Reading Between the Lines: Toward Translating Verbose Patient-authored Messages into Clinician-Formulated Questions

Sarvesh Soni, Madeline Bittner, Dina Demner-Fushman
Division of Intramural Research
National Library of Medicine, National Institutes of Health, Bethesda, MD, USA
{sarvesh.soni, madeline.bittner}@nih.gov, ddemner@mail.nih.gov

Abstract

Patient portal messages often embed clinical questions inside long, emotionally nuanced narratives, requiring clinicians to infer the underlying information need. We study the task of rewriting verbose patient-authored narratives into concise, clinician-interpreted questions framed as if querying an electronic health record (EHR) system. We evaluate a lightweight LLM-based rewrite pipeline that constrains outputs to 10–15 words and uses rule-based validation with regeneration. We test the approach on 140 distinct patient questions drawn from the ArchEHR-QA dataset and shared task. Each system output is double-annotated by two annotators for quality (Good/Ok/Bad) and error types (Generic, Malformed, Tangential, Hallucination). Results show that while models follow output constraints, they often produce overly generic or tangential questions, and occasional hallucinations introduce unsupported clinical details. Across both clinician-question and patient-narrative comparison settings, automatic metrics show substantial overlap across human quality labels; in pairwise meta-evaluation, BERTScore is the strongest proxy for human preferences. We release our code and annotations to support future work[1].

1 Introduction

Patients frequently contact healthcare teams via patient portals with questions about tests, diagnoses, and care plans (Tai-Seale et al., 2017). In practice, these questions are often embedded in lengthy narratives with extraneous detail, uncertainty, and emotion (Table 1), forcing clinicians to "read between the lines" to determine what the patient is actually asking (Dymek et al., 2021; Yan et al., 2021). Translating patient-authored narratives into concise clinician-interpreted questions is key for downstream systems: it clarifies the clinical information needs, enables targeted retrieval of supporting electronic health record (EHR) evidence, and reduces the risk that automated assistance addresses the wrong intent. We view this reformulation as an auxiliary representation that complements, rather than replaces, the original patient message.

Patient Question: Hi I ... started shaking ... woken up ... confused ... been told that there is a shadow on my brain ...
Clinician-interpreted Question: What caused the patient's confusion and short-term memory loss?
Output #1 (**Quality:** *Ok*; **Error:** *Generic*): What are possible causes of a brain shadow after seizure activity?
Output #2 (**Quality:** *Bad*; **Error:** *Hallucination*): What is the likely cause of the new cerebral lesion on imaging?

Table 1: Abbreviated example of question reformulation and human evaluation; full version in Appendix Table 5.

Prior work on medical question summarization targets generic consumer health queries answerable from general medical knowledge (Yadav et al., 2022). In contrast, answers to patients' EHR-related questions depend on hospitalization records. Therefore, reformulation must (i) remain faithful to patient intent, (ii) support patient-record retrieval, and (iii) avoid medical details not supported by the message. Moreover, existing methods rarely model patient-to-clinician perspective shift. Prior natural language processing work on portal messages has emphasized triage or generic response generation (Anderson et al., 2025; Biro et al., 2025); however, reformulating EHR-grounded patient narratives into clinician-style questions is underexplored. We make three contributions:

1. We present a simple large language model (LLM)-based system with constrained output, rule-based validation, and regeneration for patient-to-clinician question reformulation.

2. We conduct double-annotated human evaluation of LLM outputs, rating quality (*good*,

[1] https://github.com/soni-sarvesh/patient-question-reformulation

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 481–489

ok, or *bad*) and categorizing errors (*generic*, *malformed*, *tangential*, *hallucination*).

3. We analyze automatic metrics in two comparison settings (against the reference clinician question and the patient-authored narrative) and meta-evaluate which metrics best track human preferences for model comparison.

2 Related Work

Medical question summarization. Generating clinician-interpreted questions from patient narratives is question summarization, which is widely studied for consumer health questions answerable from general medical knowledge rather than patient-specific records (Ben Abacha et al., 2021).

EHR question answering (QA) and information retrieval (IR). Most clinical QA and IR work focuses on clinician-authored questions over EHR content (Bardhan et al., 2024; Sivarajkumar et al., 2024). Differently, ArchEHR-QA (Soni and Demner-Fushman, 2026a) is a publicly available resource that pairs patient-authored narratives with clinician-interpreted questions and annotated EHR evidence for patient-specific QA. We build on this line of work by isolating the upstream *reformulation* step, focusing on translating patient language into a clinician-formulated query suitable for downstream retrieval and grounding.

