UIC at ArchEHR-QA 2025: Tri-Step Pipeline for Reliable Grounded Medical Question Answering

Mohammad Arvan[1] **Anuj Gautam**[2] **Mohan Zalake**[1] **Karl M. Kochendorfer**[1]
{marvan3, zalake, kkoche1}@uic.edu, agautam@uillinois.edu

[1]University of Illinois at Chicago
[2]University of Illinois

Abstract

Automated response generation from electronic health records (EHRs) holds potential to reduce clinician workload, but it introduces important challenges related to factual accuracy and reliable grounding in clinical evidence. We present a structured three-step pipeline that uses large language models (LLMs) for evidence classification, guided response generation, and iterative quality control. To enable rigorous evaluation, our framework combines traditional reference-based metrics with a claim-level "LLM-as-a-Judge" methodology. On the ArchEHR-QA benchmark, our system achieves 82.0 percent claim-level evidence faithfulness and 51.6 percent citation-level factuality, demonstrating strong performance in generating clinically grounded responses. These findings highlight the utility of structured LLM pipelines in healthcare applications, while also underscoring the importance of transparent evaluation and continued refinement. All code, prompt templates, and evaluation tools are publicly available.

1 Introduction

Artificial intelligence (AI) holds transformative potential for healthcare, particularly in automating routine clinical tasks. A significant challenge in contemporary clinical practice is managing patient messages efficiently, a process that often requires clinicians to synthesize information from electronic health records (EHRs) and compose personalized, accurate responses. This time-consuming task imposes substantial cognitive and emotional burdens on medical professionals, contributing to burnout and potentially diminishing the quality of patient care (Shanafelt et al., 2022).

The ArchEHR challenge addresses this critical need by focusing on automated clinical response generation from EHRs. This process presents two primary technical challenges: the accurate extraction of relevant information from patient historical records, and the generation of factual, faithful, context-appropriate responses suitable for patient communication. Large language models (LLMs) have shown promising capabilities in medical question answering, with some studies reporting that they match or exceed clinicians in empathy and communication quality (Ayers et al., 2023). However, their real-world deployment remains constrained by risks of factual errors, hallucinations (i.e., the generation of incorrect or fabricated information), and misunderstandings of medical context.

A fundamental challenge in advancing this field lies in the evaluation of AI-generated responses. Traditional text similarity metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) have demonstrated poor correlation with human judgments across various tasks and scenarios (Liu et al., 2016; Lowe et al., 2017; Xu et al., 2023; Fabbri et al., 2021; Ernst et al., 2023). This limitation necessitates novel approaches to ensure the reliability and safety of automated clinical communication systems.

To address these challenges, we present a grounded medical question-answering system specifically designed for the ArchEHR challenge. Our approach innovates by treating evidence classification as a multiple-choice task, where an LLM selects among predefined clinical evidence categories. This classification then informs a structured clinical response generation process, followed by automated quality control and iterative revision to enhance response adherence to the required format and citation standards.

The remainder of this paper is organized as follows: Section 2 reviews related work on medical question answering, the spectrum of LLM usage, and evaluation methodologies for natural language generation. Section 3 details our methodology, including the LLM-based classification system, response generation process, and evaluation

BioNLP 2025 Shared Tasks, pages 110–117

framework. Section 4 presents experimental results on the ArchEHR-QA dataset. We then discuss the implications and limitations of our findings in Section 5. Finally, we conclude with a summary and directions for future work. Our source code, prompts, and evaluation scripts are available at https://github.com/mo-arvan/grounded-medical-question-answering.

2 Related Work

Our work intersects three fundamental areas: medical question answering (QA) using large language models (LLMs), the spectrum of LLM usage strategies, and evaluation methods for natural language generation (NLG). Together, these domains support the development of a reliable medical QA system. In this section, we summarize recent research in each area to contextualize our contributions.

Medical QA with LLMs Recent advances in large language models have significantly transformed medical QA, demonstrating strong performance in few-shot and zero-shot settings (Kung et al., 2023; Nori et al., 2023; Brin et al., 2023; Singhal et al., 2022). Despite their strengths, these models continue to face critical challenges. Chief among these are hallucinations, referring to generated statements that are not supported by underlying medical evidence or knowledge sources (Zhang et al., 2023; Yang et al., 2024), and difficulties in maintaining accurate, up-to-date clinical knowledge (Zhou et al., 2023; Gao et al., 2023). Our work addresses these limitations through a combination of targeted constraints and comprehensive evaluation protocols designed to ensure response faithfulness.

