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

Akshat Dasula [1 2] **Prasanna Desikan** [1] **Jaideep Srivastava** [2] **Shivali Dalmia** [1] **Abhishek Mukherji** [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 models achieve a Cohen's κ values around 0.5 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.

[1]Centific AI Research, Redmond, USA [2]Department of Computer Science & Engineering, University of Minnesota-Twin Cities, Minneapolis, USA. Correspondence to: Akshat Dasula <dasul001@umn.edu>.

Proceedings of the Workshop on Structured Data for Health at the 43rd International Conference on Machine Learning, Seoul, South Korea. Copyright 2026 by the author(s).

Discharge Summary Generation and Evaluation. Most prior work focuses on automated generation of discharge 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 and Appendix A. 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, see Appendix B) yielding six LLM calls per summary to avoid context degradation. Each prompt employs an indirect Chain-of-Thought (CoT) strategy (Wei et al., 2022) (asks for justification but doesn't prompt to think step-by-step), 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, with 95% bootstrap confidence intervals (1,000 resamples) for accuracy, κ, weighted F1, and Spearman ρ. Claude Sonnet 4.5, Gemini 3 Flash, DeepSeek V3, and GPT-5.4 form a top cluster with overlapping κ confidence intervals (range [.380, .533]), making them statistically indistinguishable on overall agreement despite point-estimate differences. The locally deployed Qwen 2.5-7B ($\kappa = 0.226$) substantially underperforms, and Phi-4's κ CI of [.003, .089] approaches zero, indicating its outputs are effectively uncorrelated with clinician labels. All models remain below the $\kappa = 0.6$ threshold typically considered "good" agreement, indicating that zero-shot auditing is feasible but far from solved.

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.

The clinician used `Unclear` sparingly across the full set of labels (38 of 2,300, 1.7%), with the label concentrated in questions where partial documentation is clinically common, pre-hospitalization functional status alone accounts for 11 of the 38 `Unclear` labels, followed by medication restart plans (4), and social and surgical history (3 each). A finer-grained look at these disagreements (Appendix C) is informative on the 38 pairs where the clinician selected `Unclear` , the two top-tier models force the choice in opposite directions, Gemini 3 Flash labels 20 as `Yes` and 17 as `No`, while Sonnet 4.5 labels 20 as `No` and 15 as `Yes`. That two strong models systematically disagree on exactly the cases the clinician flagged as ambiguous indicates documentation ambiguity in these instances, strengthening the interpretation of `Unclear` as a real documentation signal. 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.

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. The clinician-rated mean completeness across the 50 summaries is **79.0%**, this is the primary evidence of documentation gaps in MIMIC-IV summaries. Model-assessed completeness scores are reported as exploratory validation of whether LLMs can recover this signal. Model-assessed means range from 54.9% (Qwen 2.5-7B) to 74.2% (Gemini 3 Flash), with all eleven models systematically *under*-estimating completeness relative to the clinician, the closest model (Gemini) is 4.8 percentage points below, and the locally deployed Qwen underestimates by more than 20 percentage points. This under-estimation suggests that current zero-shot LLM auditors are more likely to flag false positive documentation gaps than to miss real ones, an error mode that may increase clinician alert fatigue but is unlikely to create false confidence in incomplete documentation. These model-assessed completeness scores should be checked alongside component-level agreement, as a model can produce a plausible aggregate completeness score while disagreeing with the clinician on which specific elements are present. Among models, Gemini 3 Flash achieves the strongest Spearman correlation with clinician completeness rankings ($\rho = 0.537$, 95% CI $[.321, .706]$); Grok 4.1, Qwen 2.5-7B, and Phi-4 show ρ confidence intervals that include zero, indicating their per-summary completeness scores are not reliable proxies for clinician judgment.

Component-Level Agreement. Figure 2 shows Cohen's κ broken down by DISCHARGED component and model. Four distinct patterns emerge. First, Demographics (D)

Table 2. Overall agreement between each LLM and the clinician on $n = 50$ MIMIC-IV discharge summaries. Values are point estimates with 95% bootstrap confidence intervals (1,000 resamples). Models ranked by κ. n = number of evaluated (summary, question) pairs out of $50 \times 46 = 2{,}300$.

