LCDS: A Logic-Controlled Discharge Summary Generation System Supporting Source Attribution and Expert Review

Cheng Yuan[1*], **Xinkai Rui**[1,2*], **Yongqi Fan**[1], **Yawei Fan**[1], **Boyang Zhong**[1], **Jiacheng Wang**[1], **Weiyan Zhang**[1†], **Tong Ruan**[1†]
[1]East China University of Science and Technology, Shanghai 200237, China
[2]Ruijin Hospital, Shanghai Jiaotong University School of Medicine, Shanghai 200025, China
{ruantong,weiyanzhang}@ecust.edu.cn

Abstract

Despite the remarkable performance of Large Language Models (LLMs) in automated discharge summary generation, they still suffer from hallucination issues, such as generating inaccurate content or fabricating information without valid sources. In addition, electronic medical records (EMRs) typically consist of long-form data, making it challenging for LLMs to attribute the generated content to the sources. To address these challenges, we propose **LCDS**, a **L**ogic-**C**ontrolled **D**ischarge **S**ummary generation system. LCDS constructs a source mapping table by calculating textual similarity between EMRs and discharge summaries to constrain the scope of summarized content. Moreover, LCDS incorporates a comprehensive set of logical rules, enabling it to generate more reliable silver discharge summaries tailored to different clinical fields. Furthermore, LCDS supports source attribution for generated content, allowing experts to efficiently review, provide feedback, and rectify errors. The resulting golden discharge summaries are subsequently recorded for incremental fine-tuning of LLMs. Our project and demo video are in the GitHub repository https://github.com/ycycyc02/LCDS.

1 Introduction

The discharge summary (DS) is the final section of an electronic medical record (EMR) that consolidates essential patient information, such as admission details, medical history, diagnoses, treatments, medications, and follow-up recommendations (Xiong et al., 2019). It plays a critical role in ensuring continuity of patient care, facilitating communication between healthcare providers and patients, and supporting clinical decisions (Lenert et al., 2014; Kripalani et al., 2007; Li et al., 2013; Walraven et al., 2002). Traditionally, discharge summaries are manually written by physicians, making the process time-consuming, labor-intensive, and susceptible to subjective biases (Xu et al., 2024; Hartman et al., 2023; Rink et al., 2023). Recently, large language models (LLMs) have shown great promise in automating discharge summary generation by leveraging retrieval, reasoning, and fine-tuning techniques (Van Veen et al., 2024). For example, Liu et al. (2022) propose Re3Writer, which simulates physician workflows through medical knowledge retrieval and reasoning. Similarly, Lyu et al. (2024) integrate extractive methods with generative techniques, combining named entity recognition (NER) and prompt-tuned text generation.

Despite these advancements, several critical challenges remain in automated discharge summary generation using LLMs.

Precise Content Localization: EMRs typically consist of long-form, complex, and heterogeneous data spanning multiple sections (Wu et al., 2024). Directly feeding complete EMRs into LLMs can exceed their context limits, thus degrading the quality of generated summaries and increasing interference from irrelevant or redundant information.

Accuracy and Hallucination Reduce: Although LLMs demonstrate remarkable performance, they still suffer from hallucination issues, generating inaccurate or fabricated content lacking valid sources (Maynez et al., 2020; Zhang et al., 2023b; Ji et al., 2023). In the medical domain, this can significantly compromise patient safety and care quality. Effective strategies to impose logical constraints to mitigate these hallucinations remain underexplored.

Adaptability to Different Clinical Departments: While discharge summaries share a general structure across medical specialties, their detailed content requirements vary significantly. Current automated generation methods often lack adapt-

[*]Equal Contribution.
[†]Co-corresponding Author.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), pages 284–294

ability to specific departmental needs, risking the omission of crucial clinical information.

Traceability and Trustworthiness: As discharge summaries directly influence patient care decisions, medication guidance, and follow-up treatments, ensuring content traceability is essential. However, current LLM-based generation systems lack explicit source attribution mechanisms, making it challenging for medical professionals to verify and trust the generated content.

To address these challenges, we propose A **LCDS** (**L**ogic-**C**ontrolled **D**ischarge **S**ummary Generation) System, featuring source attribution, logical constraints, and expert review:

- **Source Mapping for Precise Content Localization**: LCDS constructs a source mapping table by calculating textual similarity between EMRs and discharge summaries, effectively constraining content selection and enhancing summary accuracy.
- **Logic-Controlled Summary Generation**: LCDS incorporates structured prompts guided by medical-domain logical rules, significantly improving factual accuracy and reducing hallucinations in generated discharge summaries.
- **Attribution-Based Expert Review**: LCDS segments generated summaries at the sentence level, explicitly attributing content to original EMR sources. This mechanism supports expert verification, facilitates error correction, and enhances clinical reliability.

Our system implements all proposed functionalities, demonstrating a complete pipeline for discharge summary generation from EMRs. Moreover, we conducted experiments using real-world clinical data from 15 medical departments. Experimental results show that LCDS outperforms existing methods in terms of accuracy, coherence, and clinical applicability of the generated discharge summaries, significantly reducing hallucinations and improving content traceability.

