CareTransition-Audit: A Benchmark to Audit Discharge Summaries for Efficient Care Transitions

Anonymous Authors[1]

Abstract

Incomplete or inconsistent discharge documentation drives care fragmentation and avoidable readmissions. Despite its critical role in patient safety, auditing discharge summaries relies on manual review and does not scale. We propose an automated framework for auditing discharge summaries using large language models (LLMs). Our approach operationalizes the DISCHARGED framework into a checklist of 46 questions. Using 50 summaries from the MIMIC-IV database, with clinician ground-truth labels, we benchmark 11 LLMs. Model-assessed mean documentation completeness ranges from 54.9% to 74.2%, and the best-performing model achieves a Cohen's κ of 0.496 against clinician labels, indicating moderate agreement. All models struggle to identify ambiguous documentation (`Unclear`), highlighting a key gap in current automated auditing. This work provides a clinician-validated benchmark and zero-shot baselines for systematic quality improvement in clinical documentation.

1. Introduction

High-quality discharge documentation is essential for patient safety and is a key determinant of readmission risk after hospitalization, as it facilitates a seamless transition from hospital to home through the effective transfer of necessary information (Sakaguchi & Lenert, 2015; Agency for Healthcare Research and Quality, 2017) on medications and follow-up, and preventable adverse events in the early post-discharge period. Studies of patients hospitalized with heart failure and other high-risk conditions have shown that discharge summaries containing key content elements such as medication changes, pending tests, and clear follow-up plans, are associated with lower odds of 30-day readmission (Al-Damluji et al., 2015), showing the importance of patient-centered documentation in safe transitions of care. In most clinical workflows, physicians write the formal discharge note while clinical nurses are responsible for educating the patient and family before discharge, using that same documentation as a reference. When the discharge note is vague or incomplete, it increases the risk of miscommunication and missed education points. Manual auditing of these notes is time and labor intensive making it difficult to perform at scale.

Our work addresses this gap by formulating 46 atomic audit questions from the DISCHARGED framework (Ng et al., 2025), validated by a clinical expert, and applying them to 50 MIMIC-IV (Johnson et al., 2023) discharge summaries restricted to surviving discharges. We benchmark eleven LLMs in a zero-shot setting and compare against clinician ground-truth labels. We make three contributions: (1) a clinically validated 46-question evaluation framework for discharge documentation completeness; (2) a preliminary benchmark dataset of 50 summaries with clinician-verified labels; and (3) zero-shot baselines across eleven LLMs establishing reference performance for automated auditing.

2. Related Works

NLP for Clinical Documentation. Automated processing of clinical text has a long history in biomedical informatics. Early work focused on extracting diagnoses, medications, and adverse events from free-text EHR data using rule-based and supervised machine learning approaches (Meystre et al., 2007; Habehh & Gohel, 2021), which typically require extensive feature engineering and task-specific annotations. The introduction of transformer-based models substantially advanced the field, ClinicalBERT (Alsentzer et al., 2019) adapted bidirectional encoders to clinical corpora, while domain-specific generative models (Singhal et al., 2022; Nazi & Peng, 2024; Christophe et al., 2024) have demonstrated strong performance on clinical information extraction and summarization tasks. However, the predominant focus of these models has been on clinical prediction or information retrieval rather than on evaluating the documentation quality or completeness.

Discharge Summary Generation and Evaluation. Most prior work focuses on automated generation of discharge

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

summaries (Rodrigues & Lopes, 2025; Hartman et al., 2023), typically through fine-tuning to produce summaries that conform to predefined templates (Ellershaw et al., 2024) or to regenerate specific sections such as the Brief Hospital Course by removing these sections from the discharge summary and using the rest as input (Liu et al., 2024b; Li et al., 2026). The *Discharge Me!* shared task (Xu et al., 2024) is an established benchmark for this, attracting contributions from multiple teams (Wu et al., 2024; Liu et al., 2024b). However, evaluation in this line of work has relied predominantly on surface-level metrics such as ROUGE and BERTScore, or on LLM-as-a-Judge protocols (Croxford et al., 2025a) and expensive clinician review. Efforts to improve factual reliability, such as PDSQI-9 (Croxford et al., 2025b) and hallucination detection methods (Asgari et al., 2025), address an important dimension of generation quality but remain insufficient for ensuring patient-centric completeness, as a summary can be fully factual yet still omit critical discharge elements.