Patient portal messaging. Prior work on patient portal messages has emphasized message classification and triage to reduce clinician burden and improve routing (Tafti et al., 2019; Ren et al., 2023; Anderson et al., 2025). More recently, large language models (LLMs) have been integrated for drafting responses to patient messages, raising both efficiency opportunities and safety concerns (Chen et al., 2024; Garcia et al., 2024; Biro et al., 2025). Clinician question reformulation is complementary: by surfacing the inferred clinical information need, it can support safer downstream retrieval of patient-specific evidence and grounded response drafting.

3 Methods

3.1 Data

We use patient-authored questions and corresponding clinician-interpreted questions from ArchEHR-QA (Soni and Demner-Fushman, 2026a), which links questions from public health forums about recent hospitalizations to de-identified clinical notes from MIMIC (Johnson et al., 2016, 2023). To preserve clinical fidelity, the clinical notes are unedited; the patient questions are minimally edited to match surface details in the associated note (e.g., pronoun changes such as "he" to "she"). In total, we evaluate on 140 distinct patient questions. This set combines the 108 unique questions in the original ArchEHR-QA release (134 question–note instances before deduplication) with 32 additional questions from the ArchEHR-QA 2026 Shared Task data (Soni and Demner-Fushman, 2026b).

3.2 Model

We use a simple LLM-based generation approach with rule-based validation. Given a patient-authored narrative, an instruction-tuned LLM is prompted to produce a concise clinician-interpreted question that captures inferred clinical intent and is suitable for downstream EHR-oriented use. The prompt constrains the output to a single 10–15-word clinician question to promote uniformity (Table 4). Validation uses deterministic string processing to extract the clinician question from the prompted markdown format and verify that exactly one non-empty question is present. If extraction fails (i.e., the generated output does not conform to the prompted format), the same prompt is reused for up to five attempts; the first generation that passes the format check is retained. In our runs, every retained output passed validation within five attempts. This rule-based check only enforces output format, not the 10–15-word length; the length constraint is specified in the prompt and influences generation, but is not strictly enforced post-hoc. Beyond this, we do not use retrieval, external medical knowledge, or hand-crafted rewriting rules. We evaluate two open-weight LLMs with different parameter scales: Llama 3.1 8B and 3.3 70B (Grattafiori et al., 2024); decoding and inference settings are reported in Appendix A.1.

3.3 Evaluation

Manual evaluation. Three annotators with clinical/biomedical expertise (an MD, a clinical informaticist, and a biomedical researcher) evaluated model outputs for *Quality* (*Good*: faithful and specific; *Ok*: usable but imperfect; *Bad*: unusable) and assigned *Error* labels as needed. We used four non-mutually-exclusive error categories: *Generic* (too general), *Malformed* (ill-formed), *Tangential* (not focused on the core patient intent), and *Hallucination* (unsupported details). Annotators could view

3-class *Good / Ok / Bad*		Binary {*Good, Ok*} / *Bad*	
P_o	κ	P_o	κ
36.8%	0.162	71.1%	0.234

Table 2: Inter-rater agreement on *Quality* annotations for 3-class (*Good*/*Ok*/*Bad*) and a binary recoding (*Acceptable*/*Bad*). Exact match P_o and Cohen's κ are reported.

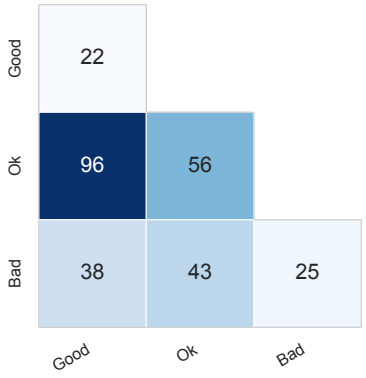

Figure 1: Pairwise agreement/disagreement matrix for *Quality* ratings between annotators ($N = 280$). Diagonal cells show agreement counts; off-diagonal cells show disagreement counts for each label pair.

the reference clinician question, but overlap with the reference was not a rating criterion, since multiple reformulations could be valid; annotators were blinded to model identity. All 280 outputs (140 cases×2 models) were double-annotated. For the first 50 cases (100 outputs), annotators discussed disagreements to calibrate guideline interpretation, but did not revise labels for subjective disagreements; the remaining cases were annotated independently, and we retain both labels without adjudicating subjective disagreements.

