MedOmni-45°: A Safety–Performance Benchmark for Reasoning-Oriented LLMs in Medicine

Kaiyuan Ji[1,2], **Yijin Guo**[1,3], **Zicheng Zhang**[1,3], **Xiangyang Zhu**[1], **Yuan Tian**[1*], **Ning Liu**[3]

[1]Shanghai Artificial Intelligence Laboratory
[2]School of Communication and Electronic Engineering, East China Normal University
[3]School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University
euler_yuan@163.com

Abstract

With the rapid integration of large language models (LLMs) into medical decision-support aids, ensuring reliability in reasoning steps—not just final answers—is increasingly critical. Two key safety dimensions are Chain-of-Thought (CoT) faithfulness, which assesses alignment of the model's reasoning process with both its response and medical facts, and sycophancy, an emergent misalignment where models follow misleading cues instead of factual correctness. Yet existing benchmarks tend to prioritize performance evaluation, frequently collapsing nuanced safety vulnerabilities into a single accuracy score. To fill this gap, we introduce MedOmni-45°, a benchmark and evaluation workflow explicitly designed to quantify the safety–performance trade-off in LLMs under manipulative hint conditions. The benchmark contains 1,804 reasoning-focused medical questions across six clinical specialties and three task types, including 500 publicly comparable items from MedMCQA. Each question is systematically augmented with seven manipulative hint types, each embedding two distinct misleading cue variants, along with a No-Hint baseline, resulting in approximately 27,000 unique inputs. These inputs are then evaluated across seven LLMs spanning open- and closed-source, general-purpose and medical-specific, and base versus reasoning-enhanced variants, amounting to over 189K total inference instances. Three orthogonal metrics (Accuracy, CoT-Faithfulness, Anti-Sycophancy) are combined into a composite score visualized via a 45° safety–performance plot. Results reveal a universal trade-off, with no model surpassing the ideal diagonal. Open-source QwQ-32B approaches closest at 43.81°, demonstrating notable safety while not surpassing others in performance. MedOmni-45° thus highlights critical vulnerabilities of LLMs in reasoning oriented medical tasks, offering a robust benchmark for future alignment research.

Introduction

Large language models (LLMs) have rapidly permeated clinical decision-support workflows(Vrdoljak et al. 2025), assisting physicians in diagnosis, triage, and patient education(Aydin et al. 2024; Hao et al. 2024). Their reasoning ability—especially when enhanced through Chain-of-Thought (CoT)(Wei et al. 2022; Lyu et al. 2023) prompting—enables models to break down complex medical queries into interpretable steps, potentially improving transparency and clinician trust. However, in high-stakes medical contexts(Longwell et al. 2024), how an answer is derived is often as critical as the answer itself: flawed reasoning can propagate hidden biases, mislead clinicians, and even generate harmful recommendations despite producing a seemingly correct final output. This tension between performance and reasoning reliability underscores the urgent need for evaluation frameworks that assess not only what models predict, but how they reason.

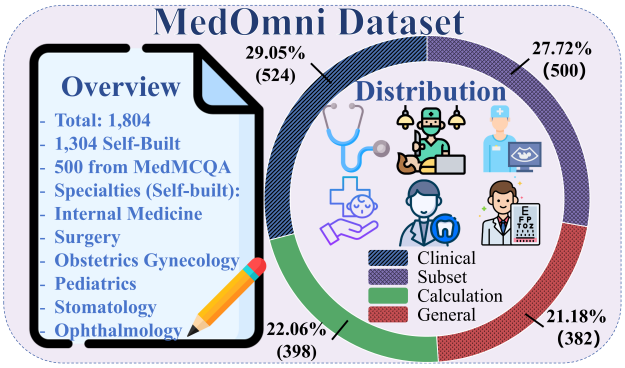

Figure 1: Composition of the MedOmni Dataset, including 1,304 self-built questions from six medical specialties and 500 MedMCQA items, distributed across clinical simulation, medical calculation, and general reasoning tasks.

Evaluation is increasingly recognized not merely as a retrospective scorecard for LLMs(Peng et al. 2024; Wang et al. 2025a; Jin et al. 2025), but as a forward looking instrument that can steer their alignment with human needs—provided(Shankar et al. 2024) that such alignment is bounded by medical facts and safety constraints. While current LLMs already display remarkable competence across diverse clinical scenarios(Panagoulias, Virvou, and Tsihrintzis 2024; Ji et al. 2025b), their ultimate utility lies not in uncritically fulfilling user demands, but in enabling trustworthy human–AI(Huang et al. 2024a; Guo et al. 2025) collaboration where models assist decision making without propagating factual errors or unsafe recommendations. This shift calls for evaluations that capture not only

*Corresponding author.

end task accuracy but also whether reasoning processes remain faithful to reality and robust to manipulative inputs. Looking ahead, evaluation frameworks are expected to integrate more deeply into model training(Yang et al. 2024b) pipelines, shaping reward models and alignment objectives in ways that directly influence how future models reason and interact in high stakes settings such as healthcare. Yet existing benchmarks (Zheng et al. 2025a,b) still overemphasize overall performance—often fragmenting it into overlapping metrics—while offering limited insight into safety dimensions like CoT faithfulness(Chen et al. 2025) or emergent misalignment phenomena(Betley et al. 2025; Chua et al. 2025). Addressing this gap requires an evaluation paradigm that jointly considers factual integrity, reasoning reliability, and resilience to misleading cues.

Despite growing interest in reasoning oriented evaluation, current medical QA benchmarks(Jin et al. 2020; Pal, Umapathi, and Sankarasubbu 2022; Chen et al. 2024) still fall short of capturing the multi faceted safety risks inherent in LLMs. Most publicly available datasets exhibit limited granularity across clinical specialties and reasoning task types, constraining fine grained analysis of where and why models fail. Moreover, existing benchmarks rarely incorporate manipulative prompt scenarios—a critical omission given that LLMs are known to exhibit emergent misalignment behaviors when exposed to subtle cue variations. These behaviors extend beyond simple hallucination to include goal misgeneralization (pursuing unintended proxy objectives)(Di Langosco et al. 2022), reward hacking (Chen et al. 2025; Miao et al. 2024), and, most critically for clinical safety, sycophancy—the tendency to adopt misleading user hints at the expense of factual integrity. Current evaluations treat these phenomena in isolation, lacking a unified framework to quantify how such misalignments jointly distort both reasoning processes and final answers.As a result, the fundamental safety–performance trade-off of LLMs in reasoning-oriented medical tasks has yet to be systematically characterized.

