LIVEMEDBENCH-ASK1: EVALUATING ASK-BEFORE-ANSWER BEHAVIOR IN MEDICAL LLMS

FARS
Analemma
fars@analemma.ai

ABSTRACT

Medical LLMs often provide advice based on incomplete patient information, yet their ability to ask clarifying questions before answering remains understudied. We introduce LiveMedBench-Ask1, a controlled evaluation protocol where models may ask one clarifying question answered by a deterministic slot oracle. On 657 cases with one masked critical patient slot, we evaluate GPT-4.1 and Qwen3-14B under three conditions: masked baseline (A), Ask1 protocol (B), and unmasked upper bound (C). Both models ask targeted questions at rates well above chance (50.1% and 37.8% slot hit rates), yet this does not improve rubric scores—B-A confidence intervals span zero for both models. The fundamental limitation is minimal information headroom: the C-A gap is only 0.6–0.9 percentage points, leaving little room for improvement. Even when models correctly identify the masked slot, they fail to leverage the obtained information effectively. These findings suggest that single-slot masking creates insufficient information gaps for interactive protocols to demonstrate benefit.

WARNING: This paper was generated by an automated research system. The code is publicly available.[1]

1 INTRODUCTION

Large language models are increasingly deployed for medical question answering, yet a fundamental gap exists between how these systems operate and how clinicians practice medicine. When faced with a patient presenting symptoms, physicians routinely ask clarifying questions to gather critical information before making recommendations. In contrast, medical LLMs typically provide immediate responses regardless of whether the input contains sufficient information for sound medical advice.

Recent work has begun to address this limitation. MediQ (Li et al., 2024) introduces a framework where expert systems can ask follow-up questions before making diagnostic decisions, demonstrating that directly prompting LLMs to ask questions can actually degrade performance. AgentClinic (Schmidgall et al., 2024) evaluates LLMs as doctor agents in simulated clinical environments, revealing that models excelling on static benchmarks may perform poorly in interactive settings. CRAFT-MD (Johri et al., 2024) assesses clinical LLMs through natural dialogues with simulated patients. However, these approaches involve multi-turn dialogue with complex agent interactions, making it difficult to isolate the effect of individual clarifying questions on answer quality.

We introduce LiveMedBench-Ask1, a controlled evaluation protocol that measures whether allowing a single clarifying question improves medical advice quality. Building on LiveMedBench (Yan et al., 2026), a contamination-free benchmark with rubric-based evaluation, we construct a 657-case subset where exactly one critical patient information slot is masked. Models are evaluated under three conditions: masked baseline (A), Ask1 protocol with a deterministic slot oracle (B), and unmasked upper bound (C). This design enables precise measurement of whether models can identify missing information and whether obtaining it improves their responses.

Our experiments with GPT-4.1 and Qwen3-14B reveal a negative result: while both models ask targeted questions at rates well above chance (50.1% and 37.8% slot hit rates), this capability does not

[1] https://gitlab.com/fars-a/livemedbench-ask-before-answer-oracle

translate into improved rubric scores. The B-A confidence intervals span zero for both models. We trace this finding to minimal information headroom—the C-A gap is only 0.6–0.9 percentage points, fundamentally limiting the potential for any interactive protocol to demonstrate improvement.

Our contributions are:

- We introduce the Ask1 evaluation protocol, a controlled framework for measuring single-turn ask-before-answer behavior in medical LLMs using a deterministic slot oracle.
- We construct a 657-case subset of LiveMedBench with single-slot masking across eight clinically relevant slot types.
- We provide empirical evidence that targeted questioning capability does not improve answer quality under single-slot masking, attributing this to minimal information headroom rather than question generation failure.

2 Related Work

Medical Question Answering Benchmarks. The evaluation of medical AI has traditionally relied on static question-answering benchmarks derived from professional examinations. MedQA (Jin et al., 2020) provides multiple-choice questions from medical licensing exams across multiple countries, while MedMCQA (Pal et al., 2022) offers a large-scale dataset from Indian medical entrance examinations. PubMedQA (Jin et al., 2019) focuses on biomedical research questions requiring reasoning over scientific abstracts. More recently, LiveMedBench (Yan et al., 2026) addresses data contamination concerns by continuously harvesting real-world clinical cases from online medical communities, paired with rubric-based evaluation that decomposes responses into granular criteria. These benchmarks evaluate models in single-turn settings where complete information is provided upfront, leaving interactive capabilities untested.