Automated evaluation. We compute ROUGE-L (Lin, 2004), BERTScore (Zhang et al., 2019), Medcon (Yim et al., 2023), and AlignScore (Zha et al., 2023) between each generated question and two comparison texts: the reference clinician question and the patient-authored narrative. These metrics capture lexical overlap (ROUGE), embedding similarity (BERTScore), clinical concept overlap (Medcon), and factual consistency (AlignScore).

4 Results and Discussion

The average lengths of the patient narratives and reference clinician-interpreted questions are 100.4 and 11.9 words, respectively. The Llama 8B and

Set-level (multi-label) agreement	
Metric	**Score**
Mean Jaccard similarity	0.283
Exact match rate (%)	19.3

Per-error-type agreement		
Label	P_o(%)	κ
Generic	45.7	−0.086
Malformed	93.2	0.629
Tangential	60.7	0.070
Hallucination	84.6	0.230

Table 3: Inter-rater agreement for *Error* annotations. Set-level agreement for multi-label error sets uses mean Jaccard similarity and exact match; per-error-type agreement reports percent agreement P_o and Cohen's κ.

70B models generated questions averaging 15.5 and 11.5 words, respectively. Since the 10–15-word constraint is conveyed via the prompt but not enforced by validation, these averages reflect how closely each model follows the prompted length instruction. The 70B model adheres tightly to the prompted range: 95% of its outputs fall within 10–15 words (with the remaining 5% slightly under 10). The 8B model follows the lower bound reliably (100% of outputs are at least 10 words) but frequently exceeds the upper bound, with only 48.6% of outputs in the 10–15 range.

Inter-rater Agreement. On *Quality*, annotators agreed on 36.8% of instances, increasing up to 71.1% with binary recoding (Table 2). This is consistent with a subjective boundary between *Good* and *Ok* and with the difficulty of judging adequacy; Figure 1 shows the distribution of disagreements, and Cohen's κ indicates slight agreement.

For *Error* labels, set-level agreement is moderate with the mean Jaccard similarity of 0.283, while the exact match is 19.3% (Table 3). Per-label agreement varies widely: *Malformed* errors are easiest to identify with a percent agreement (P_o) of 93.2%, whereas *Generic* is hardest ($P_o = 45.7\%$), likely because judgments about whether a reformulation is overly broad or insufficiently specific are more subjective than judgments about surface forms.

Inter-rater variability is also common in clinical interpretation itself, where specialists may legitimately disagree (Novack et al., 2006). Related open-ended biomedical annotation tasks, including Medical Subject Headings (MeSH) indexing and evaluation of generated clinical notes, also report modest inter-rater agreement (Fernandez-Llimos, 2025; Moramarco et al., 2022). We therefore pre-

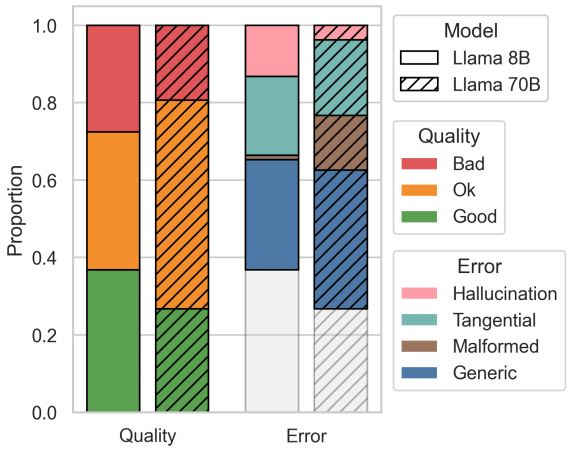

Figure 2: Human judgments of *Quality* and *Error* for questions generated by Llama 8B and 70B models.