Discharge Quality Frameworks and Auditing. Structured frameworks for discharge documentation quality such as the AHRQ's IDEAL Framework (Agency for Healthcare Research and Quality, 2017) and DISCHARGED mnemonic framework have been proposed to enumerate elements for safe discharge documentation. Despite the availability of such frameworks, their application has remained manual and small-scale, limited by the time and labor required for clinician-led chart review (Ellershaw et al., 2024). To our knowledge, no prior work has been done to audit patient-centric discharge documentation completeness for safe transitions.

3. Dataset and Study Design

This work presents a retrospective audit of hospital discharge summaries using a clinically validated set of audit questions. The objective is to assess the completeness and internal consistency of discharge documentation at scale, rather than to evaluate the appropriateness of clinical care delivered. Therefore, no new clinical content is generated and no clinical outcomes are modeled.

Data Source and Cohort. We use MIMIC-IV (Johnson et al., 2023), a publicly available, de-identified critical care database containing structured EHR data and clinical notes from Beth Israel Deaconess Medical Center. All adult inpatient admissions with an associated discharge summary were eligible for inclusion, admissions resulting in in-hospital mortality were excluded to focus on care transitions where discharge documentation directly informs downstream providers and patient education. From the eligible population, we sampled 50 discharge summaries from 50 unique patients using a stratified sampling strategy developed in consultation with clinical experts. Stratification was performed along two axes: (1) *discharge disposition*, to ensure representation across discharge locations, and (2) *ICU utilization*, a binary indicator of whether the admission included an intensive care unit stay, capturing complexity differences between ICU and non-ICU documentation. Patient ages ranged from 23 to 91 years ($\mu = 59.5$), with a gender distribution of 56% male and 44% female. The mean ICU length of stay was 2.04 days and the mean admission length of stay was 6.1 days. All summaries were annotated by a clinical expert against the full set of 46 audit questions, producing clinician-verified ground-truth labels.

3.1. Operationalizing DISCHARGED as an Audit Checklist

We operationalize the ten components of the DISCHARGED framework (Ng et al., 2025) into a structured audit checklist, as shown in Table 1. Each question is answered using one of four labels: `Yes` if the summary explicitly contains the requested information; `No` if no relevant information is present; `Unclear` if the information is partially present but insufficiently specific due to ambiguities in clinical writing or model uncertainty; and `N/A` if the question's precondition is not met (available only for specific conditional questions). Missing documentation is interpreted strictly as a documentation gap and does not imply that the corresponding clinical care was not delivered, a distinction critical for interpreting audit results without conflating documentation quality with care quality.

Table 1. Component-wise Audit Questions' Structure

Component	#Q	Audit Focus
(D)emographics	3	Identity, document placement
(I)mportant Alerts	3	Allergies, risks, precautions
(S)ocial Setup	2	Lifestyle, social context
(C)omp. History	4	Prior diagnoses, medications
(H)istory & Exams	8	Admission, vitals, exam
(A)ssessment & Course	8	Diagnoses, course, management
(R)ecorded Med Δ	6	Rationale, restart plans
(G)oals of Care	1	Advance directives
(E)xpected Follow-up	3	PCP, instructions, pending
(D)ischarge Info	8	Date, disposition, author
Total	**46**	

Prompting Strategy. Questions are divided into six prompts (<10 each), yielding six LLM calls per summary to avoid context degradation. Each prompt employs an indirect Chain-of-Thought (CoT) strategy (Wei et al., 2022), the model is instructed to answer each question with one of the designated labels, extract supporting evidence, and a brief justification linking the evidence to the label. Prompts explicitly instruct the model to recognize de-identification (e.g., masked patient identifiers) and to distinguish these from genuinely absent documentation, reducing false negatives attributable to de-identification

4. Results

We evaluate and compare eleven LLMs against clinician labels, on 50 MIMIC-IV discharge summaries using identical prompts. Models were selected to span a range of model families and parameter scales, Gemini-3-Flash-Preview (Google DeepMind, 2025), DeepSeek v3.2 (Liu et al., 2024a), Phi-4 (Abdin et al., 2024), Claude Sonnet-4.5 (Anthropic, 2025), GPT-5.4 (OpenAI, 2026), GPT-4o (Achiam et al., 2023), Grok-4.1-Fast (xAI, 2025), Llama 3.3-Nemotron-49B-v1.5 (NVIDIA, 2025), Llama 4 Maverick (Meta AI, 2025), and Nova-2-Lite-v1 (Amazon Web Services, 2025) were accessed via the OpenRouter API, Qwen 2.5-7B-Instruct (Qwen et al., 2025) was deployed locally using HuggingFace Transformers to demonstrate the feasibility of privacy-preserving on-premise auditing. No model-specific prompt tuning or few-shot examples are used.