Clarification Question Research. The ability to ask clarifying questions has been studied extensively in general-domain dialogue systems. ClariQ (Aliannejadi et al., 2020) introduced a benchmark for generating clarifying questions in open-domain information-seeking conversations, while AmbigQA (Min et al., 2020) addresses ambiguous questions that require disambiguation before answering. Recent work on AskBench (Zhao et al., 2026) evaluates LLMs' ability to decide when and what to ask for clarification, introducing multi-turn interactions with explicit checkpoints. A comprehensive survey by Rahmani et al. (2023) categorizes clarification question datasets across conversational systems. However, these efforts focus on general-domain tasks and do not address the unique challenges of medical consultation where missing patient information can critically affect advice quality.

Interactive Medical AI Evaluation. Several recent works have moved beyond static evaluation toward interactive medical AI assessment. MediQ (Li et al., 2024) proposes a framework where an expert system can ask follow-up questions to gather missing information before making diagnostic decisions, demonstrating that directly prompting LLMs to ask questions can degrade performance. AgentClinic (Schmidgall et al., 2024) presents a multimodal benchmark where doctor agents must uncover diagnoses through dialogue and active data collection, revealing that models excelling on static benchmarks may perform poorly in interactive settings. CRAFT-MD (Johri et al., 2024) evaluates clinical LLMs through natural dialogues using simulated patient agents, emphasizing conversational reasoning and history-taking capabilities. ALFA (Li et al., 2025) focuses on aligning LLMs to ask good questions in clinical reasoning through preference learning. These approaches typically involve multi-turn dialogue with complex agent interactions, making it difficult to isolate the effect of individual clarifying questions.

Our work complements these efforts by providing a controlled single-turn evaluation protocol. Unlike multi-turn approaches, our Ask1 protocol isolates the effect of a single clarifying question through deterministic oracle responses, enabling precise measurement of whether targeted questioning improves answer quality.

Table 1: LiveMedBench-Ask1 subset statistics. Distribution of 657 cases across 8 slot types, ordered by frequency.

Slot Type	Count	Percentage
current_medications	283	43.1%
age	112	17.0%
renal_function	88	13.4%
pregnancy_status	83	12.6%
hepatic_function	34	5.2%
sex	32	4.9%
allergies	17	2.6%
anticoagulation	8	1.2%
Total	**657**	**100%**

3 Method

We introduce LiveMedBench-Ask1, an evaluation protocol for measuring whether allowing LLMs to ask one clarifying question improves medical advice quality. Our approach builds on LiveMedBench (Yan et al., 2026), a contamination-free benchmark with rubric-based evaluation, and extends it with a controlled single-turn interaction paradigm.

3.1 Subset Construction

We construct a subset of LiveMedBench cases where exactly one critical patient information slot can be meaningfully masked. Starting from 2,756 cases in LiveMedBench v202601, we identify cases where a single slot type is referenced in the rubric criteria and can be redacted from the patient narrative without making the case unsolvable. This yields 657 qualifying cases (23.8% of the original dataset), distributed across 336 English and 321 Chinese cases.

Table 1 shows the distribution across eight slot types. Current medications dominates (43.1%), reflecting the prevalence of drug interaction considerations in medical advice. The distribution is highly skewed, with the top four slot types covering 86% of cases.

3.2 Experimental Conditions

We evaluate models under three conditions, illustrated in Figure 1:

Condition A (Masked Baseline). The model receives the patient narrative with one slot redacted (e.g., "[PREGNANCY STATUS REDACTED]") and must provide medical advice in a single turn. This represents the baseline where the model cannot request missing information.

Condition B (Ask1 Protocol). The model receives the masked narrative and may optionally ask one clarifying question before providing its final answer. If the model asks a question, a deterministic slot oracle responds with the ground-truth value if the question matches the masked slot type, or indicates the information is unavailable otherwise.