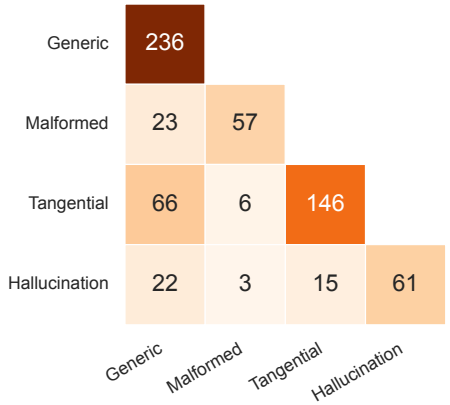

Figure 3: *Error*-label co-occurrence matrix computed over all annotator-output instances (no adjudication). Diagonal: count of each error label; off-diagonal: within-instance co-selections of error pairs.

serve these annotations as complementary annotator judgments rather than as a single gold label.

Human ratings. Figure 2 summarizes *Quality* and *Error* judgments for both models. Overall, the 70B model produces fewer *Bad* outputs, but many generations remain *Ok*, reflecting that partially correct intent capture may still be too generic or miss patient-specific nuance. Both models exhibit substantial *Generic* and *Tangential* errors; the 8B model shows a higher fraction of *Hallucination*-labeled errors, while the 70B model shows relatively more *Malformed*-labeled errors.

Error co-occurrence. Figure 3 characterizes error frequency and co-occurrence across annotated *Error* label instances (error categories are non-mutually exclusive). *Generic* is the dominant failure mode (236 instances), followed by *Tangential* (146), with *Hallucination* (61) and *Malformed* (57) occurring less frequently. *Generic* frequently co-occurs with *Tangential* (66 instances), suggesting that a model latches onto a peripheral detail and produces a broad question about it rather than the most clinically relevant information need.

Hallucinations are less common but clinically important. As shown in Table 1, hallucinations can introduce unsupported details (e.g., assuming a "lesion" when the patient only mentions a "shadow"), which would misdirect downstream EHR retrieval and response drafting. Hallucination co-occurs more often with *Generic* (22) or *Tangential* (15) than with *Malformed* (3), suggesting that hallucinations are often fluent and well-formed, making them harder to detect using format checks alone.

Automatic metrics and human quality judgments. ROUGE, BERTScore, Medcon, and AlignScore stratified by human *Quality* labels are shown in Appendix Figures 4 and 5. When computed against the reference clinician question, automatic metrics align only weakly with human *Quality* ratings. BERTScore shows the clearest upward shift for both 8B and 70B models (higher median for *Good* than *Bad*/*Ok*). ROUGE and AlignScore show heavy overlap across *Bad*/*Ok*/*Good*, and the median scores for *Good* are not consistently higher than those for *Ok* or *Bad*. Medcon is largely zero-dominated with high variance and overlap across quality levels, consistent with sparse concept overlap when only a single reference clinician question is available and when acceptable reformulations may use different clinically meaningful terms. Thus, automated metrics may flag very low-scoring outliers but do not reliably distinguish reformulations with different *Quality* labels and may miss clinically consequential errors (e.g., fluent hallucinations), reinforcing the need for expert evaluation and richer multi-reference benchmarks. Using the patient-authored narrative instead yields the same broad conclusion: metric score distributions still overlap substantially across quality labels. However, the clinician-question setting appears somewhat more informative overall, most clearly for BERTScore, while differences for the other metrics are smaller or less consistent.

Automatic metrics and human pairwise preferences. We meta-evaluate automatic metrics against human pairwise preferences, following

prior work (Freitag et al., 2024), on the subset of cases where the two systems received different human quality ratings (Appendix Table 6). When computed against the reference clinician question, BERTScore achieves the highest overall pairwise accuracy (PA; 61.0%) and is the best metric in each quality-pair bucket, whereas ROUGE-L and AlignScore remain below 50% PA overall. When computed against the patient narrative, BERTScore still ranks first (52.0%), but the gap to ROUGE-L and AlignScore narrows (both 48.0%). Medcon remains weaker, though its PA increases from 26.8% with clinician questions to 40.7% with patient narratives. Overall, BERTScore is the strongest proxy for human preferences in our setting.

5 Conclusion

In a simple LLM-based constrained generation approach to translating verbose patient-authored narratives into concise clinician-interpreted questions, with outputs assessed through double-annotated human evaluation, we found that fluent rewrites are often *generic* or *tangential*, and occasional *hallucinations* introduce unsupported clinical details. Automatic metrics do not cleanly separate more useful reformulations from less useful ones; in pairwise meta-evaluation, BERTScore provides the most informative signal for relative model comparison among the metrics we tested.