2 Related Work

Existing methods for automatic DS generation fall into three categories:

Extraction-Abstracting Methods: These methods first extract key information from medical records and then generate summaries, aiming to balance traceability and textual fluency. Representative studies include (Shing et al., 2021; VC et al., 2023; K et al., 2021). While such approaches enhance factual accuracy, they heavily rely on the quality of the source text, making them prone to information omission.

Knowledge-Enhanced Methods: This category integrates external knowledge bases or retrieval-augmented techniques to improve the reliability of summaries. Examples include reinforcement learning-based medical entity verification (Zhang et al., 2020), embedded entity retrieval alignment (Adams et al., 2024), and a three-step generate framework comprising retrieval, reasoning, and synthesis (Liu et al., 2022). However, these methods are computationally complex and constrained by the timeliness of the knowledge base.

LLM-Based Methods: These approaches leverage prompt engineering or fine-tuning techniques to adapt large models for medical applications. (Clough et al., 2024) has shown that GPT-4 and its variants can generate summaries approaching physician-level quality. However, as noted by (Williams et al., 2024; Dubinski et al., 2024; Kim et al., 2024), the generated content still requires human review to ensure clinical accuracy. Additionally, LLMs are prone to hallucinations, potentially producing misleading or erroneous information. The lack of a clear provenance mechanism further complicates the verification of generated summaries by medical professionals.

3 System Workflow and Usage Example

This section introduces the system's usage and functionality through case studies. As shown in Figure 2, the workflow consists of four steps:

Input EMR Format Conversion: LCDS converts various types of EMR documents uploaded by users into a unified JSON format, ensuring data consistency and standardization.

Reference-Guided Source-Aware Discharge Summary Generation: Key content is extracted from standardized EMRs, and a "Silver" DS is generated based on refined logical field constraints.

Attribution-Based Comparison and Review: LCDS aligns each sentence in the summary with the original EMR, allowing experts to review, compare, and modify content for a high-quality "Gold" Discharge Summary.

Iterative Optimization: Review feedback and finalized discharge summaries create an incremen-