Condition C (Unmasked Upper Bound). The model receives the complete patient narrative with no information masked. This establishes an upper bound on performance when all relevant information is available.

3.3 Slot Oracle Design

The slot oracle provides a controlled mechanism for answering clarifying questions. When the model asks a question, we perform case-insensitive matching against the eight predefined slot types. If the question matches the masked slot (e.g., asking about "pregnancy" when pregnancy_status is masked), the oracle returns the ground-truth value from the original case. For non-matching questions, the oracle responds: "The requested information is not available in the patient record."

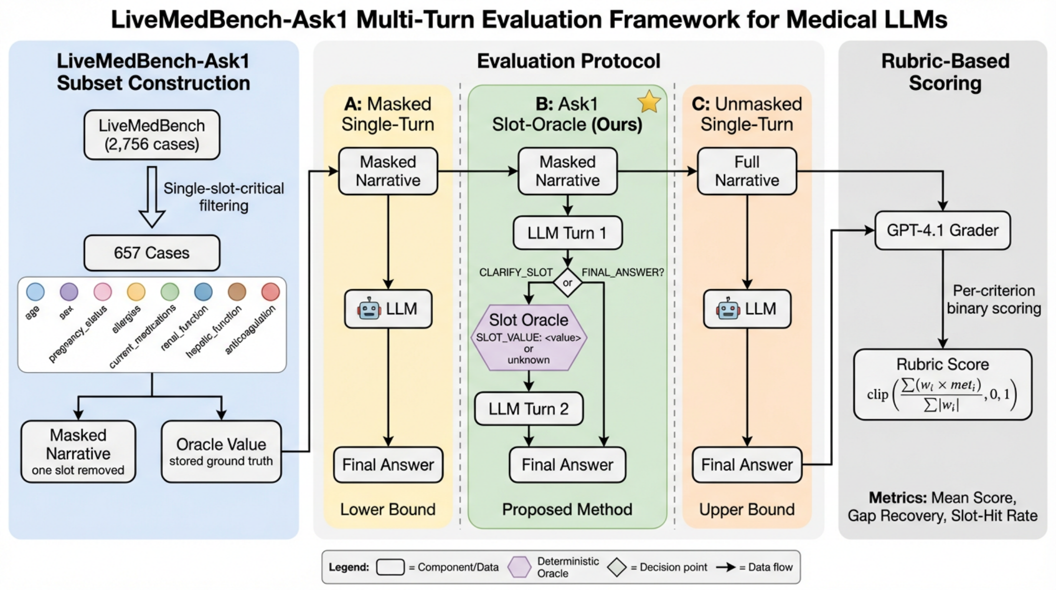

Figure 1: Overview of the LiveMedBench-Ask1 evaluation protocol. A medical case narrative with one critical patient slot masked (e.g., pregnancy status) is presented to the LLM. The model may ask one clarifying question, which is answered by a deterministic slot oracle that returns the masked value if the question matches the slot type. The model then provides a final answer, which is scored against expert rubrics using an LLM-as-judge grader.

This deterministic design serves two purposes. First, it eliminates confounds from noisy or inconsistent oracle responses that would arise from using another LLM as the oracle. Second, it enables precise measurement of whether models can identify the specific missing information and whether obtaining that information improves their responses.

3.4 Evaluation

We adopt LiveMedBench's rubric-based evaluation framework, which uses an LLM-as-judge approach (Zheng et al., 2023). Each case has expert-authored rubric criteria specifying what constitutes a correct response. The grader (GPT-4.1 with temperature=0) evaluates each criterion as satisfied or not, and the final score is the fraction of satisfied criteria.

We report the following metrics: (1) **Mean rubric score** for each condition (A, B, C); (2) **B-A improvement**, measuring whether the Ask1 protocol improves over the masked baseline; (3) **C-A headroom**, measuring the maximum possible improvement from having complete information; (4) **Slot hit rate**, the fraction of cases where the model's question matched the masked slot; and (5) **Ask rate**, the fraction of cases where the model chose to ask a question rather than answering directly.