Spectrum of LLM Usage The complexity of medical queries has prompted the adoption of distinct modeling strategies aimed at improving reasoning and accuracy. One widely used approach involves task decomposition, in which a complex problem is reformulated into smaller, sequential reasoning tasks. These are often structured as chains or directed acyclic graphs (DAGs) of intermediate steps (Wei et al., 2022; Shen et al., 2023). Although effective, these structures are typically defined in advance and lack adaptability. Alternatively, AI agents offer a more dynamic approach. These systems autonomously generate and execute plans informed by contextual cues (Kim et al., 2024). However, such flexibility introduces increased system complexity and requires more rigorous evaluation to verify reliability (Anthropic, 2025). Our framework adopts a pipeline strategy that decomposes responses into interpretable stages. This approach balances control and transparency with adaptability across diverse query types. The exploration of more autonomous agent-based approaches is deferred to future work.

Evaluation of Natural Language Generation Evaluation of generated medical text involves multiple complementary methodologies. Traditional reference-based metrics, including BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), assess surface-level lexical overlap between system outputs and gold-standard references. However, such metrics often correlate poorly with human judgments of quality and relevance (Liu et al., 2016; Lowe et al., 2017; Xu et al., 2023; Fabbri et al., 2021; Ernst et al., 2023). More recent semantic-oriented metrics, such as BERTScore (Zhang et al., 2020) and AlignScore (Zha et al., 2023), use contextual embeddings to better capture semantic equivalence, offering improved sensitivity beyond surface similarity.

LLM-based judgment frameworks, particularly those employing the "LLM-as-a-Judge" paradigm, have demonstrated greater alignment with human evaluators (Zheng et al., 2023; Ashktorab et al., 2024; Hong et al., 2024; Ru et al., 2024; Gilardi et al., 2023). These techniques often break the evaluation process into finer-grained subtasks such as claim extraction and factual verification (Ru et al., 2024). Although promising, concerns remain about evaluator bias and model inconsistency (Schroeder and Wood-Doughty, 2024; Thakur et al., 2024). Our evaluation framework integrates both reference-based and LLM-based methods for a more comprehensive analysis of text quality and reliability.

These three strands of prior work collectively inform our methodology for building a reliable and interpretable medical QA system. By integrating structured decomposition strategies, constraint-driven generation, and multi-method evaluation, we tackle key challenges in producing trustworthy, clinically relevant outputs. This approach also supports future adaptability as techniques in each domain continue to evolve.

3 Methodology

To achieve rigorous and clinically reliable automation of message generation in healthcare, we present a methodology encompassing three sequential stages: (1) evidence classification using Large Language Models (LLMs), (2) generation of clinician-facing responses with iterative quality control, and (3) comprehensive evaluation across diverse medical datasets. This structured pipeline ensures transparency through principled processing and systematic validation. It ultimately supports robust clinical decision-making.

Prompt Templates To standardize and guide LLM behavior across each stage, we employ a suite of carefully designed prompt templates publicly available at GitHub[1]. These templates include:

- **Evidence Classification**: Categorizing relevant evidence segments from EHRs.
- **Grounded Question Answering**: Generating clinician responses grounded in classified evidence.
- **Answer Revision**: Refining responses through iterative feedback.

Evidence Classification We formulate evidence classification as a multiple-choice task, wherein the LLM assigns EHR evidence segments to one of three classes: *relevant*, *supplementary*, or *not relevant*. To ensure consistency in the output, categorical labels are constrained using Enum types (Willard and Louf, 2023). Additionally, to improve interpretability and encourage faithful predictions, the model is prompted to provide a rationale before selecting its final label (Wei et al., 2022).

Response Generation and Quality Control The LLM generates responses designed for clinicians that emphasize clarity, coherence, and professional tone after identifying relevant evidence. These outputs undergo a systematic quality assurance process based on metrics such as structural consistency, citation accuracy, and length. When deficiencies are detected, iterative feedback prompts the LLM to revise and improve outputs. This feedback loop enforces adherence to clinical communication standards.