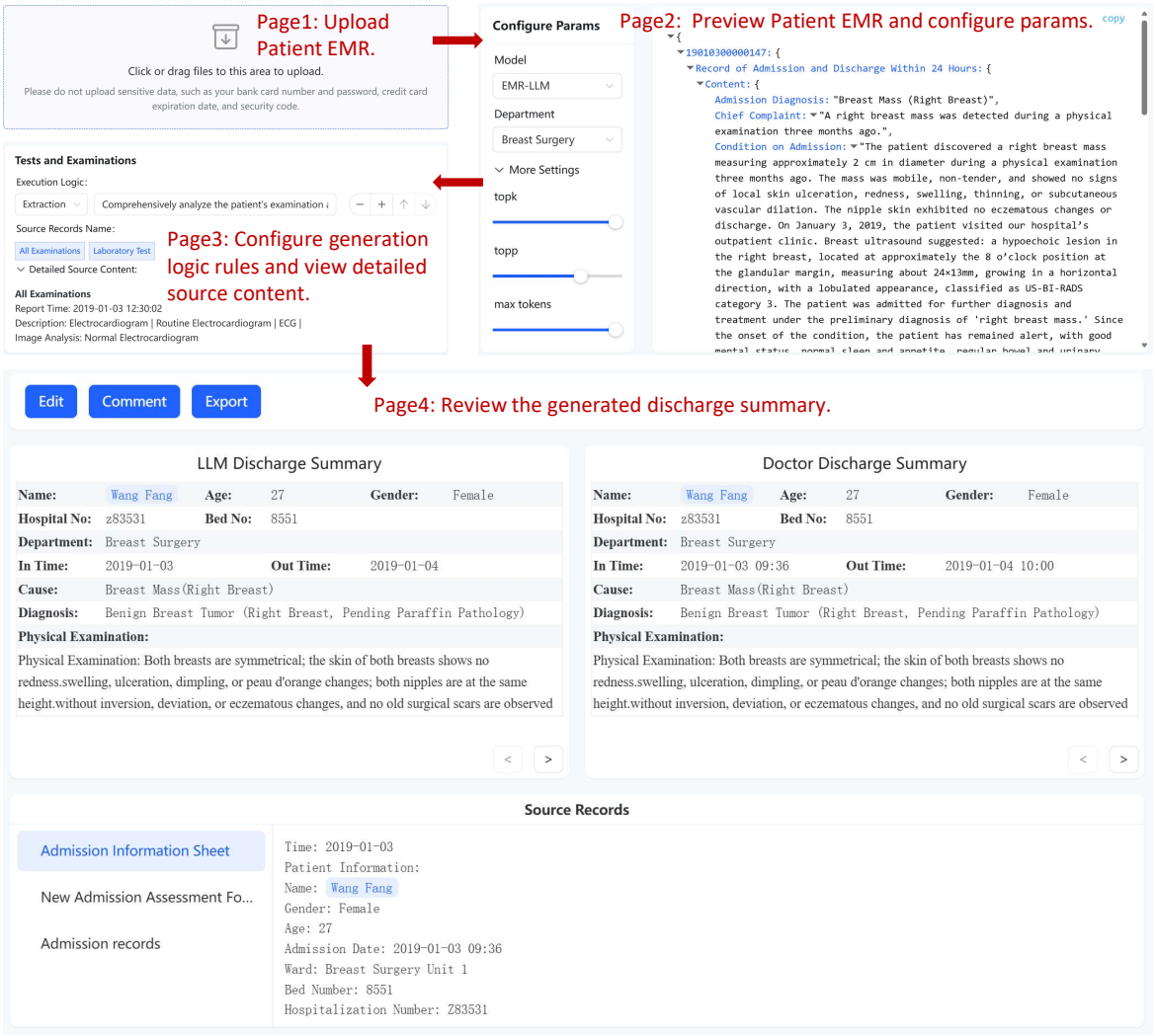

Figure 1: Screenshot of the LCDS web application, where the page functions are annotated.

tal training dataset for continuous model optimization once enough data is accumulated.

3.1 Input EMR Format Conversion

As shown in Figure 1, users begin on Page 1 by uploading multiple EMR documents via a drag-and-drop interface (see Appendix A for supported document types). LCDS preprocesses and converts these documents into a unified JSON format, facilitating consistency and accurate source attribution. The unified format simplifies downstream processing and improves processing efficiency. Upon successful conversion, users proceed to Page 2, where the right panel displays structured EMR data, summarizing all uploaded records, and the left panel offers configuration options for model selection and department-specific logical rules, allowing users to tailor generation parameters to clinical needs.

3.2 Reference-Guided Source-Aware Discharge Summary Generation

After configuration, users proceed to Page 3, where they can preview source document names, extracted key content and customize logical constraints. LCDS supports 15 medical departments, with baseline source references provided for each DS field. As shown in Page 3 of Figure 1, the *"Source Records Name"* section displays source documents for the breast surgery department's DS, while *"Detailed Source Content"* shows extracted medical content. Users can modify logical rules in the *"Execution Logic"* section, which supports extraction, reasoning, summarization, and judgment logic types. The fifth logic type, *knowledge*, generates follow-up medication recommendations based on predefined mappings of medical history and test results to department-specific guidelines.

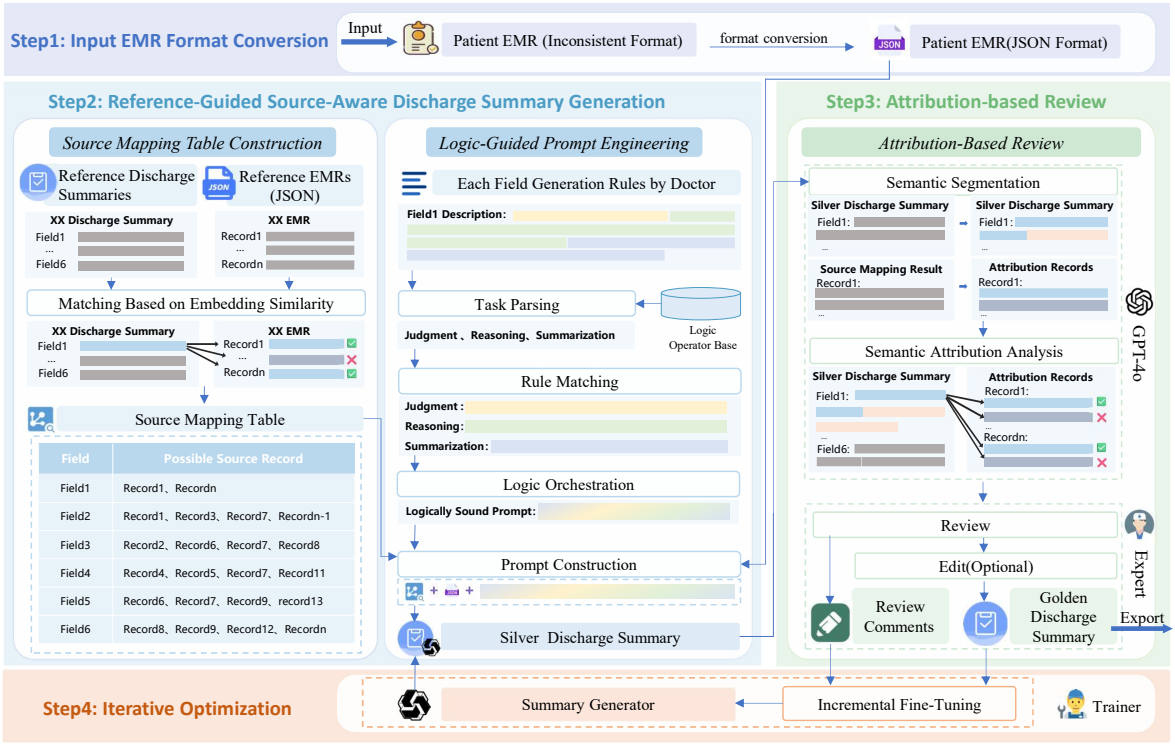

Figure 2: System workflow overview. The process includes four steps: (1) Upload and convert EMRs; (2) Extract key information, configure generation logic, and generate the discharge summary; (3) Perform attribution analysis and review; (4) Construct an incremental dataset and perform incremental learning.

