MeDiSumQA: Patient-Oriented Question-Answer Generation from Discharge Letters

Amin Dada[1], Osman Alperen Koraş[1], Marie Bauer[1], Amanda Butler Contreras[2], Kaleb E Smith[2], Jens Kleesiek[1,3,4,5], Julian Friedrich[1],
[1]Institute for AI in Medicine (IKIM), University Hospital Essen, Germany
[2]NVIDIA, Santa Clara, USA
[3]Cancer Research Center Cologne Essen (CCCE), University Medicine Essen, Germany
[4]German Cancer Consortium (DKTK, Partner site Essen), Germany
[5]Department of Physics, TU Dortmund, Germany
Correspondence: amin.dada@uk-essen.de

Abstract

While increasing patients' access to medical documents improves medical care, this benefit is limited by varying health literacy levels and complex medical terminology. Large language models (LLMs) offer solutions by simplifying medical information. However, evaluating LLMs for safe and patient-friendly text generation is difficult due to the lack of standardized evaluation resources. To fill this gap, we developed **MeDiSumQA**. **MeDiSumQA** is a dataset created from MIMIC-IV discharge summaries through an automated pipeline combining LLM-based question-answer generation with manual quality checks. We use this dataset to evaluate various LLMs on patient-oriented question-answering. Our findings reveal that general-purpose LLMs frequently surpass biomedical-adapted models, while automated metrics correlate with human judgment. By releasing **MeDiSumQA** on PhysioNet, we aim to advance the development of LLMs to enhance patient understanding and ultimately improve care outcomes.

1 Introduction

Access to health documents empowers patients and improves medical care (Greene and Hibbard, 2012; Lye et al., 2018; Ross and Lin, 2003). These documents, however, often use language too complex for patients to understand (Paasche-Orlow et al., 2005), and physicians have no time to simplify documents in a patient-friendly manner (Ammenwerth and Spötl, 2009).

This gap between healthcare providers and patients can be bridged by large language models (LLMs) (Ali et al., 2023; Jeblick et al., 2024; Zaretsky et al., 2024; Eisinger et al., 2025). Through their ability to simplify medical information, LLMs can enhance the access to health documents and ultimately improve patient care. However, assessing and comparing LLMs in their ability to generate safe and patient-friendly text remains challenging due to the lack of benchmarks and publicly available resources. Strict privacy regulations surrounding clinical data limit dataset accessibility, thereby impeding the development of open benchmarks for evaluating LLMs in medical contexts.

To address this issue, we developed **MeDiSumQA**. **MeDiSumQA** is a novel, patient-oriented question-answering (QA) dataset, a format especially suitable to improve patient understanding of clinical documents (Cai et al., 2023).

In this paper, we describe how we created, curated, and evaluated **MeDiSumQA**, crafting a standardized resource for future benchmarking. By making this task openly available to researchers, we support broader development and testing of LLMs for healthcare applications, helping address challenges of time constraints and health literacy.

2 Related Work

While several clinical QA datasets exist (Pampari et al., 2018; Lehman et al., 2022; Soni et al., 2022; Bardhan et al., 2022; Dada et al., 2024b; Kweon et al., 2024), none, to the best of our knowledge, are explicitly designed for patient-oriented use.

Prior research has explored medical text simplification, but did not focus on helping patients understand clinical documents in a QA format. Aali et al. (2024) developed a public dataset that converts MIMIC hospital course summaries into concise discharge letters. Campillos-Llanos et al. (2022) created a Spanish dataset for simplifying clinical trial texts, demonstrating the importance of multilingual resources. Trienes et al. (2022) focused on making pathology reports more understandable for patients, though their dataset remains private and does not address everyday clinical questions. Similarly, while Ben Abacha and Demner-Fushman (2019)'s MeQSum dataset transforms consumer health questions into brief medical queries, but is not based on clinical documents.

Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health), pages 124–136
May 4, 2025

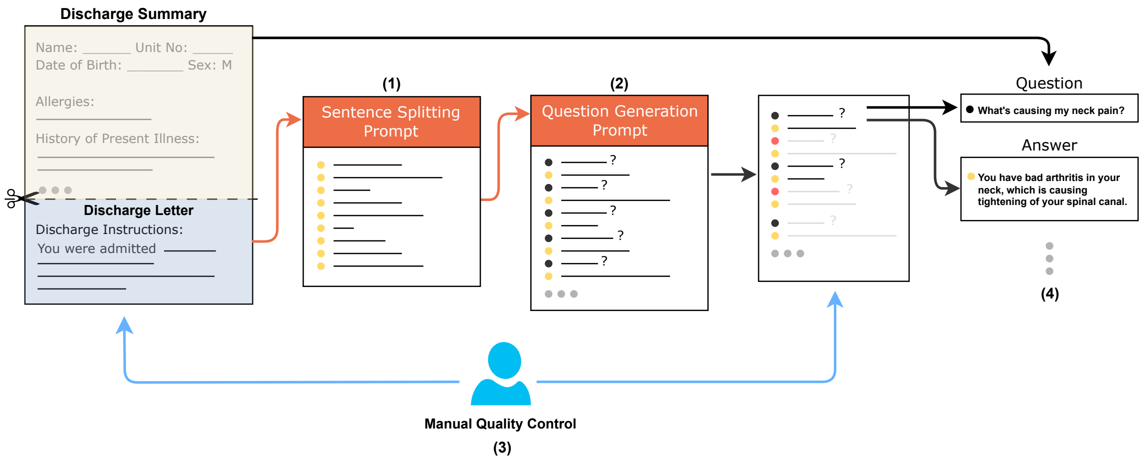

Figure 1: Generation process of **MeDiSumQA**. After identifying the discharge letter, we separate it from the main document and use an LLM to split it into sentences (1). Based on these sentences, we let an LLM generate matching questions (2). The resulting question-answer pairs were reviewed and curated by a physician, resulting in the the final **MeDiSumQA** dataset of 416 question-answer pairs (3). For inference, we provide LLMs with the discharge summary (without the bottom discharge letter) and pose the generated question. The model answer is then compared to the extracted ground truth answer (4).