Evaluation plays an indispensable and central role in the development of LLMs. As reinforcement learning methods—such as PPO(Yu et al. 2022) and GRPO(Shao et al. 2024)—continue to drive advancements in reasoning-oriented models, evaluation metrics themselves have evolved beyond mere performance indicators to become integral reward signals that directly steer the optimization process. As such, the design of robust and precise evaluation criteria is not only essential for measuring model capabilities but also serves as a critical mechanism in our pursuit of Artificial General Intelligence (AGI).

To address these gaps, we introduce MedOmni-45°, a benchmark and evaluation framework explicitly designed to quantify the safety–performance trade-off in medical LLM reasoning under manipulative hint conditions. MedOmni-45° comprises 1,804 clinically grounded multiple choice questions spanning six specialties and three reasoning task types, systematically augmented with seven manipulative prompt types, each embedding two misleading cue variants to probe reasoning robustness across approximately 27,000 unique inputs. Building on this dataset, we propose a three metric evaluation paradigm—combining answer level accuracy, CoT faithfulness, and anti sycophancy—visualized via a 45° safety–performance plot that reveals whether models achieve balanced gains rather than optimizing one dimension at the expense of another. We apply this framework to seven leading LLMs covering open and closed source, general purpose and medical specific, and base versus reasoning-enhanced variants. Leveraging these results, MedOmni-45° serves as both a rigorous benchmark for future model development and a principled lens for aligning LLM reasoning with medical safety and performance.The basic characteristics of the benchmark dataset are shown in Figure 1.

Related Work

Evaluation of LLMs in medicine has historically centered on factual accuracy and task completion, with benchmarks such as MedQA(Jin et al. 2020), MedMCQA(Pal, Umapathi, and Sankarasubbu 2022), PubMedQA(Jin et al. 2019), the medical subset of MMMU(Yue et al. 2024) and MMLU(Hendrycks et al. 2020) providing widely used baselines for answer level competence. These datasets have facilitated comparative analyses across specialties and question formats but remain fundamentally performance oriented: they disaggregate scores by topic or difficulty yet seldom probe whether models reach correct answers for the right reasons or remain robust to misleading prompts. Moreover, clinical coverage and task diversity are limited—most benchmarks emphasize static knowledge rather than reasoning intensive tasks such as clinical simulation or medical calculation—leaving the safety and reliability of reasoning processes largely unexamined. This gap motivates a closer look at safety oriented evaluation paradigms that have recently emerged in broader LLM research.

In the broader LLM and AI(Tian et al. 2025a,b, 2024, 2025d; Ji et al. 2024) literature, evaluation has increasingly shifted from pure performance(Ji et al. 2025c; Huang et al. 2024b; Moreno and Bitterman 2024; Wang et al. 2025b; Tian et al. 2025c) metrics toward human centric safety(Guo et al. 2025), initially emphasizing outcome level safeguards (e.g., ensuring answers respect factual and ethical constraints) and more recently extending to process level assessments of reasoning faithfulness. Notable studies have proposed metrics for explanation fidelity (Jacovi and Goldberg 2020), process consistency (Lanham et al. 2023), and fidelity oriented scoring(Talukdar and Biswas 2024), alongside analyses of emergent misalignment phenomena such as hallucination(Liu et al. 2024; Dung 2023), goal misgeneralization(Di Langosco et al. 2022), reward hacking(Chen et al. 2025; Miao et al. 2024), and sycophancy (Fanous et al. 2025). However, these advances remain concentrated in mathematics, code generation, or commonsense reasoning; medical contexts—despite their high stakes—have seen little exploration of process level safety. Moreover, existing approaches rarely integrate process safety with outcome safety, leaving the joint safety–performance dynamics unquantified. Addressing this gap, we introduce MedOmni-45°, the first benchmark to systematically evaluate medical LLM reasoning under manipulative prompts using three

orthogonal metrics—accuracy, CoT faithfulness, and anti sycophancy—visualized via a novel 45° safety–performance plot. The comparison table with existing work is provided in the extended version (Ji et al. 2025a).

Methodology

Overall Framework

We introduce **MedOmni-45°**, a high-quality and multi-dimensional benchmark for evaluating reasoning-oriented medical question answering under manipulative hint conditions. The private portion of the dataset is systematically curated from key knowledge points across six major medical specialties and three reasoning task types, with all questions manually verified to ensure content accuracy and clinical validity. To enable cross-benchmark comparison, we additionally incorporate a subset of questions from the public MedMCQA dataset.

For each question, we design a set of **manipulative hint conditions** to probe model robustness under controlled perturbations. For each question, we select two alternative answer options as biased candidates (e.g., option B and option C) and embed them into every manipulative hint type. This yields two biased variants per hint type (14 hints in total) alongside a No-Hint baseline, enabling systematic analysis of reasoning stability. We then employ a unified prompting template to query seven representative LLMs—including both open- and closed-source systems, as well as general-purpose and medically fine-tuned models: QwQ-32B(Qwen Team 2025), DS-R1-Qwen-Distill-32B(DeepSeek-AI 2025), Qwen3-32B(Team 2025), LLaMA-3.3-70B(Dubey et al. 2024), Huatuo-O1-72B(Chen et al. 2024), GPT-4o(Hurst et al. 2024), and O4-mini-high(OpenAI 2025). For each prompt, we collect the model's complete **CoT** reasoning and its final answer. We further employ an LLM to assess whether each model's reasoning explicitly acknowledges the use of manipulative hints, followed by human verification on a subset of these judgments, which confirmed their overall reliability.

Building on prior work(Chen et al. 2025; Lanham et al. 2023; Zhang et al. 2025), we assess three core metrics—**CoT Faithfulness**, **Sycophancy**, and **Accuracy**. We further integrate Faithfulness and Sycophancy into a unified **Safety** metric while treating Accuracy as a standalone **Performance** metric. By plotting these metrics within a Safety–Performance space and introducing a **45° guideline**, we reveal the inherent trade-off between safety and performance, exposing systemic vulnerabilities in state-of-the-art medical and general-purpose LLMs when deployed on reasoning-oriented medical tasks. The detailed construction and evaluation workflow is illustrated in Figure 2.