Limitations

First, clinician interpretation of a patient narrative is inherently subjective, and we do not adjudicate disagreements; instead, we retain multi-annotator labels. While this better reflects variation in clinician reasoning, it complicates learning and evaluation with a single "gold" label. Second, each instance currently provides only one reference clinician-interpreted question, even though multiple rewrites may be equally acceptable. This is mitigated by the judgments on the generated questions that could now be used as paraphrases in future evaluations. Third, we evaluate two open-weight models from a single family (Meta Llama 3.1 8B and 3.3 70B). This is a deliberate scope choice: the paper's contribution is a characterization of reformulation error modes and a meta-evaluation of automatic metrics, rather than a leaderboard-style model comparison. The specific error rates and metric rankings may not transfer to other open-weight, proprietary, or clinically adapted LLMs. Fourth, the 10–15-word output constraint was chosen to promote uniformity and to discourage models from echoing patient narrative verbatim, but it can force compression of multi-part patient concerns into a single question. The reference clinician questions in ArchEHR-QA average 11.9 words, and themselves often collapse compound concerns; we did not separately quantify how often important nuance is lost by the constraint. Relaxing or removing the length cap, or allowing multiple sub-questions, is worth exploring in future work. Fifth, the evaluation set contains 140 patient cases, so small between-model differences should not be over-interpreted. Sixth, we report one final valid generation per model–instance pair under a fixed prompting and decoding setup. The pipeline is intentionally simple and is not tuned to maximize task performance through extensive prompt engineering, retrieval augmentation, or model adaptation. Stronger systems or averaged multi-sample evaluations could change the absolute performance levels and possibly the relative behavior of the automatic metrics. Finally, our evaluation isolates the reformulation step and does not measure downstream effects such as evidence retrieval or answer grounding; assessing these is a natural next step.

Ethical considerations

This work uses a publicly available, de-identified dataset and releases additional annotations for research. Nevertheless, the task involves clinical language and could be misused if deployed without oversight. In particular, hallucinated clinician questions could misdirect evidence retrieval or clinical reasoning. We position our work as a benchmark and analysis of failure modes, and any real-world deployment should include clinician-in-the-loop review, monitoring for hallucinations, and safeguards against inappropriate clinical decision-making.

Acknowledgments

This research was supported by the Intramural Research Program of the National Institutes of Health (NIH) and utilized the computational resources of the NIH HPC Biowulf cluster (http://hpc.nih.gov). The contributions of the NIH authors are considered Works of the United States Government. The findings and conclusions presented in this paper are those of the authors and do not necessarily reflect the views of the NIH or the U.S. Department of Health and Human Services.

References

Blake J. Anderson, Muhammad Zia ul Haq, Yuanda Zhu, Andrew Hornback, Alison D. Cowan, Michelle Mott, Bradley Gallaher, and Arash Harzand. 2025. Development and Evaluation of a Model to Manage Patient Portal Messages. *NEJM AI*, 2(3):AIoa2400354.

Jayetri Bardhan, Kirk Roberts, and Daisy Zhe Wang. 2024. Question Answering for Electronic Health Records: Scoping Review of Datasets and Models. *Journal of Medical Internet Research*, 26(1):e53636.

Asma Ben Abacha, Yassine Mrabet, Yuhao Zhang, Chaitanya Shivade, Curtis Langlotz, and Dina Demner-Fushman. 2021. Overview of the MEDIQA 2021 Shared Task on Summarization in the Medical Domain. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 74–85, Online. Association for Computational Linguistics.

Joshua M. Biro, Jessica L. Handley, J. Malcolm McCurry, Adam Visconti, Jeffrey Weinfeld, J. Gregory Trafton, and Raj M. Ratwani. 2025. Opportunities and risks of artificial intelligence in patient portal messaging in primary care. *npj Digital Medicine*, 8(1):1–6.

Shan Chen, Marco Guevara, Shalini Moningi, Frank Hoebers, Hesham Elhalawani, Benjamin H. Kann, Fallon E. Chipidza, Jonathan Leeman, Hugo J. W. L. Aerts, Timothy Miller, Guergana K. Savova, Jack Gallifant, Leo A. Celi, Raymond H. Mak, Maryam Lustberg, Majid Afshar, and Danielle S. Bitterman. 2024. The effect of using a large language model to respond to patient messages. *The Lancet Digital Health*, 0(0).