3.3 Attribution-Based Comparison and Review

After configuring Page 3, LCDS generates the "Silver" DS and redirects users to the comparison interface on Page 4. The upper section displays the generated summary on the left, with physician-authored summaries for comparison. The lower section lists the source documents and their contents. Users can hover over the generated summary to highlight the matching content in the physician-written summary. Clicking on any part updates the lower section to show the corresponding source document and highlights related sentences. The top toolbar provides Edit, Comment, and Export functions for experts to modify content, annotate feedback, and download the final "Golden" DS in JSON format.

3.4 Iterative Optimization

Through the aforementioned steps, LCDS accumulates a dataset of "Silver" DSs and expert-reviewed "Golden" counterparts, which serves as an incremental training corpus for continuous model refinement. As data accumulates, trainers use these revised summaries for ongoing model improvement.

4 System Overview

4.1 Summary Generator

In our work, we utilize ChatGLM3-6B (GLM et al., 2024) to generate DSs. To enhance the model's understanding of task details and improve its performance in this text generation task, we construct a high-quality instruction dataset and fine-tune the model using LoRA.[1] The fine-tuned model is named EMRLLM. Since our backend model is modular, we can also replace EMRLLM with other LLMs such as Alpacare (Zhang et al., 2023c), Bentaso (Wang et al., 2023), or HuatuoGPT (Zhang et al., 2023a).

4.2 Source Mapping Table Construction

To enhance input precision, minimize hallucinations caused by excessive text scope, and improve the efficiency and accuracy of information localization, we construct a DS-EMR mapping table, which clearly defines the relationships between the DS and its corresponding source documents and relevant fields.

[1]We provide some examples of instruction dataset in https://github.com/ycycyc02/LCDS.

We collect 500 EMRs from 15 departments, each containing a physician-authored DS. These DSs serve as ground truth for localizing information from the corresponding source documents. To facilitate structured generation, we divide each DS into six distinct Fields: (1) Patient Information, (2) Discharge Diagnosis, (3) Tests and Examinations, (4) Disease Course and Treatment, (5) Condition at Discharge, and (6) Post-Discharge Medication Advice.

For short-text Fields such as "patient information", we directly use the ground truth as a keyword to search across all fields of the medical records. If a field contains the keyword, it is identified as the corresponding information source.

For long-text Fields such as "Disease Course and Treatment", content may originate from multiple medical records, and different sentences may correspond to different source documents. To address this, we first perform sentence-level semantic segmentation and then determine the source of each segment. Specifically, we employ in-context learning (ICL) for semantic segmentation, where the input consists of the "Disease Course and Treatment" text, and the output includes categorized labels and their corresponding content. For instance, if a patient's disease course involves surgery, chemotherapy, pathology, and discharge details, the output should be {Surgery: corresponding surgical description, Chemotherapy: corresponding chemotherapy description, Pathology: corresponding pathology description, Discharge Details: corresponding discharge description}. Using this approach, we break down long texts into finer-grained queries, which are then used to retrieve relevant information from all fields in the patient's EMRs.

We employ the BM25 (Robertson et al., 2009) algorithm to compute semantic similarity, ranking and filtering field contents within the same category based on similarity scores. Fields with similarity scores exceeding 0.8 are considered valid sources. For example, if chemotherapy information for patients A and B originates from Field P of Document X (with similarity scores of 0.9 and 0.85, respectively), and for patient C from Field O of Document Y (with a similarity score of 0.95), while also appearing in Field N of Document Y (with a similarity score of 0.75), only X-P and Y-O are retained as valid sources during selection. Here, X-P appears as a source in 2/3 of cases (covering patients A and B), and Y-O appears in 1/3 of cases (covering only patient C), assigning them priorities of 2/3 and 1/3, respectively. During new patient data processing, the system first extracts content from the highest-priority field. If the field is missing, it sequentially falls back to the next most relevant field.

Ultimately, this strategy leverages semantic segmentation, similarity-based retrieval, and relevance-based filtering to refine input text, ensuring that the model generates high-quality discharge summaries that better meet clinical needs within the constraints of limited scope.

4.3 Logic-Guided Prompt Engineering

To suppress hallucinations caused by free-text generation while accommodating the specific needs of different medical departments, we establish explicit generation rules and constraints for various DS content types. The generation logic is categorized into five types, with corresponding optimizations applied to each:

Extraction: Extracts deterministic information (e.g., name, hospitalization number) for data accuracy.

Summarization: Summarizes key information from multiple documents (e.g., medical history) or a concise overview.

Judgment: Evaluates input based on clinical standards (e.g., abnormal test results) and outputs compliant conclusions.

Inference: Integrates data points to infer disease progression or treatment outcomes (e.g., discharge time).