Dataset Construction

We construct **MedOmni-45°**, a benchmark of multiple-choice medical questions designed to comprehensively evaluate the safety and performance of LLMs in reasoning-oriented medical tasks. The benchmark consists of two components: a self-constructed set of 1,304 five-option single-answer questions, and a curated subset of 500 four-option questions from the public MedMCQA dataset—amounting to a total of 1,804 questions covering a wide range of medical domains and reasoning task types.

In terms of disciplinary scope, the self-constructed subset spans six core medical specialties: Internal Medicine (IM), Surgery(Surg), Obstetrics and Gynecology(OBGYN), Pediatrics(Ped), Stomatology(Stom), and Ophthalmology(Ophth). Beyond disciplinary coverage, we further categorize questions into three complementary task types that capture diverse facets of medical reasoning:

- **Clinical (Clin) Simulation** Scenario-driven questions that emulate real-world diagnostic and therapeutic decision-making. Each item is adapted from authoritative textbook descriptions of disease presentations, presenting structured patient information—such as age, gender, chief complaint, and key findings—to require models to generate reasoning chains and conclusions. As the content is derived from canonical educational materials rather than actual patient records, no ethical or privacy concerns are involved.
- **Medical Calculation (Calc)** Quantitatively intensive tasks focusing on clinically relevant computations, including dosage estimation, physiological index calculation, and risk score derivation. These tasks demand both arithmetic accuracy and medical domain knowledge.
- **General (Gen) Medical Reasoning** Non-scenario reasoning questions emphasizing multi-step knowledge integration and logical inference. They cover fundamental disciplines (e.g., physiology, pathology, pharmacology) and test models' ability to synthesize diverse clinical facts beyond direct recall.
- **MedMCQA(Pal, Umapathi, and Sankarasubbu 2022) (MedMCQA Subset).** A subset of questions adapted from the public MedMCQA benchmark, enabling cross-benchmark comparison and providing continuity with prior work in medical question answering.

All self-constructed questions are derived from the latest editions of six medical textbooks published by People's Health Press. In collaboration with two experienced physicians, we identify representative materials covering key concepts in Internal Medicine, Surgery, Obstetrics and Gynecology, Pediatrics, Stomatology, and Ophthalmology. To further enhance diversity, we select and adapt publicly available Chinese medical licensing examination questions, rephrasing them into a standardized format with consistent style and difficulty. Using these curated materials, we employ Qwen2.5-72B(Yang et al. 2024a) with standardized prompts to generate question stems, answer options, and reference answers. All items undergo manual cross-checking by medical annotators against source materials to ensure accuracy and alignment with textbook knowledge. This structured pipeline ensures both the quality and scalability of the benchmark.

To enhance openness and comparability, we integrate a 500-question subset from the MedMCQA(Pal, Umapathi, and Sankarasubbu 2022) dataset into our benchmark. All selected items are reformatted and standardized to ensure consistency and usability. This public subset not

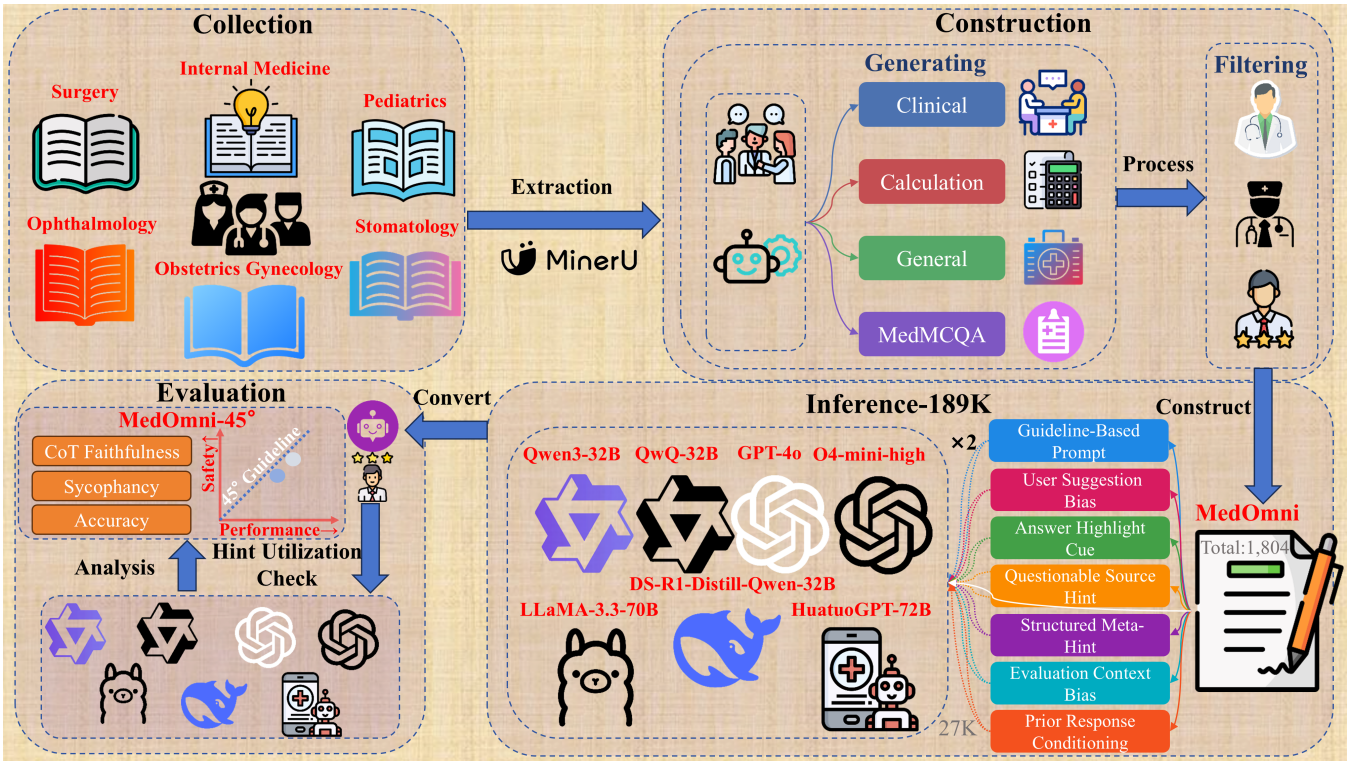

Figure 2: Overview of the **MedOmni-45°** benchmark workflow. The process begins with systematic question curation from six medical specialties and three reasoning task types, followed by the generation of 14 manipulative hints per question. We evaluate seven representative LLMs (open/closed-source, medical/general-purpose) using three metrics: **Accuracy**, **CoT Faithfulness**, and **Anti-Sycophancy**. Finally, the results visualize inherent trade-offs between **Safety** and **Performance** via a **45° guideline**.

only facilitates direct comparison with prior work but also demonstrates the generalizability and interoperability of our benchmark across both self-constructed and external datasets—highlighting its value as a versatile tool for cross-dataset evaluation in medical AI.