[1] https://github.com/mo-arvan/grounded-medical-question-answering/tree/master/prompts

Evaluation Strategy Our evaluation strategy includes two phases: benchmarking foundational medical reasoning and assessing the full clinical message pipeline.

The first phase evaluates the LLM's performance using multiple-choice datasets closely aligned with our evidence classification framework: MMLU-Pro-Med, MedQA-US, MedMCQA, and PubMedQA (Wang et al., 2024; Jin et al., 2021; Pal et al., 2022; Jin et al., 2019). These datasets collectively measure domain-specific competency.

The second phase involves a comprehensive evaluation of the pipeline. This includes evidence classification, response generation and automated quality control applied to the ArchEHR-QA dataset (Soni and Demner-Fushman, 2025b,a), which is sourced from real-world EHR scenarios. Performance is assessed across two major dimensions:

Factuality is evaluated using Precision, Recall, and F1 Scores that compare the evidence cited in the generated responses to manually annotated ground-truth evidence. A "strict" Citation F1 considers only essential evidence, whereas a "lenient" variant also incorporates supplementary evidence.

Relevance is measured by comparing generated answers to essential EHR sentences and the original clinical question. Metrics employed include BLEU, ROUGE, SARI, BERTScore, AlignScore, and MEDCON.

Faithfulness Verification via Claim-Level Triple Extraction We introduce a custom, interpretable faithfulness metric grounded in claim-level triple extraction to evaluate factual consistency. Faithfulness, defined as the extent to which generated outputs accurately reflect source evidence, is a critical factor for clinical dependability (Ru et al., 2024). However, it is often difficult to measure due to incomplete references and the resource-intensive nature of expert reviews.

Our approach extracts atomic subject-predicate-object triples from both generated responses and their supporting EHR evidence. In a fungal infection case, the following triples, for example, are identified:

> ("Yeast", "was seen with", "bacteria on initial sputum gram stain"),
> ("Torulopsis glabrata", "was identified in", "blood/fungal culture"), and
> ("Antifungal therapy", "was started after", "fungal findings were confirmed").

In this example, the first two triples are supported by evidence, while the third lacks grounding. Each claim's support is verified by a separate LLM. Faithfulness is then quantified as the proportion of claims backed by evidence—in this case, 66.7%. This metric provides scalable and explainable factuality assessment.

Summary In summary, our methodology integrates structured prompt-guided classification, coherent response generation with iterative quality checks, and a rigorous evaluation framework. These include domain-specific benchmarks and interpretable factuality metrics. This design creates reliable, transparent, and extensible automation for generating clinical messages grounded in EHR data.

4 Results

This section presents our evaluation of model performance on two complementary tasks: general medical knowledge assessment and grounded clinical question answering. We first measure accuracy on standard multiple-choice benchmarks to assess general medical knowledge competence. We then evaluate the ability of the models to generate factually grounded and contextually relevant answers to clinical questions using the ArchEHR-QA dataset.

4.1 General Medical Knowledge Assessment

Table 1 summarizes accuracy scores across four established medical knowledge benchmarks. GPT-4o consistently outperforms both GPT-4o-mini and the baseline GPT-4† across all datasets.

In particular, GPT-4o achieves 77.67% on MedMCQA, marking an 8 percentage point improvement over GPT-4. On MedQA, it attains 88.69%, surpassing GPT-4 by 5 points. For MMLU-Pro-Med, GPT-4o sets a new state of the art with 81.56% accuracy. Although performance on PubMedQA is lower at 45.80%, this is expected due to the dataset's reliance on detailed comprehension of specialized biomedical literature. The lack of retrieval capabilities particularly challenges models in this setting.

4.2 Grounded Medical Question Answering

We next evaluate models on the ArchEHR-QA dataset, which benchmarks clinical question answering grounded in patient electronic health records. To ensure comparability with prior work, we use the official evaluation scripts provided by the challenge organizers.

Table 2 reports factuality and relevance scores for GPT-4o and GPT-4o-mini on both the development and test sets. GPT-4o achieves factuality scores of 51.85% (dev) and 51.59% (test), along with relevance scores of 29.96% and 33.33%, respectively. GPT-4o-mini scores 27.27% for factuality and 29.21% for relevance on the development set. As only GPT-4o was submitted to the challenge, test set outcomes for GPT-4o-mini are unavailable.