Knowledge: Uses clinical knowledge bases to generate advisory information (e.g., follow-up departments, precautions).

To implement logic-driven DS generation, we first collaborate with medical experts to define natural language generation rules for each DS field. We then employ GPT-4o (Hurst et al., 2024) with a three-stage intelligent processing mechanism for optimization:

Task Parsing: Automatically matches generation rules with 1-4 logical structures based on predefined logic types.

Rule Matching: Assigns detailed generation rules to each logical structure.

Logic Orchestration: Integrates and generates structured, coherent, and logically sound prompt composite instructions.

Through the three-stage optimization of task parsing, rule matching, and logic orchestration, the system generates field-specific logical combination templates that comply with medical standards and

Method	ROUGE-L	LLM-as-a-Judge	Human
GPT-4o with COT	24.01	24.68	31.41
GPT-4o with LCDS	40.24	54.81	52.57
EMRLLM with LCDS	**77.60**	**75.26**	**79.45**

Table 1: Performance comparison of different methods, including GPT-4o with COT, GPT-4o with LCDS, and EMRLLM with LCDS. The results are evaluated using ROUGE-L, LLM-as-a-Judge, and human evaluation. The best results in each column are highlighted in bold.

maintain a clear logical flow. This enables an automated transformation from business directives to precise prompts. Additionally, physicians can modify the results during the rule-matching stage to meet personalized requirements. For example, if a physician wishes to include intraocular pressure test results in the DS, they can adjust the rule matching output accordingly, further optimizing the final generated content.

4.4 Attribution-Based Comparison

In the medical domain, the generation of discharge summaries requires clear content attribution for auditing and verification. To this end, we propose an attribution-based review method that establishes explicit correspondence between generated content and original medical records, ensuring accuracy and reliability.

Specifically, we first perform sentence-level segmentation on both the generated DS and the associated original medical records. Then, we leverage the GPT-4o model to process each generated sentence and determine its supporting sentence(s) within the original medical records. To ensure precise attribution, each sentence in the original records is assigned a unique identifier, and GPT-4o is instructed to return only the corresponding identifiers of supporting sentences.

On the user interface, when a user clicks on a sentence in the generated DS, the system highlights the corresponding original medical record sentences with the same identifier, facilitating easy comparison and verification.

5 Evaluation

In this section, we validate the effectiveness of LCDS through a combination of automatic and human evaluation. The experimental results are presented in Table 1.

Dataset: We collect 150 EMRs, selecting 10 from each of 15 departments.

Baseline Methods: To evaluate the effectiveness of LCDS, we compare it with the following three baseline methods: 1) GPT-4o with COT (Chain of Thought (Wei et al., 2022)): Using GPT-4o for EMR-based text generation, incorporating the COT reasoning method to enhance logical consistency. 2) GPT-4o with LCDS: Using GPT-4o within the LCDS framework to optimize its performance and enhance its applicability in the medical domain. 3) EMRLLM with LCDS: Using EMRLLM within the LCDS framework to optimize DS generation and enhance output precision.

Evaluation Metrics: We employ both automatic and human evaluation metrics. **Automatic Evaluation**: ROUGE-L (Lin, 2004) measures the longest common subsequence overlap between the generated DS and the reference DS, providing an indication of lexical similarity. LLM-as-a-Judge (Gu et al., 2024) employs DeepSeek-R1 (Guo et al., 2025) to assess the generated text along four dimensions, including accuracy, completeness, standardization, and practicality, with a combined total score of 100 points. The evaluation criteria are detailed in Appendix B. **Human Evaluation**: Medical experts assign an overall score to the generated text based on the same four dimensions, with the total score ranging from 0 to 100. Detailed evaluation guidelines are provided in Appendix C.

Evaluation Results: The results demonstrate that GPT-4o with LCDS outperforms GPT-4o with COT across all metrics, indicating that the LCDS framework contributes to improved generation quality. Furthermore, EMRLLM with LCDS achieves superior performance compared to GPT-4o with LCDS, suggesting that task-specific fine-tuning on medical datasets significantly enhances generation quality.

6 Conclusion

We present **LCDS**, a logic-controlled discharge summary generation system that integrates precise content localization, logic-guided generation, and attribution-based expert review. By accurately extracting relevant source content, LCDS effectively reduces irrelevant information, thereby improving the quality and coherence of generated summaries. Through medical domain-specific logical constraints, LCDS significantly mitigates hallucinations and adapts to varied requirements across different clinical departments. Additionally, LCDS supports content traceability, enabling efficient expert validation, feedback, and iterative improve-

ment of large language models in clinical practice. Our experiments on real-world clinical data demonstrate that LCDS consistently outperforms existing methods, highlighting its potential for reliable and trustworthy clinical deployment.

Limitations

Despite the remarkable progress achieved in discharge summary generation, our study still has several limitations. First, our approach primarily relies on a specific dataset for training and evaluation, which may limit the model's generalization ability and result in degraded performance when applied to different healthcare settings or other types of electronic medical records. Second, due to the highly specialized and complex nature of medical texts, the model may generate inaccurate or ambiguous content, affecting its applicability in clinical practice. Finally, although we employ both automated and manual evaluation methods, a more comprehensive assessment of the generated text's quality and usability remains necessary. Future work could incorporate additional expert reviews or real-world clinical testing to further refine the evaluation process.