Model	Accuracy	κ	W-F1	ρ	n
Sonnet 4.5	.804 [.788, .821]	**.496** [.455, .533]	.815 [.800, .831]	.432 [.178, .641]	2168
Gemini 3 Flash	.814 [.797, .830]	.483 [.441, .520]	.822 [.806, .838]	.537 [.321, .706]	2176
DeepSeek V3	.743 [.724, .761]	.423 [.388, .457]	.775 [.759, .791]	.413 [.146, .624]	2176
GPT-5.4	.772 [.753, .790]	.420 [.380, .459]	.790 [.774, .808]	.365 [.082, .585]	2114
Nova 2 Lite	.747 [.728, .765]	.401 [.365, .439]	.779 [.762, .795]	.359 [.102, .578]	2049
Nemotron 49B	.719 [.700, .739]	.373 [.337, .410]	.749 [.732, .765]	.474 [.241, .656]	2016
Grok 4.1	.721 [.702, .740]	.371 [.333, .406]	.743 [.727, .761]	.263 [−.028, .517]	2170
GPT-4o	.728 [.708, .745]	.370 [.333, .405]	.760 [.742, .776]	.341 [.064, .576]	2176
Llama 4 Maverick	.706 [.687, .727]	.340 [.304, .378]	.739 [.722, .757]	.335 [.042, .581]	2168
Qwen 2.5-7B†	.623 [.603, .644]	.226 [.194, .258]	.679 [.661, .698]	.254 [−.011, .482]	2176
Phi-4	.640 [.619, .662]	.046 [.003, .089]	.655 [.633, .677]	.100 [−.196, .388]	2035

†Locally deployed. Variation in n reflects N/A labels and inference parse failures.

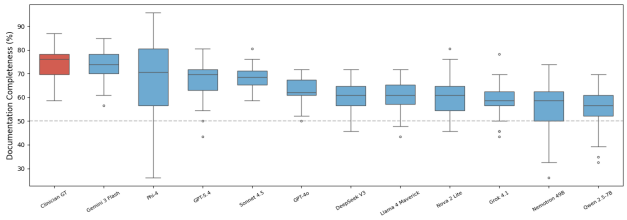

Figure 1. Per-summary completeness scores (proportion of `Yes`)

shows uniformly high κ values (.84–1.00) with very tight CIs for most models, reflecting ceiling-effect agreement where models and the clinician label nearly every question `Yes`. Second, History & Exams (H) shows the inverse pattern, κ near zero with CIs that span zero across all eleven models (e.g., Sonnet $[-.06, .13]$, Gemini $[-.04, .11]$, GPT-5.4 $[-.04, .10]$), indicating that models systematically fail to detect documentation gaps in admission narratives. Third, Important Alerts (I) is a less consistent failure: most models hover near zero with wide CIs (e.g., Sonnet $[.05, .34]$, GPT-5.4 $[-.26, .02]$), reflecting genuine difficulty with this component. Fourth, the highest robust agreement, where CIs are entirely positive and bounded away from zero, appears in Assessment (A) for the top-tier models (DeepSeek $[.54, .74]$, Sonnet $[.39, .59]$, Gemini $[.44, .69]$) and Additional / Discharge Info (Sonnet $[.51, .68]$, Gemini $[.48, .66]$).

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. **Completeness is necessary but not sufficient for documentation quality.** The 46-item checklist measures whether the informational elements required for safe care transitions are present in a discharge summary, not whether they are well-organized, internally consistent, or clinically correct. A summary can satisfy every checklist item and still be poorly structured, contain contradictions