Overall Agreement with Clinician Validated Labels. Table 2 reports agreement between each model and the clinician across all 46 questions. Claude Sonnet 4.5 achieves the highest Cohen's κ (0.496) and GPT-5.4 and Gemini 3 Flash follow closely, while the locally deployed Qwen 2.5-7B ($\kappa = 0.226$) and Phi-4 ($\kappa = 0.046$) substantially underperform. All models remain below the $\kappa = 0.6$ threshold typically considered "good" agreement, indicating that zero-shot auditing is feasible but far from solved.

Table 2. Overall agreement. Models ranked by κ. ρ = Spearman correlation of per-summary completeness scores.

Model	Acc.	κ	W-F1	ρ	p
Sonnet 4.5	.804	**.496**	.815	.334	.018
Gemini 3 Flash	.814	.483	.822	.480	<.001
DeepSeek V3	.743	.423	.775	.382	.006
GPT-5.4	.772	.420	.790	.406	.003
Nova 2 Lite	.747	.401	.779	.343	.015
Nemotron 49B	.719	.373	.749	.377	.007
Grok 4.1	.721	.371	.743	.234	.103
GPT-4o	.728	.370	.760	.333	.018
Llama 4 Maverick	.706	.340	.739	.298	.036
Qwen 2.5-7B†	.623	.226	.679	.189	.190
Phi-4	.640	.046	.655	.106	.465

†Locally deployed. W-F1 = weighted F1. ρ = Spearman completeness correlation.

Per-Label Analysis. A consistent pattern emerges across all models: `Yes` labels are predicted with high precision and recall (0.80–0.94 and 0.66–0.88 respectively), while `No` achieves moderate performance (0.33–0.60 precision, 0.52–0.78 recall). The most notable finding concerns the `Unclear` label, where all models achieve near-zero precision and recall (≤ 0.08 and ≤ 0.26). This disagreement is bidirectional: models frequently assign definitive `Yes` or `No` labels to questions that the clinician marked `Unclear`, suggesting overconfidence in resolving genuine clinical ambiguity. Conversely, models sometimes produce `Unclear` for the questions which the clinician answered definitively, indicating unnecessary hedging. Because the `Unclear` category captures precisely the cases where documentation is partial or ambiguous, the scenarios most likely to cause misinterpretation during care transitions, this inability of current models represents a key challenge for automated auditing. To understand the nature of these disagreements, our work captures free-text justifications from both the clinician and each model for every label assignment.

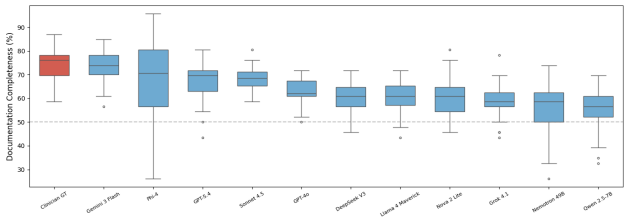

Figure 1. Per-summary completeness scores (proportion of `Yes`)

Ongoing analysis of these paired justifications will enable fine-grained characterization of *why* models and clinicians diverge on ambiguity, whether the disagreement stems from differing interpretations of clinical language, incomplete evidence extraction, or genuine boundary cases in documentation quality, informing targeted improvements to prompting strategies and providing supervision signal for future fine-tuned auditor models.

Documentation Completeness. Across 50 summaries, model-assessed mean completeness scores range from 54.9% (Qwen) to 74.2% (Gemini), confirming that substantial documentation gaps exist in MIMIC-IV discharge summaries regardless of which model performs the audit (Figure 1). Gemini achieves the strongest Spearman correlation with clinician completeness rankings ($\rho = 0.480$, $p < 0.001$), while Phi-4 and Qwen show no statistically significant correlation ($p > 0.1$), meaning their summary-level scores are unreliable proxies for clinical judgment.