Christine Dymek, Bryan Kim, Genevieve B Melton, Thomas H Payne, Hardeep Singh, and Chun-Ju Hsiao. 2021. Building the evidence-base to reduce electronic health record–related clinician burden. *Journal of the American Medical Informatics Association*, 28(5):1057–1061.

Fernando Fernandez-Llimos. 2025. Consistency of Medical Subject Headings assignment: A test-retest reliability analysis. *Research in Social and Administrative Pharmacy*, 21(10):784–789.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs Breaking MT Metrics? Results of the WMT24 Metrics Shared Task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

Patricia Garcia, Stephen P. Ma, Shreya Shah, Margaret Smith, Yejin Jeong, Anna Devon-Sand, Ming Tai-Seale, Kevin Takazawa, Danyelle Clutter, Kyle Vogt, Carlene Lugtu, Matthew Rojo, Steven Lin, Tait Shanafelt, Michael A. Pfeffer, and Christopher Sharp. 2024. Artificial Intelligence–Generated Draft Replies to Patient Inbox Messages. *JAMA Network Open*, 7(3):e243201.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 Herd of Models. *Preprint*, arXiv:2407.21783.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Liwei H. Lehman, Leo A. Celi, and Roger G. Mark. 2023. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Liwei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):160035.

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.

Francesco Moramarco, Alex Papadopoulos Korfiatis, Mark Perera, Damir Juric, Jack Flann, Ehud Reiter, Anya Belz, and Aleksandar Savkov. 2022. Human Evaluation and Correlation with Automatic Metrics in Consultation Note Generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5739–5754, Dublin, Ireland. Association for Computational Linguistics.

Victor Novack, Lone S. Avnon, Alexander Smolyakov, Rachel Barnea, Alan Jotkowitz, and Francisc Schlaeffer. 2006. Disagreement in the interpretation of chest radiographs among specialists and clinical outcomes of patients hospitalized with suspected pneumonia. *European Journal of Internal Medicine*, 17(1):43–47.

Yang Ren, Dezhi Wu, Aditya Khurana, George Mastorakos, Sunyang Fu, Nansu Zong, Jungwei Fan, Hongfang Liu, and Ming Huang. 2023. Classification of Patient Portal Messages with BERT-based Language Models. In *2023 IEEE 11th International Conference on Healthcare Informatics (ICHI)*, pages 176–182.

Sonish Sivarajkumar, Haneef Ahamed Mohammad, David Oniani, Kirk Roberts, William Hersh, Hongfang Liu, Daqing He, Shyam Visweswaran, and Yanshan Wang. 2024. Clinical Information Retrieval: A Literature Review. *Journal of Healthcare Informatics Research*, 8(2):313–352.

Sarvesh Soni and Dina Demner-Fushman. 2026a. A Dataset for Addressing Patient's Information Needs

related to Clinical Course of Hospitalization. *Scientific Data*, 13(1):523.

Sarvesh Soni and Dina Demner-Fushman. 2026b. Overview of the ArchEHR-QA 2026 shared task on grounded question answering from electronic health records. In *Proceedings of the Third Workshop on Patient-Oriented Language Processing (Cl4health)*, Palma, Mallorca (Spain). ELRA.

Ahmad P. Tafti, Sunyang Fu, Aditya Khurana, George M. Mastorakos, Kenneth G. Poole, Stephen J. Traub, James A. Yiannias, and Hongfang Liu. 2019. Artificial intelligence to organize patient portal messages: A journey from an ensemble deep learning text classification to rule-based named entity recognition. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1380–1387.

Ming Tai-Seale, Cliff W. Olson, Jinnan Li, Albert S. Chan, Criss Morikawa, Meg Durbin, Wei Wang, and Harold S. Luft. 2017. Electronic Health Record Logs Indicate That Physicians Split Time Evenly Between Seeing Patients And Desktop Medicine. *Health Affairs*, 36(4):655–662.

Shweta Yadav, Deepak Gupta, Asma Ben Abacha, and Dina Demner-Fushman. 2022. Question-aware transformer models for consumer health question summarization. *Journal of Biomedical Informatics*, 128:104040.

