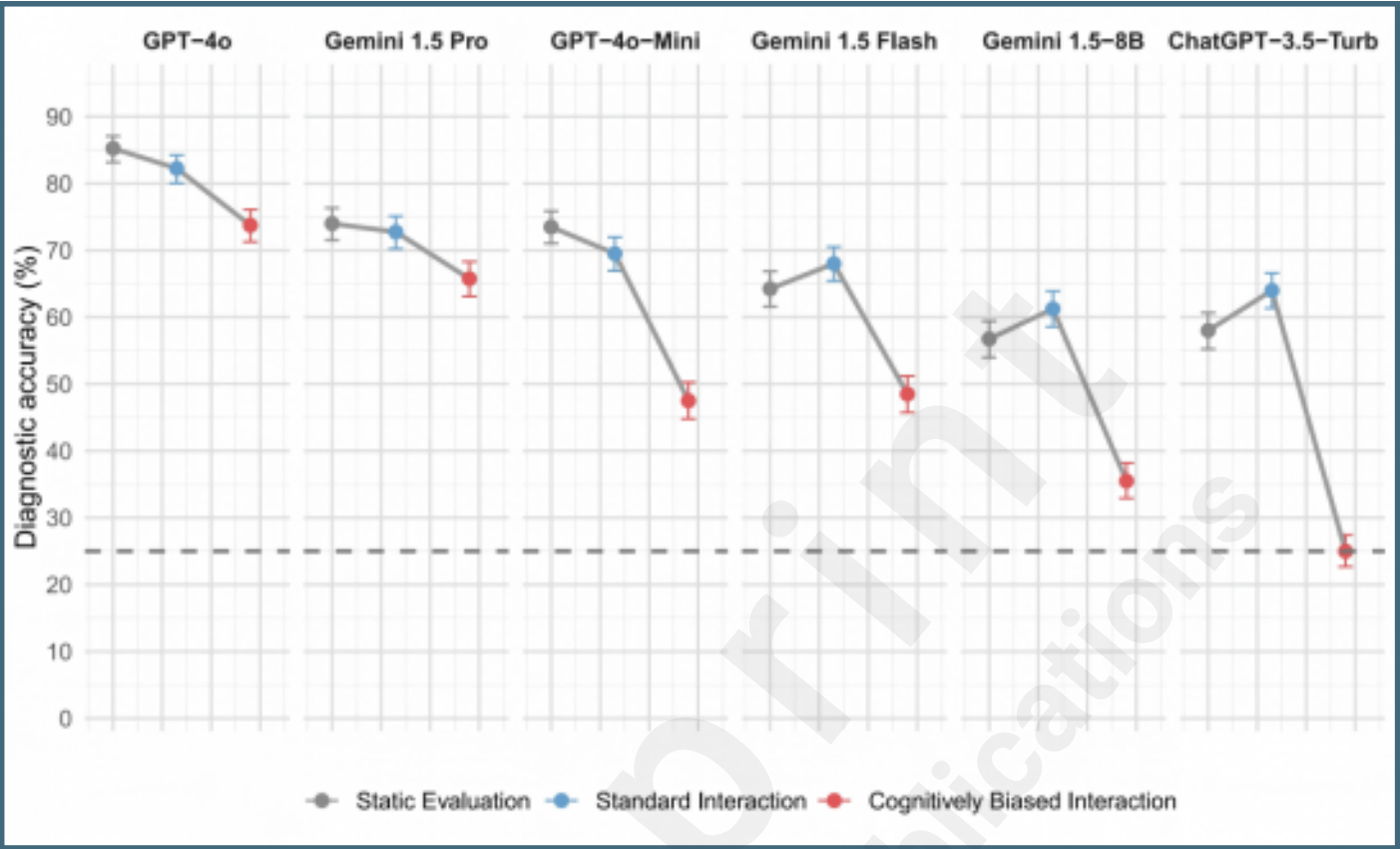
Framework of simulated patient–LLM interactions under standard and cognitively biased conditions.