In addition to factuality and relevance, we assess response faithfulness. As shown in Table 3, GPT-4o attains 76.1% on the development set and 82.0% on the test set. GPT-4o-mini achieves a lower score of 65.6% on the development set.

These results collectively indicate that GPT-4o not only generates responses that are more accurate and pertinent but also maintains a strong alignment with provided clinical evidence.

5 Discussion

Our findings show that large generative models, such as GPT-4o, demonstrate superior performance on medical question-answering tasks, excelling across both knowledge-based and clinically grounded queries. Furthermore, the model is maintaining a high degree of factual consistency in evidence-grounded outputs.

Despite these advances, key trade-offs emerge between extractive and generative approaches. Evaluation metrics employed by ArchEHR-QA emphasize lexical overlap with reference texts, thereby favoring extractive methods. Generative models, by contrast, tend to produce more fluent and coherent responses but may not replicate the precise phrasing found in reference answers. To better capture the factual accuracy of generative outputs, we adopted a structured evaluation using the LLM-as-a-Judge framework. This approach enables scalable verification by assessing whether individual assertions in a generated response are supported by underlying evidence.

However, assessing factual consistency alone does not guarantee citation-level reliability. Recent studies highlight that large language models can incorrectly attribute statements to references that do not actually support them, introducing risks in high-stakes domains like healthcare. Notably, prior evidence shows that up to 30% of model-generated

Model	MedMCQA	MedQA	MMLU-Pro-Med	PubMedQA
GPT-4o	**77.67**	**88.69**	**81.56**	**45.80**
GPT-4o-mini	68.13	74.39	74.07	44.80
GPT-4†	69.88	83.97	-	39.60

Table 1: Performance comparison (accuracy %) across medical knowledge datasets. Results marked with † are baseline results from Xiong et al. (2024).

Model	Set	Factuality	Relevance
GPT-4o	Dev	51.85	29.96
GPT-4o	Test	51.59	33.33
GPT-4o-mini	Dev	27.27	29.21

Table 2: Factuality and relevance scores for GPT-4o and GPT-4o-mini on development and test sets of ArchEHR.

Model	Set	Faithfulness
GPT-4o	Dev	76.1
GPT-4o	Test	82.0
GPT-4o-mini	Dev	65.6

Table 3: Faithfulness scores for GPT-4o and GPT-4o-mini on development and test sets.

statements may contain unsupported reference citations (Wu et al., 2025). Ensuring that all cited sources genuinely substantiate the content remains a critical challenge.

Given these limitations, the safe and responsible deployment of LLMs in clinical environments requires comprehensive validation and routine monitoring using scalable methods like those employed in this study. Importantly, expert human review remains essential, particularly in scenarios where accuracy and reliability are paramount for patient safety.

A practical advantage in the clinical setting is that real-time response generation is often not required. This relaxed time constraint allows the system to proactively generate multiple candidate questions and corresponding answers for each incoming patient message in advance. Consequently, clinicians are not burdened with crafting questions themselves and can instead select from a curated list of contextually appropriate Q&A pairs. This workflow-integrated approach streamlines clinical decision-making and promotes more efficient patient communication.

Looking ahead, integrating external knowledge retrieval with interactive clinical tools presents a promising avenue to enhance both model performance and usability. Future research should also examine the impact of such systems on key outcomes, including clinician workload, as existing evidence in this area remains mixed (Garcia et al., 2024). In addition, comprehensive human preference studies comparing outputs from extractive and generative systems will be essential to align evaluation frameworks with the practical expectations and needs of clinicians.

6 Conclusion

Our work shows that generative models such as GPT-4o perform well across a range of clinical question answering tasks. These models also demonstrate strong factual alignment with source evidence when evaluated using structured, claim-level assessment methods.

However, several important challenges remain. Distinguishing between claim faithfulness, which assesses whether individual assertions align with evidence, and citation faithfulness, which considers whether referenced sources support the claims, continues to be difficult. In addition, label consistency and the design of evaluation frameworks require further improvement to ensure more reliable assessments.