Our work addresses these limitations by introducing a public, patient-centered QA dataset based on clinical MIMIC-IV discharge summaries, creating a benchmark to evaluate LLMs.

3 Methods

3.1 Dataset Generation

In the MIMIC-IV dataset (Johnson et al., 2023), some discharge summaries conclude with a discharge letter that summarizes key information and follow-up instructions in patient-friendly language. We used these discharge letters as the foundation for generating QA pairs in the following manner (Figure 1):

First, we identified discharge summaries containing discharge letters by searching for the string[1] that indicates the start of a discharge letter. We split each discharge letter into sentences using *Meta's Llama-3-70B-Instruct* (Dubey et al., 2024), which proved more accurate than traditional sentence splitters like NLTK, especially when handling irregular formatting and placeholders introduced by anonymization. To ensure accuracy, we prompted the LLM to preserve the original sentence structure and wording, which we subsequently verified by confirming that each processed sentence could be matched exactly with its source in the original discharge letter via exact string matching.

[1]"You were admitted to the hospital"

Afterwards, we fed these sentences into an *Meta's Llama-3-70B-Instruct* to generate matching questions from a patient's perspective. The LLM was allowed to reformulate the answer to match the question, but was instructed to reference the source sentence. We then manually checked these references to confirm that no information from the source document was altered. Since the answers are directly derived from the discharge letters written by medical professionals, this method maintains both medical accuracy and patient-friendly language. All mentioned prompts are listed in Appendix A.

The resulting QA candidates were then manually reviewed by a physician who selected high-quality examples based on the following criteria:

Factual correctness Question-answer pairs had to be logically connected. Answers that did not match their questions (e.g., "What medication should I avoid taking due to a possible allergy?" - "You were prescribed ibuprofen") were excluded.

Completeness Answers had to be complete. Partial answers (e.g., "What medications were started for me?" - "You were started on Vancomycin 1gm IV every 24 hours" when additional antibiotics were prescribed) were discarded.

Safety Answers needed to be safe. Potentially

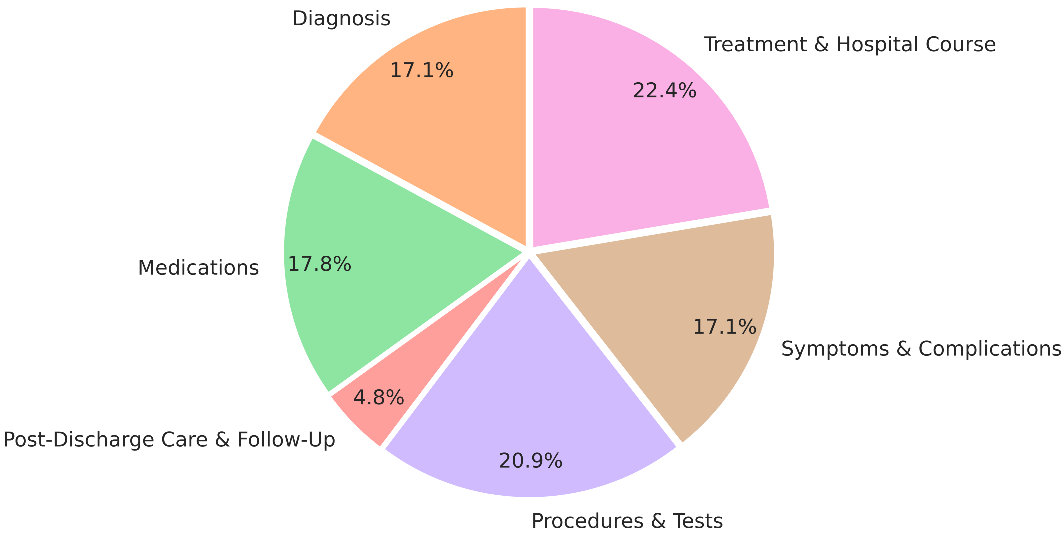

Figure 2: Frequency of question-answer categories in **MeDiSumQA**.

harmful instructions (e.g., "Take Coumadin 3 mg daily" without mentioning INR monitoring) were excluded.

Consistency Questions had to be answerable from both the discharge letter and discharge summary. Questions whose answers relied solely on information from the discharge letter were excluded.

Complexity Question-answer pairs had to be sufficiently complex. Obvious answers or overly specific questions that gave the answer away (e.g. "Did I receive Ciprofloxacin?" - "You received Ciprofloxacin.") were excluded.

As a final step, we removed the discharge letters from their summaries and combined the remaining summaries with their matching QA pairs. This resulted in three components, forming **MeDiSumQA**:

1. **A question** that serves as input for LLMs.
2. **An abbreviated discharge** summary without the discharge letter that LLMs use to answer the input question
3. **A ground truth answer** for comparison with generated responses

3.2 QA Categories

In **MeDiSumQA**, we identified six QA categories:

- Symptoms & Complications
- Procedures & Tests
- Diagnosis
- Treatment & Hospital Course
- Medications
- Post-Discharge Care & Follow-Up

To assign each QA pair to one of these categories, we used Meta's *Llama-3.3-70B-Instruct* (Dubey et al., 2024).