Ethics Statement

This study strictly adheres to ethical guidelines, ensuring that all data usage complies with relevant privacy protection and data security regulations. The datasets employed have been anonymized to prevent the exposure of sensitive patient information. Additionally, we acknowledge the potential risks associated with generative models in automated medical text generation, including the possibility of producing inaccurate or misleading content. Therefore, we emphasize that the model should be used solely as an assistive tool and that all generated outputs must be rigorously reviewed and validated by medical professionals.

Acknowledgments

We sincerely thank the anonymous reviewers for their valuable comments and suggestions. We also appreciate the support from Ruijin Hospital, Shanghai Jiaotong University School of Medicine, for this work.

References

Griffin Adams, Jason Zucker, and Noémie Elhadad. 2024. Speer: Sentence-level planning of long clinical summaries via embedded entity retrieval. *arXiv:2401.02369*.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.

Shuyang Cao and Lu Wang. 2024. Verifiable generation with subsentence-level fine-grained citations. *arXiv preprint arXiv:2406.06125*.

Yung-Sung Chuang, Benjamin Cohen-Wang, Shannon Zejiang Shen, Zhaofeng Wu, Hu Xu, Xi Victoria Lin, James Glass, Shang-Wen Li, and Wen-tau Yih. 2025. Selfcite: Self-supervised alignment for context attribution in large language models. *arXiv preprint arXiv:2502.09604*.

Reece Alexander James Clough, William Anthony Sparkes, Oliver Thomas Clough, Joshua Thomas Sykes, Alexander Thomas Steventon, and Kate King. 2024. Transforming healthcare documentation: harnessing the potential of ai to generate discharge summaries. *BJGP open*, 8(1):BJGPO.2023.0116.

Benjamin Cohen-Wang, Harshay Shah, Kristian Georgiev, and Aleksander Madry. 2024. Contextcite: Attributing model generation to context. *Advances in Neural Information Processing Systems*, 37:95764–95807.

Daniel Dubinski, Sae-Yeon Won, Svorad Trnovec, Bedjan Behmanesh, Peter Baumgarten, Nazife Dinc, Juergen Konczalla, Alvin Chan, Joshua D. Bernstock, Thomas M. Freiman, and Florian Gessler. 2024. Leveraging artificial intelligence in neurosurgery-unveiling chatgpt for neurosurgical discharge summaries and operative reports. *Acta Neurochirurgica*, 166(1):38.

Constanza Fierro, Reinald Kim Amplayo, Fantine Huot, Nicola De Cao, Joshua Maynez, Shashi Narayan, and Mirella Lapata. 2024. Learning to plan and generate text with citations. *arXiv preprint arXiv:2404.03381*.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. 2024. Chatglm: A family of large language

models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Vince C Hartman, Sanika S Bapat, Mark G Weiner, Babak B Navi, Evan T Sholle, and Thomas R Campion Jr. 2023. A method to automate the discharge summary hospital course for neurology patients. *Journal of the American Medical Informatics Association*, 30(12):1995–2003.

Lucas Torroba Hennigen, Shannon Shen, Aniruddha Nrusimha, Bernhard Gapp, David Sontag, and Yoon Kim. 2023. Towards verifiable text generation with symbolic references. *arXiv preprint arXiv:2311.09188*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Krishna K, Khosla S, Bigham J, and Lipton ZC. 2021. Generating soap notes from doctor-patient conversations using modular summarization techniques. *Association for Computational Linguistics*, pages 4958–4972.

Hanjae Kim, Hee Min Jin, Yoon Bin Jung, and Seng Chan You. 2024. Patient-friendly discharge summaries in korea based on chatgpt: Software development and validation. *Journal of Korean Medical Science*, 39(16):e148.

Sunil Kripalani, Amy T Jackson, Jeffrey L Schnipper, and Eric A Coleman. 2007. Promoting effective transitions of care at hospital discharge: a review of key issues for hospitalists. *Journal of hospital medicine: an official publication of the Society of Hospital Medicine*, 2(5):314–323.

Leslie A Lenert, Farrant H Sakaguchi, and Charlene R Weir. 2014. Rethinking the discharge summary: a focus on handoff communication. *Academic Medicine*, 89(3):393–398.

Jordan YZ Li, Tuck Y Yong, Paul Hakendorf, David Ben-Tovim, and Campbell H Thompson. 2013. Timeliness in discharge summary dissemination is associated with patients' clinical outcomes. *Journal of evaluation in clinical practice*, 19(1):76–79.

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

Fenglin Liu, Bang Yang, Chenyu You, Xian Wu, Shen Ge, Zhangdaihong Liu, Xu Sun, Yang Yang, and David Clifton. 2022. Retrieve, reason, and refine: Generating accurate and faithful patient instructions. *Advances in Neural Information Processing Systems*, 35:18864–18877.