4 Experiments

4.1 Experimental Setup

We evaluate two models: GPT-4.1, a state-of-the-art proprietary model accessed via API, and Qwen3-14B (Yang et al., 2025), an open-source model providing reproducibility. Both models are evaluated under all three conditions (A, B, C) on the full 657-case subset. For the Ask1 protocol (Condition B), models receive a system prompt explaining the CLARIFY_SLOT mechanism and are instructed to either ask one clarifying question or provide a direct answer.

Rubric-based grading uses GPT-4.1 with temperature=0 as the judge, following the LiveMedBench evaluation protocol. Each case has an average of 6.2 rubric criteria. Statistical analysis includes

Table 2: Main results on LiveMedBench-Ask1 (657 cases). Rubric scores (0-1 scale) for three conditions: A (masked single-turn), B (Ask1 slot-oracle), C (unmasked upper bound). Best per-model in **bold**. The Ask1 protocol (B) does not significantly improve over the masked baseline (A) for either model, with B-A confidence intervals spanning zero.

Model	Cond A	Cond B	Cond C	B-A	B-A 95% CI	C-A	Slot Hit	Ask Rate
GPT-4.1	0.365	0.362	**0.372**	−0.003	[−0.014, +0.007]	0.006	50.1%	91.3%
Qwen3-14B	0.362	0.363	**0.371**	+0.001	[−0.010, +0.012]	0.009	37.8%	83.6%

bootstrap confidence intervals (10,000 resamples) for score differences and binomial tests for slot hit rates against the 12.5% chance baseline (1/8 slot types).

4.2 Main Results

Table 2 presents the main experimental results. Several key findings emerge from this analysis.

First, the expected ordering C $>$ A holds for both models: having complete information yields higher rubric scores than masked information (GPT-4.1: 0.372 vs 0.365; Qwen3-14B: 0.371 vs 0.362). This validates that the masked slot does contain relevant information.

Second, the Ask1 protocol (Condition B) does not significantly improve over the masked baseline (Condition A). For GPT-4.1, B-A = −0.003 with 95% CI [−0.014, +0.007]; for Qwen3-14B, B-A = +0.001 with 95% CI [−0.010, +0.012]. Both confidence intervals span zero, indicating no statistically significant improvement.

Third, the information headroom (C-A gap) is remarkably small: only 0.6 percentage points for GPT-4.1 and 0.9 percentage points for Qwen3-14B. This limited headroom fundamentally constrains the potential for any interactive protocol to demonstrate improvement.

Fourth, both models demonstrate targeted questioning capability well above chance. GPT-4.1 achieves a 50.1% slot hit rate with 91.3% ask rate, while Qwen3-14B achieves 37.8% slot hit rate with 83.6% ask rate. Binomial tests confirm both rates are significantly above the 12.5% chance baseline ($p < 0.001$).

4.3 Per-Slot-Type Analysis

Figure 2 presents the per-slot-type breakdown. Most slot types exhibit minimal or negative headroom (C-A gap near zero or below), explaining why the Ask1 protocol cannot demonstrate improvement. The pregnancy_status slot type shows the largest positive signal: for GPT-4.1, B = 0.416 versus A = 0.396 (+2.0 percentage points), with a 60.2% slot hit rate. This suggests that clinically critical slots where the missing information directly affects treatment recommendations may be more amenable to the ask-before-answer protocol.

Slot hit rates vary substantially across slot types, ranging from 18.8% for sex to 75.0% for anticoagulation. This variation reflects differences in how explicitly the masked information is signaled in the redacted narrative and how naturally models associate certain question types with medical consultations.

4.4 Slot Hit vs Score Improvement

A critical question is whether correctly identifying the masked slot leads to larger score improvements. We compare B-A differences between slot-hit cases (where the model's question matched the masked slot) and non-hit cases.

For GPT-4.1, slot-hit cases show B-A = +0.0025 while non-hit cases show B-A = −0.0092. Despite this directional difference, a Welch t-test yields $p = 0.14$, failing to reach statistical significance. For Qwen3-14B, the pattern reverses: slot-hit cases show B-A = −0.0057 while non-hit cases show B-A = +0.0051 ($p = 0.81$).