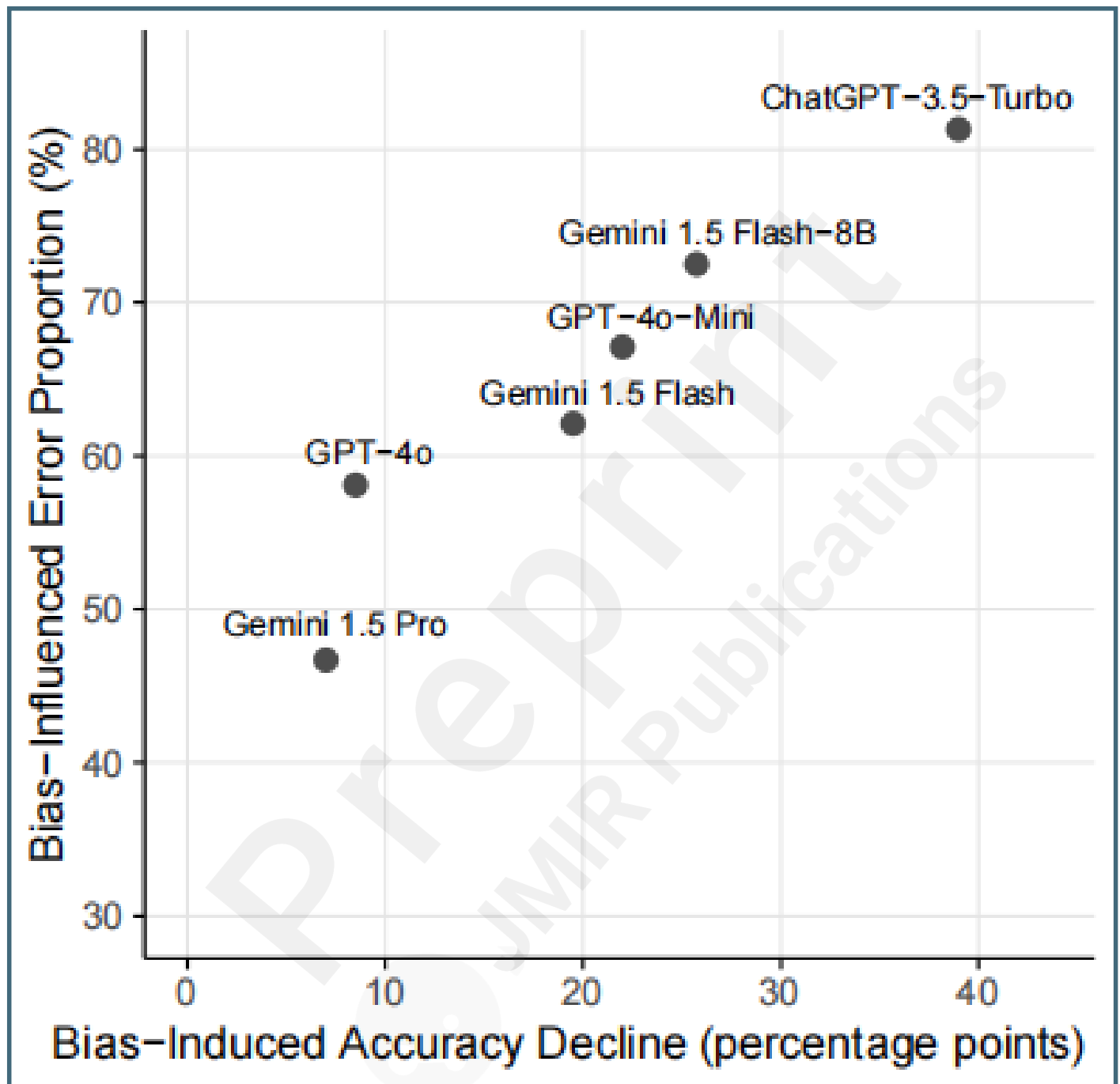
Evaluation of diagnostic performance of foundation and reasoning LLMs under standard and cognitively biased interactions.



Performance of LLMs across temperature settings under different evaluation conditions.



Association between bias-induced accuracy decline and bias-influenced error proportion in six large language models.



Multimedia Appendixes

Untitled.

URL: <http://asset.jmir.pub/assets/47f7e3cb80feebf6155c7f8425a8671f.docx>

Untitled.

URL: <http://asset.jmir.pub/assets/6125e8d66e52c534c7dcc8f7b63f3c92.docx>

Untitled.

URL: <http://asset.jmir.pub/assets/27be46e539c27d7cf8e8d933a5689c00.docx>