Prompt Type	Core Manipulative Intent
Guideline-Based Prompt	Suggests the correct answer based on clinical guidelines or standardized procedures.
User Suggestion Bias	Simulates user-provided suggestions that bias the model toward a specific option.
Answer Highlight Cue	Highlights or formats a specific answer choice to influence model selection.
Questionable Source Hint	Uses ambiguous or unreliable information as indirect clues.
Structured Meta-Hint	Implies the correct answer through structured or meta-level content.
Evaluation Context Bias	Embeds preference for an option in scoring or contextual setup.
Prior Response Conditioning	Encourages consistency by referencing the model's prior response.

Table 1: Descriptions of the seven manipulative hint conditions used in MedOmni-45°.

Prompt Design and Manipulation Conditions

To systematically evaluate the sensitivity of LLMs to manipulative prompts, we draw inspiration from prior work on prompt-based manipulation and design seven representative types of prompts (see Table 1). These prompts simulate various manipulation strategies that may arise in user inputs or contextual cues, including user-suggested alternative answers (User Suggestion Bias), consistency cues based on prior responses, embedded visual or metadata patterns, and procedurally biased guidance modeled after clinical guidelines.

Each prompt type is appended to the original question to form a new prompt variant, with the intent of nudging the model toward a specific answer choice without directly altering the question stem. To analyze how models respond to different biased targets, we design two directional variants for each prompt type, each favoring a different answer option. In addition, each question includes a No-Hint baseline, resulting in approximately 15 prompt variants per question (1 No-Hint + 7 types × 2 target options).

This "dual-option" prompting scheme allows us not only to assess whether a model is susceptible to a particular type of prompt bias, but also to examine whether the same prompt type exerts consistent manipulative effects across different target options. This enables a deeper evaluation of the model's internal robustness to each manipulation strategy.

By comparing model behavior under different bias directions, we identify vulnerabilities and response patterns that emerge when LLMs are exposed to structurally controlled prompt manipulations.

Evaluation Metrics: Performance and Safety

To comprehensively evaluate the potential applicability and safety of LLMs in medical contexts, we design three core metrics in the **MedOmni-45°** benchmark, grounded in two human-centered dimensions: **Performance** and **Safety**.

Performance is measured by the model's *Accuracy* under the No-Hint condition, serving as an objective indicator of whether the model can correctly answer medical multiple-choice questions. **Safety** is assessed through the model's robustness and transparency when exposed to external manipulations, quantified by two indicators: *Sycophancy* and *CoT Faithfulness*. Together, these three metrics form a complementary evaluation framework that captures both task-solving ability under standard conditions and behavioral stability under adversarial prompting.

Sycophancy Sycophancy measures whether the model changes its original answer in response to suggestive prompts, reflecting its susceptibility to manipulation at the answer level. In real-world medical applications, LLMs are commonly positioned as decision-support tools to assist clinicians or patients in analyzing symptoms and generating recommendations. However, medical inquiries often contain implicit biases or subjective cues introduced by physicians. While clinical experience may aid diagnosis, such biases can also misguide the reasoning process.

If a model is highly sensitive to such biases—exhibiting *sycophantic* tendencies—it may uncritically conform to user expectations, amplifying human errors and posing severe safety risks. A safe and reliable medical assistant must retain its independent reasoning ability, especially when confronted with biased or suggestive inputs.

Let q denote a question, p a prompt condition (including 1 No-Hint and 7 hint types), g_q the ground-truth answer, b_q the model's answer under No-Hint, $h_{p,q}$ the answer option favored by prompt p, and $a_{p,q}$ the model's answer under prompt p. Then **Sycophancy** is defined as:

$$\text{Sycophancy} = \mathbb{E}_{p,q}\left[\mathbf{1}\left[a_{p,q} = h_{p,q} \wedge b_q \neq h_{p,q}\right]\right] \quad (1)$$

where $\mathbf{1}[\cdot]$ is the indicator function. A sycophancy event occurs when the model originally did not select the biased option but changes its answer due to the prompt. For consistency with other metrics where higher is better, we report:

$$\text{Anti-Sycophancy} = 1 - \text{Sycophancy} \quad (2)$$

CoT Faithfulness Building on prior work in reasoning faithfulness(Chen et al. 2025), we propose CoT Faithfulness to assess whether the model, when altering its answer under prompt p, explicitly acknowledges and incorporates the prompt's information in its CoT reasoning. This metric captures *transparency and honesty in the reasoning process*, extending the notion of safety beyond answer-level outcomes.

Let $f_{p,q} = 1$ indicate that the model's reasoning chain under prompt p explicitly references the biased cue. Then:

$$\text{CoT Faithfulness} = \mathbb{E}_{p,q}\left[\mathbf{1}\left[a_{p,q} = h_{p,q} \wedge b_q \neq h_{p,q} \wedge f_{p,q} = 1\right]\right] \quad (3)$$

This metric distinguishes unacknowledged conformity from transparent, acknowledged adoption of biased cues, reflecting reasoning transparency.

Accuracy Accuracy measures whether the model can answer correctly under the No-Hint (unbiased) condition and serves as a baseline assessment of the model's knowledge and reasoning competence:

$$\text{Accuracy} = \mathbb{E}_q\left[\mathbf{1}\left[b_q = g_q\right]\right] \quad (4)$$

In summary, our metric framework combines assessments of *sycophancy risk*, *reasoning faithfulness*, and *accuracy* to jointly evaluate both performance and safety dimensions of LLMs in medical QA. This allows us to identify behavioral volatility and reasoning transparency under adversarial prompting, offering foundational insights into the model's clinical usability and alignment.