Qi Yan, Zheng Jiang, Zachary Harbin, Preston H Tolbert, and Mark G Davies. 2021. Exploring the relationship between electronic health records and provider burnout: A systematic review. *Journal of the American Medical Informatics Association*, 28(5):1009–1021.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. Aci-bench: A Novel Ambient Clinical Intelligence Dataset for Benchmarking Automatic Visit Note Generation. *Scientific Data*, 10(1):586.

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)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

A Implementation Details

This section reports decoding settings and the complete generation prompt.

You are provided with a patient question under `[PATIENT NARRATIVE]`.

`[PATIENT NARRATIVE]:`

`{...}`

Your task is to generate a clear and concise clinician question that a clinician would interpret from the patient narrative. Assume that the clinician is asking this question to a smart electronic health record system to help themselves in formulating a response to the patient. Restrict the length of the clinician question between 10 to 15 words.

Format instructions: Output your response in a markdown format, containing only the clinician question. For example:

```clinician-question
This is a clinician-version of the patient question.
```

Table 4: Complete prompt used to automatically generate the clinician-interpreted questions.

A.1 Hyperparameters

We use fixed decoding parameters for all runs (HuggingFace Transformers text-generation pipeline): do_sample=True, temperature=0.9, top-p=0.6, top-k=50, repetition_penalty=1.05, and max_new_tokens=512; we do not perform hyperparameter search. For each model and patient narrative, we generate a single final clinician question (with up to five retries only when the output fails format/length validation); we do not sample multiple valid generations or average across random seeds.

A.2 Prompt

Table 4 gives the complete prompt used to generate clinician-interpreted questions.

B Supplementary Example

Table 5 provides the full version of the abbreviated example in Table 1.

C Supplementary Analyses

This section provides supplementary figures, tables, and analyses that complement the results in the main text.

C.1 Quality-rating disagreements

Figure 1 shows that disagreements were mostly between adjacent labels (*Good* vs. *Ok*: 96; *Bad* vs. *Ok*: 43) rather than the extremes (*Good* vs. *Bad*: 38), supporting the use of multi-annotator labels rather than a single adjudicated "gold" label.

Patient Question: Hi I was in hospital in a critical care unit after having a fit and started shaking on one side of my body. I spaced out. I have woken up but I am confused, I don't remember what happened. I have been told that there is a shadow on my brain what could this be
Clinician-interpreted Question: What caused the patient's confusion and short-term memory loss?
System Response #1: What are possible causes of a brain shadow after seizure activity? **Quality:** *Ok* | **Error:** *Generic* **Reason:** Generated question is general.
System Response #2: What is the likely cause of the new cerebral lesion on imaging? **Quality:** *Bad* | **Error:** *Hallucination* **Reason:** Patient question does not mention a lesion.

Table 5: Complete example of patient-to-clinician question reformulation and human evaluation of model outputs.

C.2 Automatic metric distributions

Figures 4 and 5 show distributions of ROUGE-L, BERTScore, Medcon, and AlignScore stratified by human *Quality* labels, computed against the reference clinician-authored question and the patient-authored narrative, respectively.

C.3 Pairwise metric meta-evaluation

Table 6 reports pairwise accuracy results for automatic metrics against human preferences under both comparison settings.

Additional observations. Table 6 also shows a few additional trends. In the clinician question setting, Medcon yields ties for more than half of the comparisons (72/123), although its accuracy rises to 64.7% when ties are excluded. These ties likely reflect Medcon's frequent zero scores, since the compared texts are short and often share few or no clinical entities. In the patient narrative setting, Medcon still produces many ties (27/123), although fewer than in the clinician question setting. We also observe greater asymmetry by the human-preferred system in the patient narrative setting, suggesting that some metrics are more sensitive to cases where the 70B output is preferred than to cases favoring the 8B output.