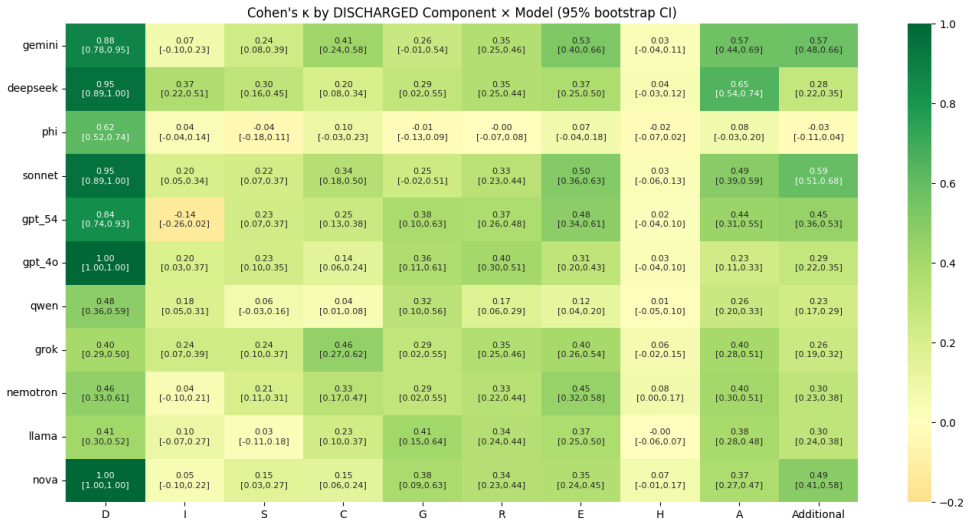

Figure 2. Cohen's κ by DISCHARGED component and model

between sections, bury critical information in narrative paragraphs, or misrepresent the clinical course. Conversely, a summary missing one or two checklist items may be highly readable and clinically sound. Our framework is therefore a screen for documentation gaps, completeness auditing is most useful as a first-pass filter that flags missing elements for clinician review, leaving questions of accuracy and clinical coherence to subsequent evaluation, whether by clinicians or by complementary frameworks such as PDSQI-9 (Croxford et al., 2025b), or by hallucination detection methods (Asgari et al., 2025).

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. This will allow direct measurement of whether the `Unclear` category is reproducible across annotators. 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 care transitions. We emphasize that the 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.

All experiments use MIMIC-IV under the PhysioNet Credentialed Health Data License, completing the required CITI training prior to data access. API-based model evaluations were performed through OpenRouter using its zero-data-retention configuration, restricted to providers contractually committed to no logging, no human review, and no use of inputs or outputs for model training. Qwen 2.5-7B was deployed locally, demonstrating the feasibility of a fully on-premise auditing pipeline for institutions where third-party API routing of clinical text is not permitted. No discharge summary text or patient-level identifiers appear in this paper or its supplementary materials.

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/.

A. Question Set

The 46 audit questions are partitioned into six prompts rather than issued as a single 46-question prompt or as 46 separate calls, balancing two competing constraints. Issuing all 46 questions in a single call risks context degradation, while issuing each question independently incurs prohibitive latency. We therefore group questions from DISCHARGED components into the same prompt, keeping each prompt under ten questions to remain within reliable context limits, and issue six LLM calls per discharge summary. The resulting partition is fixed across all models and summaries; see Table 3.

Table 3. Full 46-question DISCHARGED audit checklist, grouped by prompt. `N/A` is admissible only for conditional questions.

#	ID	Question	N/A	Condition
Prompt 1: Demographics + Important Alerts				
1	D1	Are basic patient demographics (age or date of birth, and sex) documented in the discharge summary?	No	—
2	D2	Is a patient identifier (e.g. name, medical record number, or patient identification number) documented, even if de-identified?	No	—
3	D3	Is patient contact information (e.g. address or phone number) documented, even if de-identified or blank?	No	—
4	I1	Is the patient's allergy status documented (either specific allergies listed, or an explicit statement such as NKDA/NDA/no known allergies)?	No	—
5	I2	If specific allergies are listed, are the allergens and their reaction types (e.g. rash, anaphylaxis) documented?	Yes	Patient has no allergies
6	I3	Are any other clinical alerts documented, such as adverse drug reactions, special risks, or precautions?	No	—
Prompt 2: Social Setup + Comprehensive History + Goals of Care				
7	S1	Does the discharge summary document any social history (e.g. smoking status, alcohol use, substance use, occupation, or living situation)?	No	—
8	S2	Does the discharge summary describe the patient's pre-hospitalization functional status (e.g. independence, mobility, baseline exercise tolerance)?	No	—
9	C1	Does the discharge summary state the patient's past medical history (e.g. previous diagnoses or chronic conditions)?	No	—
10	C2	Does the discharge summary state the patient's past surgical history?	Yes	Explicit "no prior surgeries"
11	C3	Is a pre-admission medication list documented?	No	—
12	C4	If a pre-admission medication list is documented, does it include doses and frequencies (not just drug names)?	Yes	C3 = No
13	G1	Is there any documentation of goals of care, advance directives, code status, or advance care planning?	No	—
Prompt 3: Recorded Medication Changes + Expected Follow-up				
14	R1	Is a discharge medication list documented?	No	—
15	R2	If a discharge medication list is documented, does it include the purpose or indication for each medication?	Yes	R1 = No
16	R3	If a discharge medication list is documented, does it include dose, route, and/or frequency information?	Yes	R1 = No