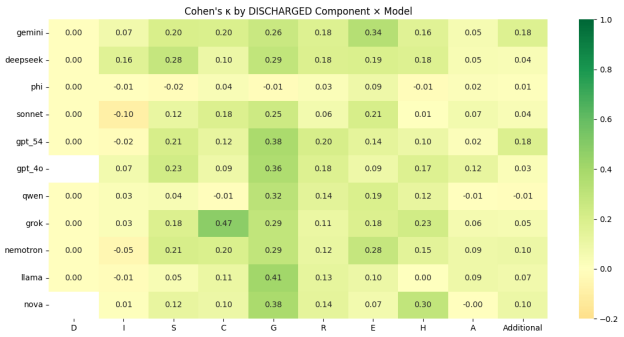

Figure 2. Cohen's κ by DISCHARGED component and model

Component-Level Agreement. Figure 2 shows Cohen's κ broken down by DISCHARGED component and model. Follow-up instructions (E) and medications (R) show the highest agreement, as these map to clearly templated, eas-

ily verifiable sections of MIMIC-IV summaries. Assessment (A) and important alerts (I) show the lowest agreement. Demographics (D) shows near-zero κ not due to disagreement but due to ceiling effects, both models and clinician label these questions `Yes` for nearly every summary, producing no variance for κ to measure.

5. Discussion

Our framework evolved iteratively with clinical input. We initially considered auditing against the AHRQ's IDEAL framework (Agency for Healthcare Research and Quality, 2017). However, we realized that the discharge summary documentation does not capture the entirety of this discharge planning process. Because many IDEAL components describe clinical *processes*, such as *Ask open-ended questions to elicit questions and concerns of the patient*, that are not documented in discharge summaries even when performed in practice, auditing them would require evaluating clinical workflows at various points. This mismatch between process oriented guidelines and a documentation oriented evaluation would have led to unreliable audit results. The DISCHARGED framework (Ng et al., 2025), designed specifically to guide summary *writing*, mapped more naturally to documentation content. The initial 34 questions were a direct mapping of framework components, during pilot annotation, compound questions that evaluated multiple elements introduced labeling ambiguity. We decomposed these into 46 atomic sub-questions, reducing annotator ambiguity and improving audit granularity. Format-specific questions were excluded to maintain flexibility across the diverse documentation styles observed in clinical practice.

Clinical experts highlighted that discharge summaries are authored incrementally by multiple contributors across care transitions (e.g., ICU to floor, resident rotations), with much content added near discharge. This multi-author process can produce excessive detail in some sections and omission of critical information in others, and the level of detail varies by service, obstetric discharges are substantially more standardized than complex ICU admissions. Experts emphasized that summaries should prioritize ongoing care needs over exhaustive inpatient chronicles. Notably, nurses actively use physician-written summaries for patient education, positioning automated auditing as a practical intervention at multiple workflow stages rather than merely retrospective measurement. However, even complete documentation does not guarantee effective transitions, as downstream providers may not read the summary due to time constraints or cross-institutional barriers.

Limitations. Our preliminary benchmark reflects documentation practices at a single institution (MIMIC-IV), generalizability to other healthcare systems remains to be validated. Clinician-labeled answers reflect a single expert's assessment and would benefit from multi-annotator agreement studies to quantify inter-rater reliability. The cohort includes all surviving discharges regardless of disposition, and documentation requirements may differ across disposition types such as home versus skilled nursing facility. Results are based on 50 summaries, which may not capture the full distribution of documentation patterns across clinical services. The use of de-identified data introduces artifacts such as masked names and dates that may affect model transferability to real-world settings. `Unclear` labels represent a composite of genuine clinical ambiguity in the documentation and the model's own uncertainty, making it difficult to disentangle documentation quality from model limitations without additional clinician justification.

6. Future Work

Several directions emerge from this work. First, we are working with clinical experts to expand the question-set and the cohort. We also plan to recruit additional annotators for multi-annotator agreement studies. Second, because discharge documentation varies substantially by clinical specialty and service, further research is needed to determine the stratification strategy for different clinical contexts, enabling finer-grained analysis of documentation quality. Third, we also plan to incorporate structured MIMIC-IV data (medications, laboratory results) as supplementary auditing context, enabling assessment of factual consistency by cross-referencing narratives against structured records, identifying cases where documented information contradicts or omits elements present in the underlying EHR data. Fourth, using the expanded clinician-labeled dataset, we aim to train a locally deployable supervised fine-tuned (SFT) auditor for real-time documentation evaluation while preserving patient privacy. This would enable scalable auditing, and serving as a real-time auditor during physician note completion or as an evaluator when an LLM generates a discharge summary.