3.3 Evaluation

We evaluated the following models on **MeDiSumQA**: *Mistral-7B-Instruct-v0.1* (Jiang et al., 2023), *Meta-Llama-3-8B-Instruct*, *Meta-Llama-3.1-8B-Instruct* (Dubey et al., 2024), and four biomedical models derived from previously mentioned general-purpose language models: *BioMistral-7B* (Labrak et al., 2024), *Llama3-Med42-8B* (Christophe et al., 2024), *Llama3-Aloe-8B-Alpha* (Gururajan et al., 2024), and *Meditron3-8B* (OpenMeditron, 2024). We evaluated model performance on the **MeDiSumQA** dataset through automatic and manual assessments to ensure a comprehensive analysis.

3.3.1 Automatic Evaluation

We evaluated the models using established similarity metrics that capture both n-gram overlap and semantic similarity. The temperature was set to 1.0

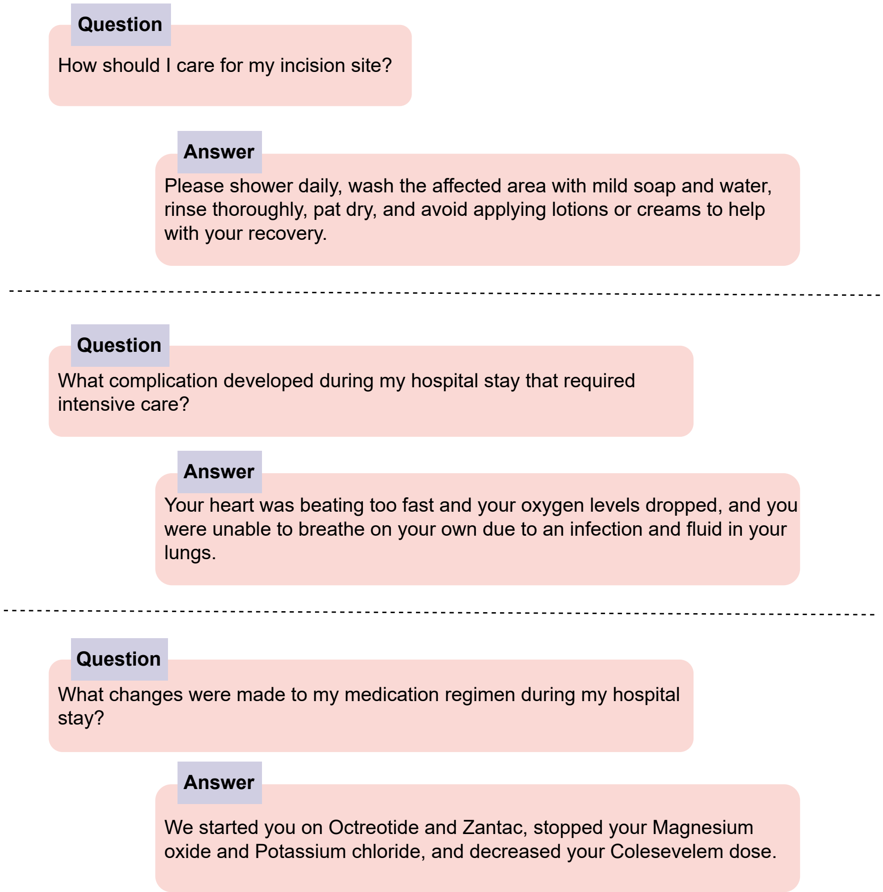

Figure 3: Example of QA pairs in **MeDiSumQA** dataset.

for all models. Due to the long input length, the models were prompted with a one-shot example. Additional details about the prompts are described in Appendix A.

Specifically, we used ROUGE-1, ROUGE-2, and ROUGE-L (Lin, 2004) to measure lexical overlap at varying levels of granularity, as well as BERT Score (Zhang* et al., 2020) to evaluate semantic similarity using contextual embeddings. For the BERT Score we tuned the rescaling baselines for MIMIC-IV discharge summaries using *Bio_ClinicalBERT* (Alsentzer et al., 2019). We also used the Unified Medical Language System (UMLS) parser *scispaCy* (Neumann et al., 2019) to assess the alignment of biomedical entities between predictions and ground truth answers, computing a UMLS F1 score.

As baselines for these metrics, we calculated both lower and upper bounds. To provide a lower bound for meaningful model predictions, we compute the similarity between the question and ground truth. For the upper bound, we paraphrased ground truth answers using *Llama-3.3-70B-Instruct* and measured their similarity to the original ground truth.

3.3.2 Manual Evaluation

To complement the automatic evaluation, we manually assessed 100 generated answers from two models: *Mistral-7B-Instruct-v0.1*, a lower-scoring model, and *Meta-Llama-3.1-8B-Instruct*, a higher-scoring model. For each model, we sorted the answers by the average similarity score across all automatic metrics. We then divided them into five equal-sized bins, with the lowest 20% placed in bin 1, the next 20% in bin 2, up to bin 5 containing the highest 20%. We then sampled ten predictions from each bin.

The answers were rated by a physician on five critical aspects:

- **Factuality:** Accuracy of medical information, rated on a scale from 1 to 5.