continued on next page

Table 3 continued from previous page

#	ID	Question	N/A	Condition
17	R4	Are any medication changes (new medications started, medications stopped, or dose adjustments) clearly documented?	No	—
18	R5	For documented medication changes, is the specific clinical rationale for each change provided?	Yes	R4 = No
19	R6	For medications stopped during the stay, is there a clear plan for whether or when they should be restarted?	Yes	No medications stopped
20	E1	Are follow-up instructions or appointments included in the discharge summary?	No	—
21	E2	Are there clear instructions regarding which outstanding investigations or pending results need to be reviewed or traced in the outpatient setting?	No	—
22	E3	Is the contact information for the Primary Care Provider (PCP) listed in the summary, even if de-identified or blank?	No	—
Prompt 4: History & Examinations				
23	H1	Does the discharge summary document the reason for the patient's admission?	No	—
24	H2	Does the discharge summary mention the admission date?	No	—
25	H3	Does the discharge summary document the source of referral or mode of admission (e.g. self-referral, ED, transfer from another facility)?	No	—
26	H4	Does the discharge summary document vital signs or clinical parameters on presentation?	No	—
27	H5	Does the discharge summary document targeted physical examination findings on presentation?	No	—
28	H6	Is the presenting symptom characterized with any detail (e.g. nature, onset, duration, progression, alleviating/exacerbating factors)?	No	—
29	H7	Are associated symptoms or significant negatives (especially to rule out red-flag symptoms) documented?	No	—
30	H8	Is relevant surgical history, drug history, or family history documented where pertinent to the presenting complaint?	No	—
Prompt 5: Assessment & Clinical Course				
31	A1	Are medical diagnoses given in the summary (actual medical diagnosis, not just symptoms)?	No	—
32	A2	Is the severity or complication level of the main diagnoses clearly described (e.g. KDIGO stage for AKI)?	No	—
33	A3	Where appropriate, does the summary include a brief one-sentence problem representation explaining the key features that support the diagnosis?	No	—
34	A4	Are clinical investigations listed (i.e. blood tests, lab tests, imaging, diagnostic procedures)?	No	—
35	A5	Is there a concise description of the patient's hospital course or clinical trajectory during admission?	No	—

continued on next page

Table 3 continued from previous page

#	ID	Question	N/A	Condition
36	A6	Does the summary describe the management plan for each main problem, including conservative measures, pharmacologic treatments, and any procedures or surgeries?	No	—
37	A7	Is the response to treatment documented for each major problem (e.g. resolution of symptoms, improvement in oxygen requirement, trending of creatinine)?	No	—
38	A8	If recommended investigations or treatments were withheld or stopped, is the reason documented (e.g. patient preference, goals of care, futility, risk > benefit)?	Yes	None withheld or stopped
Prompt 6: Discharge Information (Additional)				
39	Add1	Is the date of discharge documented?	No	—
40	Add2	Is the specialty of the doctor that discharged the patient included in the summary?	No	—
41	Add3	Is the discharge disposition documented (e.g. home, rehab, skilled nursing facility, step-down care)?	No	—
42	Add4	Is the type of discharge documented (e.g. normal, against medical advice, abscondment)?	No	—
43	Add5	Is the condition of the patient at discharge described (e.g. stable, improved, critical)?	No	—
44	Add6	Is hospital contact information listed for patient perusal, even if de-identified or blank?	No	—
45	Add7	Is information about the discharge summary writer included, even if de-identified?	No	—
46	Add8	Is the attending physician or discharging provider identified in the summary, even if de-identified?	No	—

B. Prompts

All six prompts share a common structure: an auditor persona statement scoped to the prompt's components, a note about de-identified content, a rule set governing the four allowed labels (`Yes`, `No`, `Unclear`, `N/A`), the audit questions for that prompt, and a strict JSON output schema requiring an answer, verbatim evidence from the discharge summary, and a brief justification for each question. The discharge summary text is appended at the end of each prompt. All prompts are issued zero-shot, with no system message and no model-specific tuning. The placeholder `{discharge_summary}` is replaced at runtime with the verbatim discharge summary text. The full audit checklist with conditional logic is given in Table 3.