Mengxian Lyu, Cheng Peng, Daniel Paredes, Ziyi Chen, Aokun Chen, Jiang Bian, and Yonghui Wu. 2024. Uf-hobi at" discharge me!": A hybrid solution for discharge summary generation through prompt-based tuning of gatortrongpt models. *arXiv preprint arXiv:2407.15359*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.

Lesley C Rink, Tolu O Oyesanya, Kathryn C Adair, Janice C Humphreys, Susan G Silva, and John Bryan Sexton. 2023. Stressors among healthcare workers: a summative content analysis. *Global qualitative nursing research*, 10:23333936231161127.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Han-Chin Shing, Chaitanya Shivade, Nima Pourdamghani, Feng Nan, Philip Resnik, Douglas Oard, and Parminder Bhatia. 2021. Towards clinical encounter summarization: Learning to compose discharge summaries from prior notes. *arXiv preprint arXiv:2104.13498*.

Aviv Slobodkin, Eran Hirsch, Arie Cattan, Tal Schuster, and Ido Dagan. 2024. Attribute first, then generate: Locally-attributable grounded text generation. *arXiv preprint arXiv:2403.17104*.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, et al. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature medicine*, 30(4):1134–1142.

Hartman VC, Bapat SS, Weiner MG, and et al. 2023. A method to automate the discharge summary hospital course for neurology patients. *Journal of the American Medical Informatics Association*, 12:12.

Carl Van Walraven, Ratika Seth, Peter C Austin, and Andreas Laupacis. 2002. Effect of discharge summary availability during post-discharge visits on hospital readmission. *Journal of general internal medicine*, 17:186–192.

Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. 2023. Huatuo: Tuning llama model with chinese medical knowledge. *Preprint*, arXiv:2304.06975.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Christopher Y. K. Williams, Jaskaran Bains, Tianyu Tang, Kishan Patel, Alexa N. Lucas, Fiona Chen, Brenda Y. Miao, Atul J. Butte, and Aaron E. Kornblith. 2024. Evaluating large language models for drafting emergency department discharge summaries. *medRxiv: The Preprint Server for Health Sciences*.

Haotian Wu, Paul Boulenger, Antonin Faure, Berta Céspedes, Farouk Boukil, Nastasia Morel, Zeming Chen, and Antoine Bosselut. 2024. Epfl-make at "discharge me!": An llm system for automatically generating discharge summaries of clinical electronic health record. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 696–711.

Ying Xiong, Buzhou Tang, Qingcai Chen, Xiaolong Wang, and Jun Yan. 2019. A study on automatic generation of chinese discharge summary. In *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1681–1687. IEEE.

Justin Xu, Zhihong Chen, Andrew Johnston, Louis Blankemeier, Maya Varma, Jason Hom, William J Collins, Ankit Modi, Robert Lloyd, Benjamin Hopkins, et al. 2024. Overview of the first shared task on clinical text generation: Rrg24 and" discharge me!". In *BioNLP@ ACL*.

Xi Ye, Ruoxi Sun, Sercan Ö Arik, and Tomas Pfister. 2023. Effective large language model adaptation for improved grounding and citation generation. *arXiv preprint arXiv:2311.09533*.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Jianquan Li, Guiming Chen, Xiangbo Wu, Zhiyi Zhang, Qingying Xiao, et al. 2023a. Huatuogpt, towards taming language model to be a doctor. *arXiv preprint arXiv:2305.15075*.

Jingyu Zhang, Marc Marone, Tianjian Li, Benjamin Van Durme, and Daniel Khashabi. 2024. Verifiable by design: Aligning language models to quote from pre-training data. *arXiv preprint arXiv:2404.03862*.

Nan Zhang, Yusen Zhang, Wu Guo, Prasenjit Mitra, and Rui Zhang. 2023b. Famesumm: investigating and improving faithfulness of medical summarization. *arXiv preprint arXiv:2311.02271*.

Xinlu Zhang, Chenxin Tian, Xianjun Yang, Lichang Chen, Zekun Li, and Linda Ruth Petzold. 2023c. Alpacare: Instruction-tuned large language models for medical application. *arXiv preprint arXiv:2310.14558*.

Yuhao Zhang, Derek Merck, Emily Tsai, Christopher D. Manning, and Curtis Langlotz. 2020. Optimizing the factual correctness of a summary: A study of summarizing radiology reports. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5108–5120. Online.

A Details on Document Types

Our system encompasses eight types of EMR documents, including medical records, nursing records, examinations, laboratory tests, medical orders, pathology reports, diagnoses, and vital sign records. The specific content of each document type is detailed in Table 2, with representative examples available in our public repository.

To ensure consistent data representation and enable effective cross-source integration, all documents are transformed into a standardized JSON format via predefined conversion scripts upon upload. This conversion framework is designed to be both highly generalizable and configurable: by implementing tailored scripts for specific data types, we achieve precise format mapping and data normalization. Consequently, our system exhibits strong adaptability, enabling flexible application to a wide range of EMR datasets.