Model	Biomedical	Avg	R-L	R-1	R-2	BERT F1	UMLS F1
Lower Bound	-	20.93	13.11	15.76	2.82	60.22	12.74
Upper Bound	-	44.72	41.55	45.13	16.82	81.35	38.75
BioMistral-7B	Yes	23.69	15.1	19.67	5.29	64.24	14.13
Llama3-Med42-8B	Yes	29.27	21.2	26.84	8.65	68.45	21.21
Llama3-Aloe-8B-Alpha	Yes	19.47	8.94	12.11	3.81	61.83	10.66
Meditron3-8B	Yes	29.00	21.1	26.63	8.63	68.01	20.62
Mistral-7B-Instruct-v0.1	No	23.24	14.55	19.00	5.08	64.15	13.42
Meta-Llama-3-8B-Instruct	No	28.75	20.78	26.51	8.72	67.69	20.06
Meta-Llama-3.1-8B-Instruct	No	**31.43**	**24.1**	**29.93**	**10.24**	**69.35**	**23.55**

Table 1: Automatic evaluation of seven models on **MeDiSumQA**.

- **Brevity:** Conciseness of the response, rated on a scale from 1 to 5.
- **Patient-Friendliness:** Clarity and accessibility of the response for laypersons, rated on a scale from 1 to 5.
- **Relevance:** Alignment of the response with the question, rated on a scale from 1 to 5.
- **Safety:** Potential for harm or dissemination of misleading information, rated as a binary score (unsafe [0]/safe [1]).

Using the same sampling scheme and models, we collected 100 additional model-generated answers. These answers were then compared to their ground truth by a physician in a blinded fashion, indicating the preferred answer for each pair.

4 Results

4.1 MeDiSumQA Description

Initially, we generated 500 QA pairs, which were reduced to 416 pairs after manual curation. Figure 3 shows three examples of the resulting QA pairs.

Analysis of the QA categories in **MeDiSumQA** show a fairly even distribution across most categories (Figure 2). *Treatment & Hospital Course* make up the largest portion at 22.4%. *Procedures & Tests*, *Medications*, *Symptoms & Complications*, and *Diagnosis* each range between 17.1% and 20.9%. *Post-Discharge Care & Follow-Up* questions are notably underrepresented at only 4.8%.

4.2 Automatic Evaluation

Automatic evaluation across different LLMs reveals varying performance on **MeDiSumQA** (Table 1).

Meta-Llama-3.1-8B-Instruct performed best among all tested metrics, achieving the highest scores despite being a general-domain model without specific biomedical adaptation.

Comparing biomedical-adapted models with their general-domain counterparts reveals mixed results. Some biomedical adaptations showed only marginal improvements over their base models: *BioMistral-7B* marginally outperformed its base model *Mistral-7B-Instruct-v0.1* with a small increase of 0.45 points, while *Llama3-Med42-8B* showed a similar pattern with a slight improvement of 0.52 points over *Meta-Llama-3-8B-Instruct*.

However, several biomedical adaptations performed notably worse. Most striking is the case of *Llama3-Aloe-8B-Alpha*, which showed a substantial decrease of 9.28 points compared to its base model *Meta-Llama-3-8B-Instruct*. Similarly, *Meditron3-8B* exhibited a considerable decline of 2.43 points relative to *Meta-Llama-3.1-8B-Instruct*.

4.3 Manual Evaluation

Manual comparison of *Llama-3.1-8B-Instruct* and *Mistral-7B-Instruct-v0.1* across factuality, brevity, patient-friendliness, relevance, and safety revealed differences between the lower and higher scoring models (Figure 4).

In terms of factuality, *Llama-3.1-8B-Instruct* demonstrated consistently high performance, maintaining scores above 4.0 across all bins, with minimal variation. In contrast, *Mistral-7B-Instruct-v0.1* showed a gradual improvement from bin 1 (score 2.5) to bin 5 (score 4.3).

In the brevity metric, both models showed improved scores in higher bins. *Llama-3.1-8B-Instruct* maintained generally higher brevity scores throughout, starting at approximately 4.0 in bin