Addressing these challenges, together with incorporating direct feedback from clinicians, is essential for enabling trustworthy and effective deployment of these models in real-world biomedical settings.

Limitations

One important challenge identified during Phase 1 of our evaluation involved testing models on general medical knowledge benchmarks, including MMLU-Pro-Med (Wang et al., 2024), MedQA-US (Jin et al., 2021), MedMCQA (Pal et al., 2022), and PubMedQA (Jin et al., 2019). A central limitation in this context is the lack of transparency

surrounding the training data used by commercial large language models. Without clear documentation of training corpora, there is a significant risk of data leakage, where benchmark content may inadvertently overlap with training inputs. This overlap can lead to inflated performance metrics, which misrepresent a model's generalizability and complicate direct comparisons between models. Because these benchmarks aim to assess a broad range of medical knowledge and reasoning skills, even partial contamination reduces the credibility of conclusions drawn from model performance. Although commercial LLMs exhibit strong capabilities, the opacity of their training data sources remains a fundamental barrier to reproducible and trustworthy evaluation. This limitation underscores the need for greater dataset transparency or the development of evaluation strategies that explicitly control for training-evaluation separation.

In addition, this study did not include expert validation of the model-generated responses. Due to time constraints, we were unable to engage licensed medical professionals in a systematic review process. While our structured framework incorporates LLM-as-a-Judge assessments, the absence of expert oversight limits our ability to confirm the clinical accuracy and safety of model outputs. Future work should incorporate formal expert evaluation to ensure that responses meet professional standards and are suitable for use in healthcare settings.

Ethical Considerations

The system was developed using Azure OpenAI Services in accordance with PhysioNet's responsible use guidelines [2]. We avoided using any protected health information during development.

References

Anthropic. 2025. Building effective agents. Accessed: 2025-02-18.

Zahra Ashktorab, Michael Desmond, Qian Pan, James M. Johnson, Martin Santillan Cooper, Elizabeth M. Daly, Rahul Nair, Tejaswini Pedapati, Swapnaja Achintalwar, and Werner Geyer. 2024. Aligning human and LLM judgments: Insights from evalassist on task-specific evaluations and ai-assisted assessment strategy preferences. *CoRR*, abs/2410.00873.

JW Ayers, A Poliak, M Dredze, EC Leas, Z Zhu, JB Kelley, DJ Faix, AM Goodman, CA Longhurst, M Hogarth, and 1 others. 2023. Comparing physician and artificial intelligence chatbot responses to patient questions posted to a public social media forum. jama intern med 183 (6): 589–596.

Dana Brin, Vera Sorin, Akhil Vaid, Ali Soroush, Benjamin S Glicksberg, Alexander W Charney, Girish Nadkarni, and Eyal Klang. 2023. Comparing chatgpt and gpt-4 performance in usmle soft skill assessments. *Scientific Reports*, 13(1):16492.

Ori Ernst, Ori Shapira, Ido Dagan, and Ran Levy. 2023. Re-examining summarization evaluation across multiple quality criteria. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13829–13838, Singapore. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryscinski, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir R. Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Trans. Assoc. Comput. Linguistics*, 9:391–409.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *CoRR*, abs/2312.10997.

Patricia Garcia, Stephen P Ma, Shreya Shah, Margaret Smith, Yejin Jeong, Anna Devon-Sand, Ming Tai-Seale, Kevin Takazawa, Danyelle Clutter, Kyle Vogt, and 1 others. 2024. Artificial intelligence–generated draft replies to patient inbox messages. *JAMA Network Open*, 7(3):e243201–e243201.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *CoRR*, abs/2303.15056.

Giwon Hong, Aryo Pradipta Gema, Rohit Saxena, Xiaotang Du, Ping Nie, Yu Zhao, Laura Perez-Beltrachini, Max Ryabinin, Xuanli He, Clémentine Fourrier, and Pasquale Minervini. 2024. The hallucinations leaderboard - an open effort to measure hallucinations in large language models. *CoRR*, abs/2404.05904.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577, Hong Kong, China. Association for Computational Linguistics.