Finally, we envision extending from documentation evaluation to documentation generation through a three-phase pipeline: (1) longitudinal temporal EHR representation, encoding the patient's full clinical trajectory using long-context reasoning; (2) generative LLM summarization, producing discharge summaries from the ground up rather than reformatting existing text; and (3) auditor-based optimization, using the SFT auditor as a reward signal within a Reinforcement Learning from AI Feedback (RLAIF) (Lee et al., 2023) framework. By rewarding summaries that satisfy the 46-point clinical checklist, this approach would produce documentation that is not only fluent and factually grounded but also demonstrably complete with respect to the informational elements required for safe care transitions.

Impact Statement

This work aims to improve patient safety by enabling scalable evaluation of discharge documentation quality, supporting nurses and clinicians in delivering efficient patient-care, education and transitions. However, we emphasize that this system is designed as a decision-support tool, not a replacement for clinical judgment. Automated audits should be reviewed by clinicians before acting on their outputs, as false negatives (missed gaps) could create false confidence in incomplete documentation, and false positives could increase alert fatigue.

References

Abdin, M., Aneja, J., Behl, H., Bubeck, S., Eldan, R., Gunasekar, S., Harrison, M., Hewett, R. J., Javaheripi, M., Kauffmann, P., et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Agency for Healthcare Research and Quality. IDEAL Discharge Planning Overview, Process, and Checklist, December 2017. URL https://www.ahrq.gov/patient-safety/patients-families/engagingfamilies/strategy4/index.html. AHRQ Publication No. 13-0051-EF.

Al-Damluji, M. S., Dzara, K., Hodshon, B., Punnanithinont, N., Krumholz, H. M., Chaudhry, S. I., and Horwitz, L. I. Association of discharge summary quality with readmission risk for patients hospitalized with heart failure exacerbation. *Circulation: Cardiovascular Quality and Outcomes*, 8(1):109–111, 2015. doi: 10.1161/CIRCOUTCOMES.114.001476. URL https://www.ahajournals.org/doi/abs/10.1161/CIRCOUTCOMES.114.001476.

Alsentzer, E., Murphy, J., Boag, W., Weng, W.-H., Jindi, D., Naumann, T., and McDermott, M. Publicly available clinical bert embeddings. In *Proceedings of the 2nd clinical natural language processing workshop*, pp. 72–78, 2019.

Amazon Web Services. Amazon Nova foundation models. https://aws.amazon.com/ai/generative-ai/nova/, 2025.

Anthropic. Claude sonnet 4.5. https://www.anthropic.com/news/claude-sonnet-4-5, 2025.

Asgari, E., Montana Brown, N., Dubois, M., Khalil, S., Balloch, J., Yeung, J., and Pimenta, D. A framework to assess clinical safety and hallucination rates of llms for medical text summarisation. *npj Digital Medicine*, 8, 05 2025. doi: 10.1038/s41746-025-01670-7.

Christophe, C., Kanithi, P. K., Raha, T., Khan, S., and Pimentel, M. A. Med42-v2: A suite of clinical llms. *arXiv preprint arXiv:2408.06142*, 2024.

Croxford, E., Gao, Y., First, E., Pellegrino, N., Schnier, M., Caskey, J., Oguss, M., Wills, G., Chen, G., Dligach, D., Churpek, M., Mayampurath, A., Liao, F., Goswami, C., Wong, K., Patterson, B., and Afshar, M. Evaluating clinical ai summaries with large language models as judges. *npj Digital Medicine*, 8, 11 2025a. doi: 10.1038/s41746-025-02005-2.

Croxford, E., Gao, Y., Pellegrino, N., Wong, K., Wills, G., First, E., Schnier, M., Burton, K., Ebby, C., Gorski, J., et al. Development and validation of the provider documentation summarization quality instrument for large language models. *Journal of the American Medical Informatics Association*, 32(6):1050–1060, 2025b.

Ellershaw, S., Tomlinson, C., Burton, O., Frost, T., Hanrahan, J., Khan, D. Z., Layard Horsfall, H., Little, M., Malgapo, E., Starup-Hansen, J., Ross, J., Woodward, G., Vella-Baldacchino, M., Noor, K., Shah, A., and Dobson, R. Automated generation of hospital discharge summaries using clinical guidelines and large language models. 02 2024.

Google DeepMind. Gemini 3 flash: Frontier intelligence built for speed. https://blog.google/products-and-platforms/products/gemini/gemini-3-flash/, 2025.