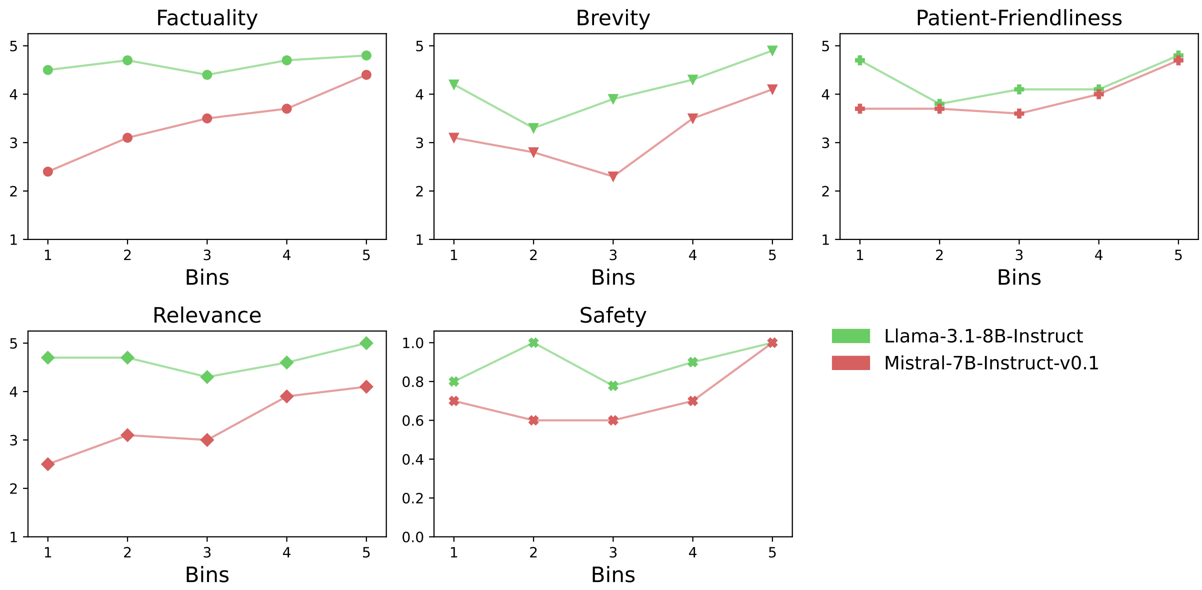

Figure 4: Physicians' evaluation of model generated answers on **MeDiSumQA**. Generated answers by *Llama-3.1-8B-Instruct* (green) and *Mistral-7B-Instruct-v0.1* (red) were sorted by their average automatic evaluation scores and divided into 5 bins. From each bin, 10 examples per model were sampled and rated by a physician across *Factuality*, *Brevity*, *Patient-Friendliness*, *Relevance*, and *Safety*. Each subplot displays scores either between 1 and 5 [*Factuality*, *Brevity*, *Patient-Friendliness*, *Relevance*] or 0 and 1 [*Safety*].

1 and reaching nearly 5.0 in bin 5. *Mistral-7B-Instruct-v0.1* displayed more variable performance, with a notable dip in bin 3 before recovering in bins 4 and 5.

Patient-friendliness scores converged for both models in the higher bins, with both achieving scores near 4.5 in bin 5. *Llama-3.1-8B-Instruct* showed initially higher scores in the lower bins, while *Mistral-7B-Instruct-v0.1* maintained relatively consistent scores around 3.5 before improving in the higher bins.

Regarding relevance, *Llama-3.1-8B-Instruct* consistently outperformed its counterpart, maintaining scores above 4.5 across all bins. *Mistral-7B-Instruct-v0.1* showed a gradual improvement from approximately 2.5 in bin 1 to 4.0 in bin 5.

Safety scores for both models were relatively high, with *Llama-3.1-8B-Instruct* showing slightly better performance, particularly in bins 2 and 3.

When a physician rated preferences between ground truth and model-generated answers, ground truth responses were generally preferred, though the patterns differed between models (Figures 5a, 5b.

For *Mistral-7B-Instruct-v0.1*, ground truth answers were strongly preferred across all bins, with model-generated answers favored only in exceptional cases.

For *Llama-3.1-8B-Instruct*, the results were more nuanced. Model-generated answers were preferred equally or slightly more often in cases with very high, but also with very low automatic similarity scores. In the middle ranges (bins 2, 3, and 4), ground truth answers were strongly preferred, though model-generated responses still garnered 10–40 % preference, with higher rates in the upper bins.

5 Discussion

Here, we introduce **MeDiSumQA**, a benchmark dataset designed to evaluate the ability of LLMs to answer clinical questions in a patient-friendly manner. By combining automatic and manual evaluations, our study provides insights into the strengths and limitations of LLMs for patient-oriented question answering, thus narrowing the gap between complex medical information and safe patient communication.

5.1 Characterization of the Dataset

MeDiSumQA provides a diverse and structured set of patient-oriented QA pairs derived from discharge summaries, covering key medical topics relevant to patient care. The category distribution of **MeDiSumQA** indicates comprehensive coverage across six major domains, with a particular empha-

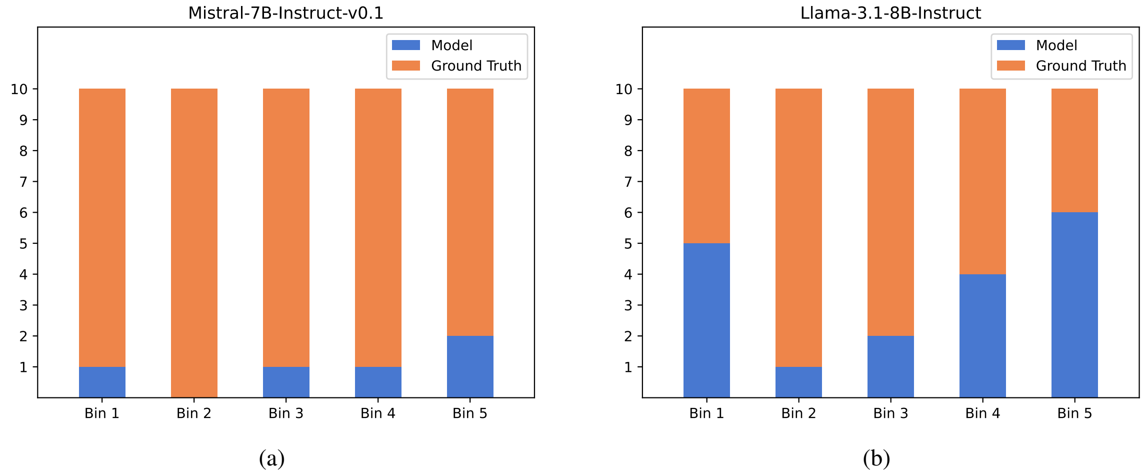

Figure 5: Physician preferences for answers generated by *Mistral-7B-Instruct-v0.1* (a) and *Llama-3.1-8B-Instruct* (b) and the ground truth answers.

sis on in-hospital care, medical interventions, and treatment courses. This suggests that the dataset aligns closely with the most immediate concerns patients may have after hospitalization, such as understanding their diagnosis, medications, and follow-up care.