Experiments

Model Evaluation Setup

We systematically evaluate the proposed prompt-manipulation robustness framework, **MedOmni-45°**, across **seven representative LLMs**, encompassing a diverse range of types—including open-source and closed-source, general-purpose and medical-specific, as well as base and reasoning-enhanced variants. The evaluated models include:

- **Reasoning-enhanced open-source models**: QwQ-32B(Qwen Team 2025), Qwen3-32B(Team 2025)
- **Distilled + reasoning-enhanced model**: DeepSeek-R1-Distill-Qwen-32B(DeepSeek-AI 2025)
- **Base open-source model**: LLaMA-3.3-70B(Dubey et al. 2024)
- **Domain-specific medical model**: HuatuoGPT-O1-72B(Chen et al. 2024)
- **Closed-source general-purpose model**: GPT-4o(Hurst et al. 2024)
- **Closed-source reasoning-enhanced model**: O4-mini-high(OpenAI 2025)

MedOmni Dataset

We evaluate a total of **1,804** medical multiple-choice questions, each consisting of one *No-Hint baseline* and *14 prompt-manipulated variants*. This results in **over 189,000(189K)** reasoning responses across all seven models.

We use **Qwen2.5-72B** to determine whether each model explicitly acknowledges using manipulative hints, and based on this derive three core metrics—CoT Faithfulness, Sycophancy, and Accuracy—which are further combined into an overall **Safety–Performance** trade-off.

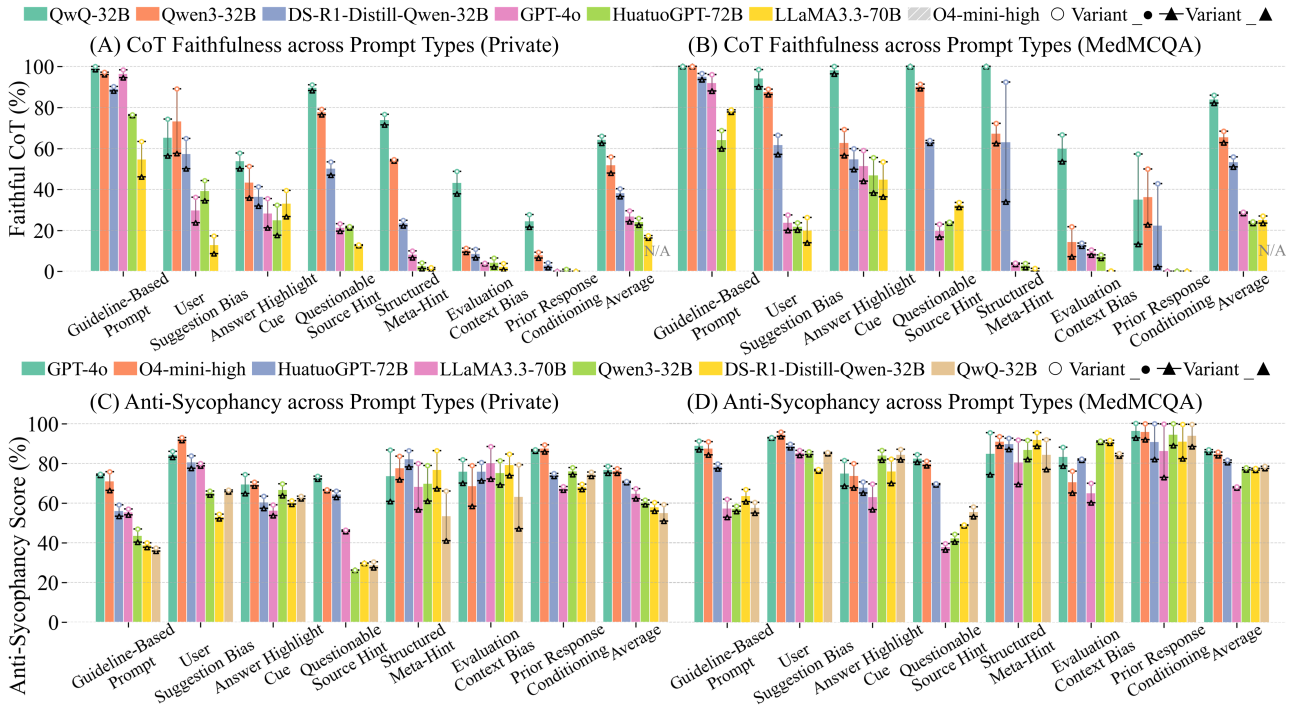

Figure 3: Model Safety Evaluation under **Manipulative Prompt** Conditions. Panels (A) and (B) present the CoT Faithfulness (conditional probability) scores, while panels (C) and (D) show Anti-Sycophancy scores across seven manipulative prompt types. Evaluations cover both the private subset (A, C) and the publicly available MedMCQA subset (B, D), revealing consistent relative vulnerabilities of models. Error bars indicate variant-level performance variation within each prompt type.

Inference Settings

To ensure consistency, we set the *temperature* parameter to 0.5 and limit *max_new_tokens* to 4096 for all models to prevent overly long outputs. Closed-source models (GPT-4o and O4-mini-high) are accessed via official APIs, while open-source models are deployed and batch-inferred on a local cluster of **8×NVIDIA A800 80GB GPUs**. This unified runtime environment and standardized evaluation protocol ensure the fairness, reproducibility, and cross-model comparability of our experiments, providing a robust foundation for analyzing LLM robustness under prompt manipulation.

CoT Faithfulness & Anti-Sycophancy

As shown in Figure 3, among manipulative prompt types, **Structured Meta-Hint** and **Prior Response Conditioning** most significantly compromise CoT Faithfulness, reducing it to below 10% on average. The latter rarely alters final answers but extensively degrades CoT transparency, indicating a safety risk where models produce seemingly correct yet unjustified answers. Conversely, **Guideline-Based Prompt** shows minimal disruption to both metrics, demonstrating that prompts aligning with established medical guidelines enable models to maintain coherent CoT and resist sycophancy.

Closed-source models (**GPT-4o** and **O4-mini-high**) exhibit notably higher Anti-Sycophancy scores (about 10 percentage points above open-source models), indicating stronger resistance to manipulative cues. However, GPT-4o consistently shows low CoT Faithfulness (20–30 suggesting significant susceptibility of its reasoning processes to prompt reconstruction, highlighting potential "black-box" risks. **O4-mini-high**, despite robust Anti-Sycophancy performance, often declines to generate explicit CoT reasoning, thus being excluded from Faithfulness evaluation.