[2] https://physionet.org/news/post/gpt-responsible-use

Yubin Kim, Chanwoo Park, Hyewon Jeong, Yik Siu Chan, Xuhai Xu, Daniel McDuff, Hyeonhoon Lee, Marzyeh Ghassemi, Cynthia Breazeal, and Hae Won Park. 2024. Mdagents: An adaptive collaboration of llms for medical decision-making. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, and 1 others. 2023. Performance of chatgpt on usmle: potential for ai-assisted medical education using large language models. *PLoS digital health*, 2(2):e0000198.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2122–2132. The Association for Computational Linguistics.

Ryan Lowe, Michael Noseworthy, Iulian Vlad Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic Turing test: Learning to evaluate dialogue responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1116–1126, Vancouver, Canada. Association for Computational Linguistics.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of GPT-4 on medical challenge problems. *CoRR*, abs/2303.13375.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Proceedings of the Conference on Health, Inference, and Learning*, volume 174 of *Proceedings of Machine Learning Research*, pages 248–260. PMLR.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Cheng Jiayang, Cunxiang Wang, Shichao Sun, Huanyu Li, Zizhao Zhang, Binjie Wang, Jiarong Jiang, Tong He, Zhiguo Wang, Pengfei Liu, Yue Zhang, and Zheng Zhang. 2024. Ragchecker: A fine-grained framework for diagnosing retrieval-augmented generation. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Kayla Schroeder and Zach Wood-Doughty. 2024. Can you trust LLM judgments? reliability of llm-as-a-judge. *CoRR*, abs/2412.12509.

Tait D Shanafelt, Colin P West, Lotte N Dyrbye, Mickey Trockel, Michael Tutty, Hanhan Wang, Lindsey E Carlasare, and Christine Sinsky. 2022. Changes in burnout and satisfaction with work-life integration in physicians during the first 2 years of the covid-19 pandemic. In *Mayo Clinic Proceedings*, volume 97, pages 2248–2258. Elsevier.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2023. Hugginggpt: Solving AI tasks with chatgpt and its friends in hugging face. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Kumar Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathaneal Schärli, Aakanksha Chowdhery, Philip Andrew Mansfield, Blaise Agüera y Arcas, Dale R. Webster, and 11 others. 2022. Large language models encode clinical knowledge. *CoRR*, abs/2212.13138.

Sarvesh Soni and Dina Demner-Fushman. 2025a. A dataset for addressing patient's information needs related to clinical course of hospitalization. *arXiv preprint*.

Sarvesh Soni and Dina Demner-Fushman. 2025b. Overview of the archehr-qa 2025 shared task on grounded question answering from electronic health records. In *The 24th Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, Vienna, Austria. Association for Computational Linguistics.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2024. Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. *CoRR*, abs/2406.12624.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding

benchmark. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Brandon T Willard and Rémi Louf. 2023. Efficient guided generation for llms. *arXiv preprint arXiv:2307.09702*, 6.

Kevin Wu, Eric Wu, Kevin Wei, Angela Zhang, Allison Casasola, Teresa Nguyen, Sith Riantawan, Patricia Shi, Daniel Ho, and James Zou. 2025. An automated framework for assessing how well llms cite relevant medical references. *Nature Communications*, 16(1):3615.

Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking retrieval-augmented generation for medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6233–6251, Bangkok, Thailand. Association for Computational Linguistics.

Fangyuan Xu, Yixiao Song, Mohit Iyyer, and Eunsol Choi. 2023. A critical evaluation of evaluations for long-form question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 3225–3245. Association for Computational Linguistics.

Yifan Yang, Qiao Jin, Qingqing Zhu, Zhizheng Wang, Francisco Erramuspe Álvarez, Nicholas Wan, Benjamin Hou, and Zhiyong Lu. 2024. Beyond multiple-choice accuracy: Real-world challenges of implementing large language models in healthcare. *CoRR*, abs/2410.18460.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. Alignscore: Evaluating factual consistency with A unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 11328–11348. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023. Siren's song in the AI ocean: A survey on hallucination in large language models. *CoRR*, abs/2309.01219.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Hongjian Zhou, Boyang Gu, Xinyu Zou, Yiru Li, Sam S. Chen, Peilin Zhou, Junling Liu, Yining Hua, Chengfeng Mao, Xian Wu, Zheng Li, and Fenglin Liu. 2023. A survey of large language models in medicine: Progress, application, and challenge. *CoRR*, abs/2311.05112.