While the dataset captures essential aspects of patient education, *Post-Discharge Care & Follow-Up* is underrepresented. This imbalance may reflect the structure of discharge summaries themselves, which tend to focus more on inpatient treatment rather than long-term care guidance. Expanding **MeDiSumQA** to include additional post-discharge documentation, such as outpatient follow-up notes or rehabilitation plans, could improve **MeDiSumQA**'s ability to support patient education beyond hospital stays.

5.2 Automatic Evaluation

MeDiSumQA requires LLMs to perform multiple skills simultaneously. Models must comprehend discharge summaries to understand patient cases, extract relevant details about hospital stays, and present this information in patient-friendly language. The discharge summaries are notably long, averaging 3,245.66 tokens with a standard deviation of 1,419.91, which is a significant challenge for LLMs due to the need for effective long-context reasoning (Li et al., 2024a). Furthermore, models must possess comprehensive medical knowledge and understanding of clinical guidelines to provide accurate follow-up advice. This complex task therefore evaluates an LLM's ability to integrate comprehension, information extraction, clear communication, and medical expertise in a patient-centered context.

Considering these antecedents, our evaluation shows that general-domain LLMs match or exceed the performance of specialized ones on biomedical tasks. Notably, *Meta-Llama-3.1-8B-Instruct* outperformed all tested biomedical domain-adapted models, raising questions about domain-specific training's effectiveness. While some biomedical models showed slight improvements over their base versions, others experienced significant performance declines, highlighting the inconsistent success of domain adaptation approaches.

These findings suggest that comprehensive pretraining on general-domain data may be more valuable than domain-specific adaptation. This challenges the conventional view that specialized tasks require domain-specific training, aligning with recent research questioning the effectiveness of biomedical adaptation (Dada et al., 2024a; Jeong et al., 2024; Dorfner et al., 2024).

5.3 Correlation of automatic and manual Evaluation

When comparing automatic with manual evaluation, our results show that calculated metrics like ROUGE and BERT Score correlate well with human judgment. Higher automated metric scores consistently corresponded to higher manual ratings and preferences, particularly for higher-scoring predictions. Conversely, answers from lower-performing models were rarely preferred by physicians and were sometimes deemed unsafe. This correlation between manual scores and physicians' assessments validates that LLMs can be well assessed in their capability to answer medical questions in a patient-friendly manner using **MeDiSumQA**.

However, manual assessment also reveals important limitations of automatic metrics, especially when models generated correct but different responses from the ground truth. Notably, in blind preference tests, *Llama-3.1-8B-Instruct* answers were sometimes preferred over ground truth answers, indicating that LLMs can generate valid alternative responses to the ground truth in **MeDiSumQA** that may be more appealing. Our manual evaluation also shows that LLMs favor safety over conciseness in their responses. These findings underscore the importance of combining human evaluation with automated scoring for thorough assessment in specialized healthcare applications.

5.4 Data Contamination

If evaluation datasets overlap with an LLM's training data, benchmark validity of these datasets is compromised due to data contamination (Li et al., 2024b; Deng et al., 2023). Such contamination can cause models to memorize rather than generalize, artificially inflating their performance. Although it is possible that some LLMs in our study have encountered parts of the MIMIC-IV dataset, this is unlikely since MIMIC-IV requires authentication for access.

A broader concern for datasets is intentional benchmark manipulation, when models are deliberately trained on evaluation datasets, which compromises dataset reliability. One solution is to generate datasets using private, inaccessible data. To facilitate this, we offer our dataset generation pipeline as open-source, allowing hospitals and other organizations to create confidential benchmarks from their own clinical reports. By releasing our **MeDiSumQA** code publicly, we enable others to develop independent datasets and conduct robust LLM evaluations using private medical data.

5.5 Outlook

We make **MeDiSumQA** available to the public, which offers an opportunity for widespread adoption in the medical AI community, enabling robust evaluations of models based on their ability to generate accurate, patient-friendly responses. This transparency can drive improvements in patient-centered AI by ensuring models are assessed against expert-validated benchmarks.

During manual evaluation, some model-generated answers were preferred over the ground truth. This presents an opportunity to refine the dataset by incorporating high-quality model-generated responses, with physicians selecting the most appropriate answers. As this approach could introduce bias toward LLMs used in the selection process, future versions of **MeDiSumQA** could involve multiple independent reviewers to ensure broader generalizability.

Lastly, expanding the dataset by applying our pipeline to a larger set of discharge summaries in different languages would enable use cases beyond single-language few-shot evaluation, including fine-tuning models for improved patient-oriented applications. Making the dataset more diverse and scalable will help develop safer, more effective AI-driven healthcare solutions.

6 Conclusion

MeDiSumQA represents another step toward enhancing patient understanding of medical documents by providing benchmarks to assess LLMs in answering medical questions in a patient-friendly manner. By evaluating models on both automated and human-centered metrics, our study demonstrates that automatic metrics correlate well with human judgment while also highlighting the potential of general-purpose LLMs in patient education. By making **MeDiSumQA** accessible on PhysioNet, we aim to foster further research into the applicability of LLMs for patient-oriented question answering and encourage advancements in this field. We hope that **MeDiSumQA** will serve as a valuable resource for the development of more patient-friendly AI systems, ultimately bridging the gap between complex medical information and safe, effective patient communication.

Limitations

Despite its strengths, **MeDiSumQA** presents challenges. The dataset primarily focuses on English-language discharge summaries, limiting its applicability to multilingual settings. Additionally, while automated metrics such as ROUGE and BERT Score provide valuable insights, our manual assessments reveal that these do not always align with human judgment, particularly in terms of brevity and relevance. Future research should explore more robust evaluation methods that incorporate real-world patient feedback.