Habehh, H. and Gohel, S. Machine learning in healthcare. *Current Genomics*, 22(4):291–300, 2021. doi: 10.2174/1389202922666210705124359. URL https://doi.org/10.2174/1389202922666210705124359.

Hartman, V., Bapat, S., Weiner, M., Navi, B., Sholle, E., and Campion, J. A method to automate the discharge summary hospital course for neurology patients, 05 2023.

Johnson, A., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T., Hao, S., Moody, B., Gow, B., Lehman, L.-w., Celi, L., and Mark, R. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10:1, 01 2023. doi: 10.1038/s41597-022-01899-x.

Lee, H., Phatale, S., Mansoor, H., Mesnard, T., Ferret, J., Lu, K., Bishop, C., Hall, E., Carbune, V., Rastogi, A., et al. Rlaif vs. rlhf: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*, 2023.

Li, W., Feng, H., Hu, C., Xu, M., and Cheng, L. Accurate discharge summary generation using fine tuned large language models with self evaluation. *Scientific Reports*, 16, 01 2026. doi: 10.1038/s41598-026-35552-z.

Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024a.

Liu, J., Nicolson, A., Dowling, J., Koopman, B., and Nguyen, A. e-health CSIRO at "discharge me!" 2024: Generating discharge summary sections with fine-tuned language models. In Demner-Fushman, D., Ananiadou, S., Miwa, M., Roberts, K., and Tsujii, J. (eds.), *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pp. 675–684, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.bionlp-1.59. URL https://aclanthology.org/2024.bionlp-1.59/.

Meta AI. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation. https://ai.meta.com/blog/llama-4-multimodal-intelligence/, 2025.

Meystre, S., Savova, G., Kipper-Schuler, K., and Hurdle, J. Extracting information from textual documents in the electronic health record: A review of recent research. *Yearb Med Inform*, pp. 128–144, 11 2007. doi: 10.1055/s-0038-1638592.

Nazi, Z. A. and Peng, W. Large language models in healthcare and medical domain: A review. In *Informatics*, volume 11, pp. 57. MDPI, 2024.

Ng, I., Tung, D., Seet, T., Yow, K., Chan, K., Teo, D., and Chua, C. E. How to write a good discharge summary: a primer for junior physicians. *Postgraduate medical journal*, 101, 02 2025. doi: 10.1093/postmj/qgaf020.

NVIDIA. Llama-3.3-nemotron-super-49b-v1.5. https://build.nvidia.com/nvidia/llama-3_3-nemotron-super-49b-v1_5, 2025.

OpenAI. Introducing GPT-5.4. https://openai.com/index/introducing-gpt-5-4/, 2026.

Qwen, :, Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Tang, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., and Qiu, Z. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Rodrigues, T. and Lopes, C. Large language model-based generation of discharge summaries, 12 2025.

Sakaguchi, F. and Lenert, L. Improving continuity of care via the discharge summary. *AMIA Annual Symposium Proceedings*, 2015:1111–1120, 11 2015.

Singhal, K., Azizi, S., Tu, T., Mahdavi, S. S., Wei, J., Chung, H. W., Scales, N., Tanwani, A., Cole-Lewis, H., Pfohl, S., et al. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*, 2022.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D., et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Wu, H., Boulenger, P., Faure, A., Céspedes, B., Boukil, F., Morel, N., Chen, Z., and Bosselut, A. EPFL-MAKE at "discharge me!": An LLM system for automatically generating discharge summaries of clinical electronic health record. In Demner-Fushman, D., Ananiadou, S., Miwa, M., Roberts, K., and Tsujii, J. (eds.), *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pp. 696–711, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.bionlp-1.61. URL https://aclanthology.org/2024.bionlp-1.61/.

xAI. Grok. https://x.ai/news/grok-4, 2025.

Xu, J., Chen, Z., Johnston, A., Blankemeier, L., Varma, M., Hom, J., Collins, W. J., Modi, A., Lloyd, R., Hopkins, B., Langlotz, C., and Delbrouck, J.-B. Overview of the first shared task on clinical text generation: RRG24 and "discharge me!". In Demner-Fushman, D., Ananiadou, S., Miwa, M., Roberts, K., and Tsujii, J. (eds.), *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pp. 85–98, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.bionlp-1.7. URL https://aclanthology.org/2024.bionlp-1.7/.