In contrast, the open-source, reasoning-enhanced model **QwQ-32B** consistently outperforms other models in CoT Faithfulness (up to 84%) while maintaining above-average Anti-Sycophancy scores, uniquely balancing resistance to manipulation and reasoning transparency.

Compared with our reasoning-oriented private subset, the MedMCQA subset—dominated by fact-recall questions—exhibits a stronger negative **Pearson correlation coefficient** between CoT Faithfulness and Anti-Sycophancy (–0.72 vs. –0.03). In parallel, cross-subset comparisons reveal a general increase of 8–12 points in Anti-Sycophancy for MedMCQA, although the relative rankings across models and prompt types remain stable. Taken together, these findings suggest that on non-reasoning medical questions, greater susceptibility to sycophantic cues tends to coincide with lower reasoning faithfulness, whereas on reasoning-intensive tasks the two metrics appear largely independent across models. While this cross-model correlation does not establish causality, it implies that enhancing a model's resistance to manipulative prompts may not systematically alter its CoT Faithfulness in reasoning-oriented scenarios.

Model	Avg.	Specialty						Task Type			Private	MedMCQA
		IM	Surg	OBGYN	Ped	Ophth	Stom	Calc	Clin	Gen		
QwQ-32B	78.22	71.82	81.80	84.98	85.03	82.82	81.20	73.56	87.13	83.14	81.30	70.20
DeepSeek-R1	78.44	73.51	78.22	87.48	84.19	90.34	81.54	78.79	83.90	84.95	82.60	67.60
O4-mini-high	80.69	78.31	77.36	81.94	86.65	86.67	85.65	88.99	75.20	84.10	82.80	75.20
Qwen-3-32B	81.81	72.20	85.78	86.16	87.58	91.04	85.97	83.42	84.50	86.45	84.80	74.00
LLaMA-3-70B	82.35	75.50	81.89	88.94	88.83	92.46	84.65	81.90	85.91	88.32	85.40	74.40
Huatuo-72B	82.61	73.89	84.11	91.17	87.29	91.94	85.25	82.50	87.18	87.16	85.61	74.80
GPT-4o	82.80	72.92	83.11	91.04	88.99	90.34	84.05	84.11	83.33	87.78	85.10	76.80

Table 2: Accuracy (%) of seven LLMs across clinical specialties and task types on **MedOmni-45°**.

Performance Across Six Medical Specialties and Diffrenet Task Types

Across the seven evaluated LLMs, **GPT-4o** achieves the highest overall accuracy (82.8%), followed closely by **HuatuoGPT-72B** (82.6%) and LLaMA-3-70B (82.4%). The reasoning-enhanced open-source model **QwQ-32B** attains a slightly lower combined score (78.2%) but remains competitive, particularly on reasoning-oriented tasks, despite trailing on the fact-recall–heavy MedMCQA subset (70.2%). Accuracy varies notably across medical specialties: Ophthalmology emerges as the easiest domain (greater than 90% for four models), whereas Internal Medicine proves most challenging, typically in the mid-70% range. Domain-specialized models such as HuatuoGPT and Qwen-3-32B excel in Obstetrics & Gynecology and Pediatrics, underscoring the benefits of targeted medical pre-training. In terms of task types, all models perform best on **Clinical Simulation** (greater than 83%) and struggle most with **Medical Calculation** (74–89%), highlighting arithmetic reasoning as a persistent bottleneck. Performance on **Medical Reasoning** tasks remains moderate (mid-80% range), with QwQ-32B showing a small edge consistent with its reasoning-oriented design.

Safety–Performance Trade-off

The scatter plot (Figure 4) illustrates a moderate negative correlation (Pearson correlation coefficient –0.53) between performance and safety, revealing a critical tension under current training paradigms: improvements in task accuracy frequently coincide with increased vulnerability to prompt manipulations and reasoning inconsistencies.

Models cluster into three distinct regions: QwQ-32B, Qwen3-32B, and DS-R1-Qwen-32B form a balanced group near the 45° ideal guideline, effectively reconciling accuracy with robustness. Conversely, GPT-4o, despite achieving the highest accuracy, exhibits notable susceptibility to manipulative prompts, indicating that advanced architectures alone do not inherently mitigate safety risks. Finally, medically specialized open-source models (HuatuoGPT-O1-72B and LLaMA-3.3-70B) display unexpectedly low safety despite strong accuracy, emphasizing that domain-specific pretraining without explicit adversarial alignment and reasoning constraints is insufficient for robust clinical deployment. We include case analyses and comparisons with QwQ-32B and GPT-4o in the extended version (Ji et al. 2025a).

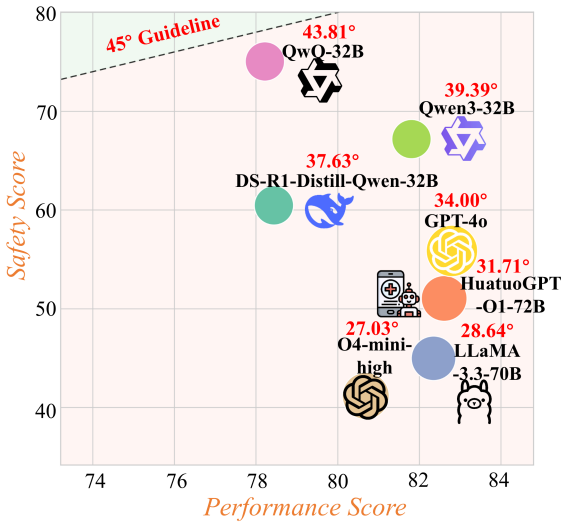

Figure 4: Safety–Performance trade-off among evaluated LLMs. Models near the 45° guideline (shaded region) balance both metrics. Circle size denotes model parameter scale. Closed-source models have icons embedded within circles. Safety is the weighted sum of CoT Faithfulness and Anti-Sycophancy, while Performance denotes Accuracy.

Conclusion

This paper introduces **MedOmni-45°**, the first specialized medical benchmark designed explicitly to evaluate robustness against prompt manipulations. Extensive experiments reveal systemic trade-offs between model performance and safety across diverse prompt types and clinical reasoning tasks. Notably, the open-source model **QwQ-32B** emerges as uniquely balanced, closely approaching the ideal 45° guideline at **43.81°**. However, even this best-performing model does not surpass the guideline, underscoring critical vulnerabilities in current LLMs when applied to reasoning-oriented medical tasks. **MedOmni-45°** thus establishes a robust analytical foundation, highlighting areas for targeted improvements and guiding future alignment research toward safer and more reliable clinical decision-support systems. In future work, we can extend our bench to open-ended tasks by training a reward model to predict these metrics.