References

Asad Aali, Dave Van Veen, Yamin Ishraq Arefeen, Jason Hom, Christian Bluethgen, Eduardo Pontes Reis, Sergios Gatidis, Namuun Clifford, Joseph Daws, Arash S Tehrani, et al. 2024. A dataset and benchmark for hospital course summarization with adapted large language models. *Journal of the American Medical Informatics Association*, page ocae312.

Stephen R Ali, Thomas D Dobbs, Hayley A Hutchings, and Iain S Whitaker. 2023. Using chatgpt to write patient clinic letters. *The Lancet Digital Health*, 5(4):e179–e181.

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Elske Ammenwerth and H-P Spötl. 2009. The time needed for clinical documentation versus direct patient care. *Methods of information in medicine*, 48(01):84–91.

Jayetri Bardhan, Anthony Colas, Kirk Roberts, and Daisy Zhe Wang. 2022. DrugEHRQA: A question answering dataset on structured and unstructured electronic health records for medicine related queries. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1083–1097, Marseille, France. European Language Resources Association.

Asma Ben Abacha and Dina Demner-Fushman. 2019. On the summarization of consumer health questions. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2228–2234, Florence, Italy. Association for Computational Linguistics.

Pengshan Cai, Zonghai Yao, Fei Liu, Dakuo Wang, Meghan Reilly, Huixue Zhou, Lingxi Li, Yi Cao, Alok Kapoor, Adarsha Bajracharya, et al. 2023. Paniniqa: Enhancing patient education through interactive question answering. *Transactions of the Association for Computational Linguistics*, 11:1518–1536.

Leonardo Campillos-Llanos, Ana R Terroba Reinares, Sofía Zakhir Puig, Ana Valverde-Mateos, and Adrián Capllonch-Carrión. 2022. Building a comparable corpus and a benchmark for spanish medical text simplification. *Procesamiento del Lenguaje Natural*, 69:189–196.

Clément Christophe, Praveen K Kanithi, Tathagata Raha, Shadab Khan, and Marco AF Pimentel. 2024. Med42-v2: A suite of clinical llms. *arXiv preprint arXiv:2408.06142*.

Amin Dada, Marie Bauer, Amanda Butler Contreras, Osman Alperen Koraş, Constantin Marc Seibold, Kaleb E Smith, and Jens Kleesiek. 2024a. Does biomedical training lead to better medical performance? *Preprint*, arXiv:2404.04067.

Amin Dada, Tim Leon Ufer, Moon Kim, Max Hasin, Nicola Spieker, Michael Forsting, Felix Nensa, Jan Egger, and Jens Kleesiek. 2024b. Information extraction from weakly structured radiological reports with natural language queries. *European Radiology*, 34(1):330–337.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. 2023. Investigating data contamination in modern benchmarks for large language models. *Preprint*, arXiv:2311.09783.

Felix J Dorfner, Amin Dada, Felix Busch, Marcus R Makowski, Tianyu Han, Daniel Truhn, Jens Kleesiek, Madhumita Sushil, Jacqueline Lammert, Lisa C Adams, et al. 2024. Biomedical large languages models seem not to be superior to generalist models on unseen medical data. *arXiv preprint arXiv:2408.13833*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Felix Eisinger, Friederike Holderried, Moritz Mahling, Christian Stegemann-Philipps, Anne Herrmann-Werner, Eric Nazarenus, Alessandra Sonanini, Martina Guthoff, Carsten Eickhoff, and Martin Holderried. 2025. What's going on with me and how can i better manage my health? the potential of gpt-4 to transform discharge letters into patient-centered letters to enhance patient safety: Prospective, exploratory study. *Journal of Medical Internet Research*, 27:e67143.

Jessica Greene and Judith H Hibbard. 2012. Why does patient activation matter? an examination of the relationships between patient activation and health-related outcomes. *Journal of general internal medicine*, 27:520–526.

Ashwin Kumar Gururajan, Enrique Lopez-Cuena, Jordi Bayarri-Planas, Adrian Tormos, Daniel Hinjos, Pablo Bernabeu-Perez, Anna Arias-Duart, Pablo Agustin Martin-Torres, Lucia Urcelay-Ganzabal, Marta Gonzalez-Mallo, et al. 2024. Aloe: A family of fine-tuned open healthcare llms. *arXiv preprint arXiv:2405.01886*.

Katharina Jeblick, Balthasar Schachtner, Jakob Dexl, Andreas Mittermeier, Anna Theresa Stüber, Johanna Topalis, Tobias Weber, Philipp Wesp, Bastian Oliver Sabel, Jens Ricke, et al. 2024. Chatgpt makes medicine easy to swallow: an exploratory case study on simplified radiology reports. *European radiology*, 34(5):2817–2825.

Daniel P Jeong, Saurabh Garg, Zachary Chase Lipton, and Michael Oberst. 2024. Medical adaptation of large language and vision-language models: Are we making progress? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12143–12170, Miami, Florida, USA. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. 2023. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.

Sunjun Kweon, Jiyoun Kim, Heeyoung Kwak, Dongchul Cha, Hangyul Yoon, Kwanghyun Kim, Seunghyun Won, and Edward Choi. 2024. Ehrnoteqa: A patient-specific question answering benchmark for evaluating large language models in clinical settings. *arXiv preprint arXiv:2402.16040*.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. BioMistral: A collection of open-source pretrained large language models for medical domains. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5848–5864, Bangkok, Thailand. Association for Computational Linguistics.