B Evaluation Criteria for LLM-as-a-Judge

Below is the translated version of the evaluation prompt for LLM-as-a-Judge:

Your task is to evaluate the quality of AI-generated discharge summaries (compared to the physician-written reference version).

Scoring range: 0–100 points

Scoring dimensions:

1. Information Accuracy

- Correctness of patient identity information (e.g., name, bed number, admission number)

- Accuracy of key time points (e.g., admission/discharge times)

- Accuracy of brief medical history and physical examination summary at admission

- Consistency of diagnostic terms with the reference answer

2. Medical Completeness

- Must include core sections: brief admission history, physical exam summary, in-hospital medical course, disease progression and treatment, discharge diagnosis, medication recommendations after discharge, patient condition at discharge

- Coverage of key data: laboratory tests, imaging results, surgical details, follow-up suggestions, medication guidance, etc. (no errors allowed in numerical values and test items related to the in-hospital course)

3. Professional Standardization

- Standardization of medical terminology

- Clear logical structure (description of diagnosis and treatment process in chronological order)

- Avoid unnecessary redundancy (e.g., full-system physical examination descriptions)

4. Clinical Practicality

- Actionability of discharge instructions (e.g., specific dressing change times, pathology report follow-up points)

- Completeness of risk warnings (e.g., signs of incision infection)

Output format:

{
"score" [overall score],
"breakdown" {
"Information Accuracy" [score]/40,
"Medical Completeness" [score]/35,
"Professional Standardization" [score]/15,
"Clinical Practicality" [score]/10
}
}

C Evaluation Criteria for Human

To ensure reliable human evaluation of discharge summaries, we developed a scoring manual with a total of 100 points. The evaluation is based on four core dimensions: accuracy, completeness, standardization, and clinical utility, with an emphasis on patient safety and clinical relevance. Each dimension is scored on a scale from 0 to its maximum value; negative scores are not permitted, and any deductions resulting in a negative value will be recorded as zero.

C.1 Accuracy of Core Information (30 points)

- **Patient Identification**: Name, admission ID, and bed number must be correct.Each error results in a 3-point deduction.

- **Time Points**: Admission and discharge dates must be accurate (minute-level precision not required).Each error results in a 3-point deduction.

- **Diagnostic Consistency**: The discharge diagnosis must fully align with the final clinical conclusion. Descriptors like "pending paraffin section" must be included if applicable. Contradictions (e.g., benign vs. malignant misclassification) result in a 15-point deduction; omission of key diagnostic content incurs a 10-point deduction.

No.	Document Name	Content Included	Structure
1	Medical Records	Admission records, surgery records, ward round records, etc.	Unstructured data with HTML tags
2	Nursing Records	Discharge summary, etc.	XML data
3	Examination	Examination information	Structured data
4	Laboratory Test	Laboratory test information	Structured data
5	Medical Orders	Tests, prescriptions, textual reminders, etc.	Structured data
6	Pathology Report	Pathology examination information and reports	Structured data
7	Diagnosis	Diagnoses given by doctors during hospitalization	Structured data
8	Vital Signs Records	Vital signs measurements during hospitalization	Structured data

Table 2: Details on Document Types

- **Admission History and Physical Exam Summary**: Should be consistent with the initial clinical documentation. Each error results in a 3-point deduction.

C.2 Completeness of Medical Content (30 points)

- **Treatment Process Description**: Must include the procedure name, specific date, anesthesia type, and key surgical details (e.g., "right breast Mammotome excision under general anesthesia"). Missing any critical element results in an 8-point deduction.
- **Key Examinations During Hospitalization**: Laboratory (e.g., CBC, liver function, hepatitis panel) and imaging reports (e.g., ultrasound, chest X-ray) should be fully documented. Missing a category of essential results incurs a 5-point deduction.
- **Post-Discharge Instructions**: Should clearly specify pathology report follow-up timing (e.g., "10 working days"), wound care details (frequency, location, contraindications), medications, signs of complications (e.g., infection), and follow-up plans. Missing any important item leads to a 6-point deduction.
- **Discharge Condition**: Should be consistent with the physician's final record; a discrepancy will result in a 5-point deduction.

C.3 Professional Standardization (25 points)

- **Terminology**: Use standardized clinical terms (e.g., "US-BI-RADS category 3"). Each error or improper abbreviation results in a 3-point deduction.
- **Logical Structure**: Clinical descriptions should follow chronological order with coherent logic. Disordered descriptions result in an 8-point deduction.
- **Content Focus**: Irrelevant details (e.g., normal neurological exams in healthy patients) should be avoided. Redundant information results in a 5-point deduction per instance.

C.4 Clinical Utility (15 points)

- **Actionable Recommendations**: Instructions must be specific (e.g., "change dressing on day 3 after surgery" rather than "change dressing regularly"). Vague advice results in a 5-point deduction.
- **Risk Mitigation**: Key complications (e.g., redness, discharge, fever) and pathology report tracking must be addressed. Missing these incurs an 8-point deduction.
- **Individualized Follow-up**: Abnormal findings (e.g., hepatitis B positive) should include tailored follow-up suggestions. Up to ±2 points may be adjusted based on appropriateness.