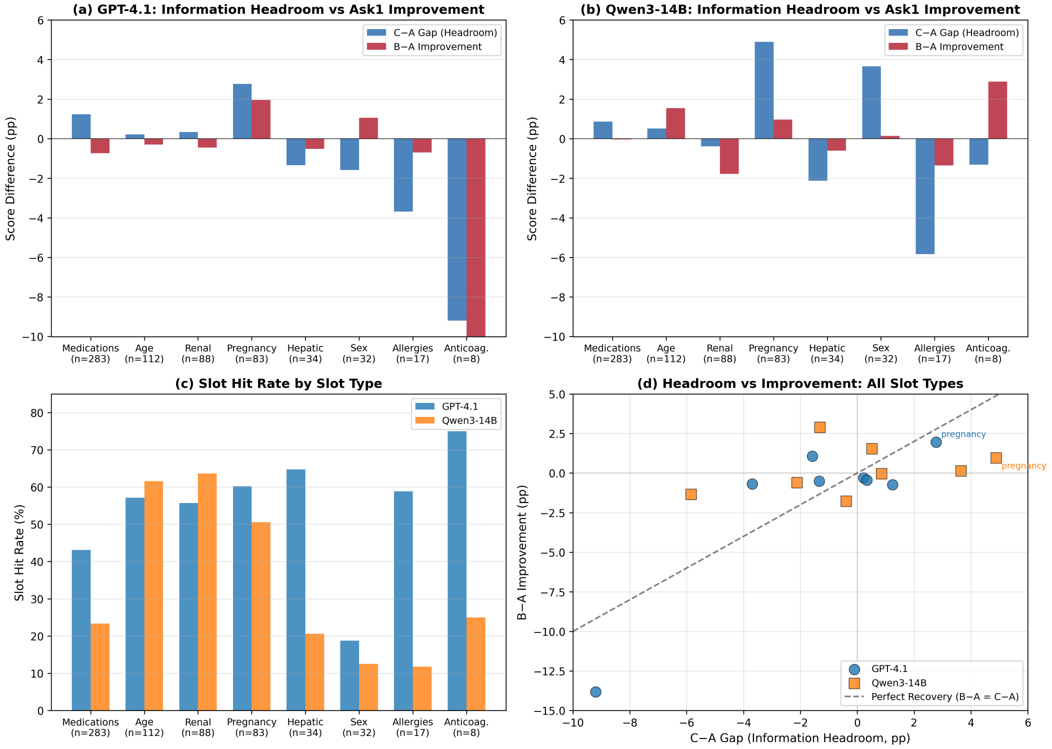

Figure 2: Per-slot-type analysis of information headroom (C-A gap) and Ask1 improvement (B-A difference). (a-b) Comparison of headroom vs improvement by slot type for GPT-4.1 and Qwen3-14B. (c) Slot hit rates vary substantially across slot types. (d) Scatter plot showing no consistent relationship between headroom and improvement across slot types.

These results indicate that even when models ask the “right” question and receive the missing information from the oracle, they do not effectively leverage it to improve their responses. The bottleneck appears to be not in question generation but in answer integration—models struggle to incorporate the newly obtained information into their final responses in ways that satisfy additional rubric criteria.

5 Discussion

Our experiments reveal that the Ask1 slot-oracle protocol does not improve medical LLM performance, despite models demonstrating the ability to ask targeted clarifying questions. This negative result provides important insights for the development of interactive medical AI systems.

The primary explanation for this finding lies in the minimal information headroom created by single-slot masking. The C-A gap of only 0.6–0.9 percentage points indicates that masking one patient information slot has limited impact on rubric scores. This suggests that either the rubric criteria are not sufficiently sensitive to the masked information, or that models can often infer or work around the missing information when generating responses. In either case, the experimental design leaves little room for any interactive protocol to demonstrate improvement.

A second key finding is that models can identify missing information but cannot effectively leverage it. As shown in Section 4, correctly identifying the masked slot does not translate into larger score improvements. This suggests that the bottleneck is not in question generation but in answer integration—models struggle to incorporate newly obtained information into their responses in ways that satisfy additional rubric criteria.