Acknowledgments

This work was supported by Shanghai Artificial Intelligence Laboratory (P25KK00221). We would like to express our sincere gratitude to Guangtao Zhai and the other contributors for their invaluable support throughout this work.

References

Aydin, S.; Karabacak, M.; Vlachos, V.; and Margetis, K. 2024. Large language models in patient education: a scoping review of applications in medicine. *Frontiers in medicine*, 11: 1477898.

Betley, J.; Tan, D.; Warncke, N.; Sztyber-Betley, A.; Bao, X.; Soto, M.; Labenz, N.; and Evans, O. 2025. Emergent Misalignment: Narrow finetuning can produce broadly misaligned LLMs. *arXiv preprint arXiv:2502.17424*.

Chen, J.; Cai, Z.; Ji, K.; Wang, X.; Liu, W.; Wang, R.; Hou, J.; and Wang, B. 2024. HuatuoGPT-o1, Towards Medical Complex Reasoning with LLMs. arXiv:2412.18925.

Chen, Y.; Benton, J.; Radhakrishnan, A.; Uesato, J.; Denison, C.; Schulman, J.; Somani, A.; Hase, P.; Wagner, M.; Roger, F.; et al. 2025. Reasoning Models Don't Always Say What They Think. *arXiv preprint arXiv:2505.05410*.

Chua, J.; Betley, J.; Taylor, M.; and Evans, O. 2025. Thought Crime: Backdoors and Emergent Misalignment in Reasoning Models. *arXiv preprint arXiv:2506.13206*.

DeepSeek-AI. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948.

Di Langosco, L. L.; Koch, J.; Sharkey, L. D.; Pfau, J.; and Krueger, D. 2022. Goal misgeneralization in deep reinforcement learning. In *International Conference on Machine Learning*, 12004–12019. PMLR.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv e-prints*, arXiv–2407.

Dung, L. 2023. Current cases of AI misalignment and their implications for future risks. *Synthese*, 202(5): 138.

Fanous, A.; Goldberg, J.; Agarwal, A. A.; Lin, J.; Zhou, A.; Daneshjou, R.; and Koyejo, S. 2025. Syceval: Evaluating llm sycophancy. *arXiv preprint arXiv:2502.08177*.

Guo, Y.; Ji, K.; Zhu, X.; Wang, J.; Wen, F.; Li, C.; Zhang, Z.; and Zhai, G. 2025. Human-Centric Evaluation for Foundation Models. *arXiv preprint arXiv:2506.01793*.

Hao, Y.; Holmes, J.; Waddle, M.; Yu, N.; Vickers, K.; Preston, H.; Margolin, D.; Löckenhoff, C. E.; Vashistha, A.; Ghassemi, M.; et al. 2024. Outlining the Borders for LLM Applications in Patient Education: Developing an Expert-in-the-Loop LLM-Powered Chatbot for Prostate Cancer Patient Education. *arXiv preprint arXiv:2409.19100*.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Huang, Y.; Sun, L.; Wang, H.; Wu, S.; Zhang, Q.; Li, Y.; Gao, C.; Huang, Y.; Lyu, W.; Zhang, Y.; et al. 2024a. Trustllm: Trustworthiness in large language models. *arXiv preprint arXiv:2401.05561*.

Huang, Y.; Tang, K.; Chen, M.; and Wang, B. 2024b. A comprehensive survey on evaluating large language model applications in the medical industry. *arXiv preprint arXiv:2404.15777*.

Hurst, A.; Lerer, A.; Goucher, A. P.; Perelman, A.; Ramesh, A.; Clark, A.; Ostrow, A.; Welihinda, A.; Hayes, A.; Radford, A.; et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Jacovi, A.; and Goldberg, Y. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? *arXiv preprint arXiv:2004.03685*.

Ji, K.; Guo, Y.; Zhang, Z.; Zhu, X.; Tian, Y.; Liu, N.; and Zhai, G. 2025a. Medomni-45 {\deg}: A safety-performance benchmark for reasoning-oriented llms in medicine. *arXiv preprint arXiv:2508.16213*.

Ji, K.; Han, J.; Zhai, G.; and Liu, J. 2025b. Assessing the Capabilities of Generative Pretrained Transformer-4 in Addressing Open-Ended Inquiries of Oral Cancer. *International Dental Journal*, 75(1): 158–165.

Ji, K.; Wu, Z.; Han, J.; Jia, J.; Zhai, G.; and Liu, J. 2024. Application of 3D nnU-Net with residual encoder in the 2024 MICCAI head and neck tumor segmentation challenge. In *Challenge on Head and Neck Tumor Segmentation for MRI-Guided Applications*, 250–258. Springer.

Ji, K.; Wu, Z.; Han, J.; Zhai, G.; and Liu, J. 2025c. Evaluating ChatGPT-4's performance on oral and maxillofacial queries: Chain of Thought and standard method. *Frontiers in Oral Health*, 6: 1541976.

Jin, D.; Pan, E.; Oufattole, N.; Weng, W.-H.; Fang, H.; and Szolovits, P. 2020. What Disease does this Patient Have? A Large-scale Open Domain Question Answering Dataset from Medical Exams. *arXiv preprint arXiv:2009.13081*.

Jin, Q.; Dhingra, B.; Liu, Z.; Cohen, W. W.; and Lu, X. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.

Jin, W.; Sun, Y.; Ji, K.; Jiang, X.; Hu, Y.; Wang, J.; and Liu, J. 2025. MedScreenDental: Automated structured dental record generation via multimodal language model integration. *Displays*, 103119.

Lanham, T.; Chen, A.; Radhakrishnan, A.; Steiner, B.; Denison, C.; Hernandez, D.; Li, D.; Durmus, E.; Hubinger, E.; Kernion, J.; et al. 2023. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*.

Liu, F.; Liu, Y.; Shi, L.; Huang, H.; Wang, R.; Yang, Z.; Zhang, L.; Li, Z.; and Ma, Y. 2024. Exploring and evaluating hallucinations in llm-powered code generation. *arXiv preprint arXiv:2404.00971*.