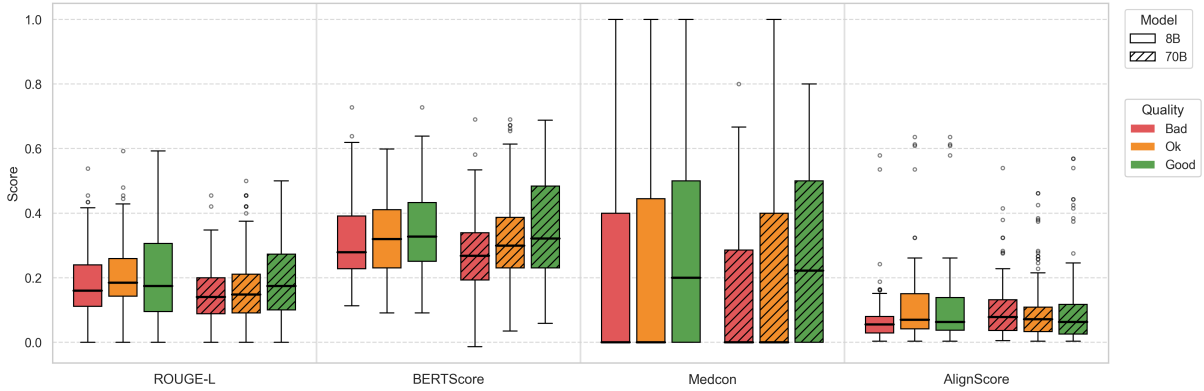

Figure 4: Distributions of automated metric scores computed between each system-generated question and the reference clinician-authored question, stratified by human *Quality* label, for outputs from Llama 8B and 70B.

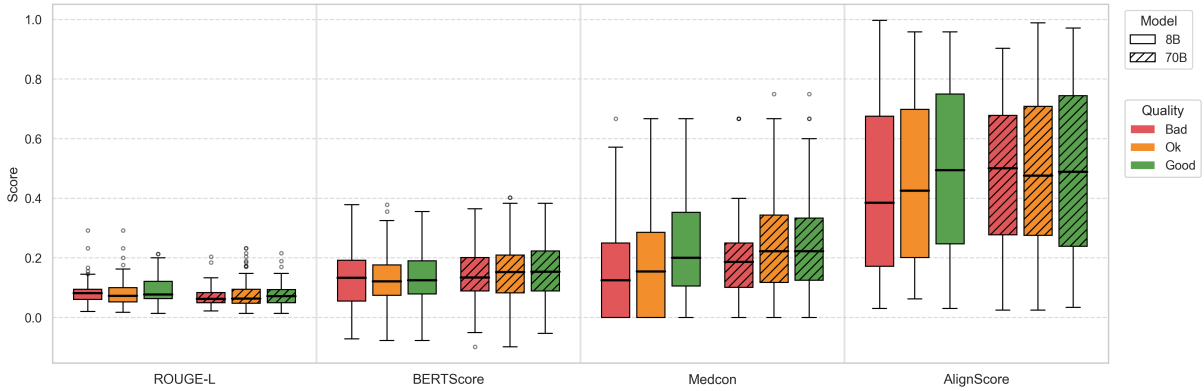

Figure 5: Distributions of automated metric scores computed between each system-generated question and the original patient-authored narrative, stratified by human *Quality* label, for outputs from Llama 8B and 70B.

Metric	Overall			PA by human winner		PA by quality pair		
	PA *N=123*	*Ties* *N=123*	*PA w/o ties* *N=123 − ties*	*8B* *N=66*	*70B* *N=57*	*good/bad* *N=28*	*good/ok* *N=58*	*ok/bad* *N=37*
Using Clinician Question								
ROUGE-L	49.6	2	50.4	51.5	47.4	60.7	44.8	48.6
BERTScore	**61.0**	0	61.0	**63.6**	**57.9**	**71.4**	**56.9**	**59.5**
Medcon	26.8	72	**64.7**	25.8	28.1	21.4	34.5	18.9
AlignScore	48.0	0	48.0	48.5	47.4	60.7	43.1	45.9
Using Patient Narrative								
ROUGE-L	48.0	4	49.6	**43.9**	52.6	**53.6**	43.1	51.4
BERTScore	**52.0**	0	52.0	39.4	**66.7**	**53.6**	**44.8**	**62.2**
Medcon	40.7	27	**52.1**	30.3	52.6	42.9	37.9	43.2
AlignScore	48.0	0	48.0	36.4	61.4	39.3	**44.8**	59.5

Table 6: Meta-evaluation of automatic metrics against human pairwise preferences on a subset of cases where the two models received different human quality ratings. PA denotes pairwise accuracy. All entries except *Ties* are percentages (%); denominators are shown below the column headers. PA treats metric ties as disagreements, whereas PA w/o ties is computed on the non-tied subset only. Columns 8B and 70B report PA conditioned on which system the human preferred, and columns good/bad, good/ok, and ok/bad report PA within each human quality-pair bucket. Bold indicates the highest overall PA within each setting.