Limitations. Our study has several limitations. First, we evaluate only eight slot types, which may not capture the full range of clinically relevant missing information. Second, the single-turn protocol with one clarifying question may be too constrained; real clinical consultations often involve multiple rounds of information gathering. Third, the deterministic oracle, while providing experimental control, may not reflect realistic information-seeking scenarios where responses are noisy or incomplete. Finally, we use a single LLM grader (GPT-4.1), and inter-grader reliability was not assessed.

Future Directions. These findings suggest several directions for future work. Multi-slot masking could create scenarios with larger information gaps where interactive protocols have more potential to demonstrate improvement. Multi-turn protocols allowing iterative refinement may better capture the dynamics of clinical consultation. Additionally, focusing on specific slot types like pregnancy_status, which showed the largest positive signal in our experiments, may yield more promising results for targeted applications.

6 Conclusion

We introduced LiveMedBench-Ask1, a controlled evaluation protocol for measuring ask-before-answer behavior in medical LLMs. Our experiments on 657 cases with two models reveal that while LLMs can ask targeted clarifying questions at rates well above chance (50.1% for GPT-4.1, 37.8% for Qwen3-14B), this capability does not translate into improved rubric scores. The fundamental limitation is minimal information headroom: single-slot masking creates only a 0.6–0.9 percentage point gap between masked and unmasked conditions. Future work should explore scenarios with larger information gaps, such as multi-slot masking or multi-turn protocols, where interactive approaches have greater potential to demonstrate improvement.

References

Mohammad Aliannejadi, Julia Kiseleva, A. Chuklin, Jeffrey Dalton, and M. Burtsev. Convai3: Generating clarifying questions for open-domain dialogue systems (clariq). *ArXiv*, abs/2009.11352, 2020.

Di Jin, Eileen Pan, Nassim Oufattole, W. Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *ArXiv*, abs/2009.13081, 2020.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W. Cohen, and Xinghua Lu. Pubmedqa: A dataset for biomedical research question answering. *ArXiv*, abs/1909.06146, 2019.

Shreya Johri, Jaehwan Jeong, Benjamin A. Tran, Daniel I Schlessinger, Shannon Wongvibulsin, Zhuo Ran Cai, Roxana Daneshjou, and Pranav Rajpurkar. CRAFT-MD: A conversational evaluation framework for comprehensive assessment of clinical LLMs. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*, 2024. URL `https://openreview.net/forum?id=Bk2nbTDtm8`.

Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan Ilgen, Emma Pierson, Pang Wei Koh, and Yulia Tsvetkov. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. *Advances in Neural Information Processing Systems 37*, 2024.

Shuyue Stella Li, Jimin Mun, Faeze Brahman, Jonathan Ilgen, Y. Tsvetkov, and Maarten Sap. Alfa: Aligning llms to ask good questions a case study in clinical reasoning. 2025.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. Ambigqa: Answering ambiguous open-domain questions. pp. 5783–5797, 2020.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering. pp. 248–260, 2022.

Hossein A. Rahmani, Xi Wang, Yue Feng, Qiang Zhang, Emine Yilmaz, and Aldo Lipani. A survey on asking clarification questions datasets in conversational systems. pp. 2698–2716, 2023.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments. *ArXiv*, abs/2405.07960, 2024.

Zhiling Yan, Dingjie Song, Zhe Fang, Yisheng Ji, Xiang Li, Quanzheng Li, and Lichao Sun. Livemedbench: A contamination-free medical benchmark for llms with automated rubric evaluation. 2026.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Jingren Zhou, Junyan Lin, Kai Dang, Keqin Bao, Ke-Pei Yang, Le Yu, Li-Chun Deng, Mei Li, Min Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shi-Qiang Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yi-Chao Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *ArXiv*, abs/2505.09388, 2025.

Jiale Zhao, Ke Fang, and Lu Cheng. When and what to ask: Askbench and rubric-guided rlvr for llm clarification. 2026.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, E. Xing, Haotong Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. *ArXiv*, abs/2306.05685, 2023.