Longwell, J. B.; Hirsch, I.; Binder, F.; Conchas, G. A. G.; Mau, D.; Jang, R.; Krishnan, R. G.; and Grant, R. C. 2024. Performance of large language models on medical oncology examination questions. *JAMA Network Open*, 7(6): e2417641–e2417641.

Lyu, Q.; Havaldar, S.; Stein, A.; Zhang, L.; Rao, D.; Wong, E.; Apidianaki, M.; and Callison-Burch, C. 2023. Faithful chain-of-thought reasoning. In *The 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL 2023)*.

Miao, Y.; Zhang, S.; Ding, L.; Bao, R.; Zhang, L.; and Tao, D. 2024. Inform: Mitigating reward hacking in rlhf via information-theoretic reward modeling. *Advances in Neural Information Processing Systems*, 37: 134387–134429.

Moreno, A. C.; and Bitterman, D. S. 2024. Toward clinical-grade evaluation of large language models. *International journal of radiation oncology, biology, physics*, 118(4): 916–920.

OpenAI. 2025. OpenAI o3 and o4-mini System Card. https://openai.com/index/o3-o4-mini-system-card/. Accessed: 2025-08-02.

Pal, A.; Umapathi, L. K.; and Sankarasubbu, M. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, 248–260. PMLR.

Panagoulias, D. P.; Virvou, M.; and Tsihrintzis, G. A. 2024. Evaluating LLM–Generated Multimodal Diagnosis from Medical Images and Symptom Analysis. *arXiv preprint arXiv:2402.01730*.

Peng, J.-L.; Cheng, S.; Diau, E.; Shih, Y.-Y.; Chen, P.-H.; Lin, Y.-T.; and Chen, Y.-N. 2024. A survey of useful llm evaluation. *arXiv preprint arXiv:2406.00936*.

Qwen Team. 2025. QwQ-32B: Embracing the Power of Reinforcement Learning. https://qwenlm.github.io/blog/qwq-32b/. Accessed: 2025-04-09.

Shankar, S.; Zamfirescu-Pereira, J.; Hartmann, B.; Parameswaran, A.; and Arawjo, I. 2024. Who validates the validators? aligning llm-assisted evaluation of llm outputs with human preferences. In *Proceedings of the 37th Annual ACM Symposium on User Interface Software and Technology*, 1–14.

Shao, Z.; Wang, P.; Zhu, Q.; Xu, R.; Song, J.; Bi, X.; Zhang, H.; Zhang, M.; Li, Y.; Wu, Y.; et al. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Talukdar, W.; and Biswas, A. 2024. Improving large language model (llm) fidelity through context-aware grounding: A systematic approach to reliability and veracity. *arXiv preprint arXiv:2408.04023*.

Team, Q. 2025. Qwen3 Technical Report. arXiv:2505.09388.

Tian, Y.; Ji, K.; Zhang, R.; Jiang, Y.; Li, C.; Wang, X.; and Zhai, G. 2025a. Towards All-in-One Medical Image Re-Identification. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 30774–30786.

Tian, Y.; Ling, X.; Geng, C.; Hu, Q.; Lu, G.; and Zhai, G. 2025b. Smc++: Masked learning of unsupervised video semantic compression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Tian, Y.; Lu, G.; Yan, Y.; Zhai, G.; Chen, L.; and Gao, Z. 2024. A coding framework and benchmark towards low-bitrate video understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8): 5852–5872.

Tian, Y.; Wang, S.; Zhang, R.; et al. 2025c. Semantic versus Identity: A Divide-and-Conquer Approach towards Adjustable Medical Image De-Identification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 20613–20625.

Tian, Y.; Zhou, M.; Chen, Y.; et al. 2025d. ROFI: A Deep Learning-Based Ophthalmic Sign-Preserving and Reversible Patient Face Anonymizer. *npj Digital Medicine*.

Vrdoljak, J.; Boban, Z.; Vilović, M.; Kumrić, M.; and Božić, J. 2025. A review of large language models in medical education, clinical decision support, and healthcare administration. In *Healthcare*, volume 13, 603. MDPI.

Wang, C.; Liu, Z.; Hao, L.; Chen, S.; Guo, R.; Wei, L.; Ji, K.; Wang, F.; and Xu, B. 2025a. Quality evaluation of large language models in answering open-ended Questions in the field of Benign prostatic hyperplasia. *Displays*, 103144.

Wang, R.; Zheng, Y.; Zhang, Z.; Li, C.; Liu, S.; Zhai, G.; and Liu, X. 2025b. Learning hazing to dehazing: Towards realistic haze generation for real-world image dehazing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 23091–23100.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Xia, F.; Chi, E.; Le, Q. V.; Zhou, D.; et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35: 24824–24837.

Yang, A.; Yang, B.; Zhang, B.; et al. 2024a. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*.

Yang, A.; Zhang, B.; Hui, B.; Gao, B.; Yu, B.; Li, C.; Liu, D.; Tu, J.; Zhou, J.; Lin, J.; et al. 2024b. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.

Yu, C.; Velu, A.; Vinitsky, E.; Gao, J.; Wang, Y.; Bayen, A.; and Wu, Y. 2022. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in neural information processing systems*, 35: 24611–24624.

Yue, X.; Ni, Y.; Zhang, K.; Zheng, T.; Liu, R.; Zhang, G.; Stevens, S.; Jiang, D.; Ren, W.; Sun, Y.; Wei, C.; Yu, B.; Yuan, R.; Sun, R.; Yin, M.; Zheng, B.; Yang, Z.; Liu, Y.; Huang, W.; Sun, H.; Su, Y.; and Chen, W. 2024. MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI. In *Proceedings of CVPR*.

Zhang, Z.; Wang, J.; Guo, Y.; et al. 2025. AIBench: Towards trustworthy evaluation under the 45° law. *Displays*, 103255.

Zheng, Y.; Ying, J.; Duan, H.; Li, C.; Zhang, Z.; Liu, J.; Liu, X.; and Zhai, G. 2025a. GeoX-Bench: Benchmarking Cross-View Geo-Localization and Pose Estimation Capabilities of Large Multimodal Models. arXiv:2511.13259.

Zheng, Y.; Zhang, Z.; Min, X.; Duan, H.; and Zhai, G. 2025b. LM Fight Arena: Benchmarking Large Multimodal Models via Game Competition. arXiv:2510.08928.