Eric Lehman, Vladislav Lialin, Katelyn Edelwina Legaspi, Anne Janelle Sy, Patricia Therese Pile, Nicole Rose Alberto, Richard Raymund Ragasa, Corinna Victoria Puyat, Marianne Katharina Taliño, Isabelle Rose Alberto, Pia Gabrielle Alfonso, Dana Moukheiber, Byron Wallace, Anna Rumshisky, Jennifer Liang, Preethi Raghavan, Leo Anthony Celi, and Peter Szolovits. 2022. Learning to ask like a physician. In *Proceedings of the 4th Clinical Natural Language Processing Workshop*, pages 74–86, Seattle, WA. Association for Computational Linguistics.

Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. 2024a. Long-context llms struggle with long in-context learning. *CoRR*, abs/2404.02060.

Yucheng Li, Frank Guerin, and Chenghua Lin. 2024b. An open source data contamination report for large language models. *Preprint*, arXiv:2310.17589.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Carolyn T Lye, Howard P Forman, Jodi G Daniel, and Harlan M Krumholz. 2018. The 21st century cures act and electronic health records one year later: will patients see the benefits? *Journal of the American Medical Informatics Association*, 25(9):1218–1220.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

OpenMeditron. 2024. Meditron3-8b model card.

Michael K Paasche-Orlow, Ruth M Parker, Julie A Gazmararian, Lynn T Nielsen-Bohlman, and Rima R Rudd. 2005. The prevalence of limited health literacy. *Journal of general internal medicine*, 20(2):175–184.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrQA: A large corpus for question answering on electronic medical records. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2357–2368, Brussels, Belgium. Association for Computational Linguistics.

Stephen E Ross and Chen-Tan Lin. 2003. The effects of promoting patient access to medical records: a review. *Journal of the American Medical Informatics Association*, 10(2):129–138.

Sarvesh Soni, Meghana Gudala, Atieh Pajouhi, and Kirk Roberts. 2022. RadQA: A question answering dataset to improve comprehension of radiology reports. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6250–6259, Marseille, France. European Language Resources Association.

Jan Trienes, Jörg Schlötterer, Hans-Ulrich Schildhaus, and Christin Seifert. 2022. Patient-friendly clinical notes: Towards a new text simplification dataset. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 19–27, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Jonah Zaretsky, Jeong Min Kim, Samuel Baskharoun, Yunan Zhao, Jonathan Austrian, Yindalon Aphinyanaphongs, Ravi Gupta, Saul B Blecker, and Jonah Feldman. 2024. Generative artificial intelligence to transform inpatient discharge summaries to patient-friendly language and format. *JAMA network open*, 7(3):e240357–e240357.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

A Prompts

Figures 6 and 7 show the prompts we use to split the discharge letter into sentences and generate question-answer pairs. For the question-answer generation we include a one shot example. Figure 8 shows the prompt we use to evaluate LLMs on **MeDiSumQA**.

System Prompt

Split the given text sentence by sentence by inserting newlines. Do not alter the text. Copy errors and mistakes. Only insert newline characters. Reply with the format '<split-text>...\n...\n...</split-text>'.

User Prompt

<text>{discharge_letter}</text>

Figure 6: Sentence Splitting prompt

System Prompt

You are a helpful assistant that is specialized in medical topics. You will be given a list of sentences that contain information about a patient. Based on the information in the statements, formulate challenging questions from the perspective of the patient. Also provide the answers which should be one or a combination of different given sentences.

1-shot example

User Prompt

Sentences:
1. You were admitted to the hospital for chest pain.
2. In the hospital you were ruled out for a heart attack and your pain was reproduced with touching the chest, suggesting a musculoskeletal cause of your discomfort.
3. You were discharged with a planned stress test as an outpatient.
4. There were no changes made to your medication regimen.

Assistant Response

Question: What was the initial reason for my hospital admission?
Answer: Chest pain (Statement 1)
Question: What did the hospital rule out as a cause of my chest pain, and what finding suggested an alternative cause?
Answer: The hospital ruled out a heart attack, by reproducing the pain with touching the chest, suggesting a musculoskeletal cause (Statement 2).
Question: What further evaluation should I plan after discharge from the hospital?
Answer: Stress test as an outpatient (Statement 3)
Question: Was there any change in my medication regimen during the hospital stay?
Answer: No (Statement 4)

User Prompt

Sentences:
{sentences}

Figure 7: MeDiSumQA generation prompt

System Prompt

You are a highly skilled assistant, specifically trained to assist lays in understanding and extracting key information from medical documents. Your primary responsibility will be to interprete discharge letters from hospitals. You will receive such a discharge letter. You should carefully review the contents and accurately answer questions related to this document. Only respond with the correct answer to the question. Answer briefly without mentioning a lot of specific details. If the question is about measurements (e.g., lab values), interpret their meaning in relation to the question, rather than writing down the values. Do not generate anything else.

1-shot example

User Prompt

--------------BEGIN DISCHARGE LETTER--------------
{discharge_summary}
--------------END DISCHARGE LETTER--------------
Question: What was the outcome of my virtual colonoscopy?

Assistant Response

Answer: We did not find any polyps, masses, or signs of inflammatory disease in your examination.

User Prompt

--------------BEGIN DISCHARGE LETTER--------------
{discharge_summary}
--------------END DISCHARGE LETTER--------------
What side effect did I experience from taking Clozapine, and how was it managed?

Figure 8: MedisumQA Inference