B.1. Prompt 1: Demographics + Important Alerts

```
You are a clinical documentation auditor who works on demographic information and patient alerts.

You will be given a discharge summary. Your task is to answer the following audit questions based ONLY on the
     information present in the discharge summary.

Note:
- You are working with a de-identified dataset, information maybe explicitly stated but the details of it maybe
     blank (e.g. contact information)
- Give justification clearly when dealing with information which has blanks or dashes

Rules:
- Do NOT infer or assume information.
- Answers must be strictly one of: "Yes", "No", "Unclear", or "N/A".
- Use "Unclear" ONLY if partial or ambiguous information is present.
- If the information is completely absent, answer "No".
```

```
- Use "N/A" ONLY when the question's precondition does not apply (e.g., a conditional question whose triggering
     condition is not met).
- Evidence must be a direct quote or exact phrase(s) from the discharge summary.
- Justification must briefly explain why the evidence supports the selected answer.
- Do NOT add any content outside the specified JSON structure.


Audit Questions for Demographic Information:
1) Are basic patient demographics (age or date of birth, and sex) documented in the discharge summary?

2) Is a patient identifier (e.g. name, medical record number, or patient identification number) documented, even
     if de-identified?

3) Is patient contact information (e.g. address or phone number) documented, even if de-identified or blank?

Audit Questions for Important Alerts:
1) Is the patient's allergy status documented (either specific allergies listed, or an explicit statement such as
     NKDA/NDA/no known allergies)?

2) If specific allergies are listed, are the allergens and their reaction types (e.g. rash, anaphylaxis)
     documented? Answer "N/A" if the patient is documented as having no
allergies.

3) Are any other clinical alerts documented, such as adverse drug reactions, special risks, or precautions?

--------------------

Output Format (STRICT - valid JSON only):

{
  "D": {
    "1": {
      "answer": "Yes/No/Unclear",
      "evidence": "Exact quoted text or Not documented",
      "justification": "Brief explanation linking the evidence to the answer"
    },
    "2": {
      "answer": "Yes/No/Unclear",
      "evidence": "Exact quoted text or Not documented",
      "justification": "Brief explanation linking the evidence to the answer"
    },
    "3": {
      "answer": "Yes/No/Unclear",
      "evidence": "Exact quoted text or Not documented",
      "justification": "Brief explanation linking the evidence to the answer"
    }
  },
  "I": {
    "1": {
      "answer": "Yes/No/Unclear",
      "evidence": "Exact quoted text or Not documented",
      "justification": "Brief explanation linking the evidence to the answer"
    },
    "2": {
      "answer": "Yes/No/Unclear/N/A",
      "evidence": "Exact quoted text or Not documented",
      "justification": "Brief explanation linking the evidence to the answer"
    },
    "3": {
      "answer": "Yes/No/Unclear",
      "evidence": "Exact quoted text or Not documented",
      "justification": "Brief explanation linking the evidence to the answer"
    }
  }
}

-------------------------
Discharge Summary:
{discharge_summary}
```

B.2. Prompt 2: Social Setup + Comprehensive History + Goals of Care

```
You are a clinical documentation auditor who works on social history, past medical history, and goals-of-care
     documentation.

You will be given a discharge summary. Your task is to answer the following audit questions based ONLY on the
     information present in the discharge summary.

Note:
```

```
- You are working with a de-identified dataset, information maybe explicitly stated but the details of it maybe
     blank (e.g. contact information)
- Give justification clearly when dealing with information which has blanks or dashes

Rules:
- Do NOT infer or assume information.
- Answers must be strictly one of: "Yes", "No", "Unclear", or "N/A".
- Use "Unclear" ONLY if partial or ambiguous information is present.
- If the information is completely absent, answer "No".
- Use "N/A" ONLY when the question's precondition does not apply (e.g., a conditional question whose triggering
     condition is not met).
- Evidence must be a direct quote or exact phrase(s) from the discharge summary.
- Justification must briefly explain why the evidence supports the selected answer.
- Do NOT add any content outside the specified JSON structure.


Audit Questions for Social Set up:
1) Does the discharge summary document any social history (e.g. smoking status, alcohol use, substance use,
     occupation, or living situation)?

2) Does the discharge summary describe the patient's pre-hospitalization functional status (e.g. whether they
     lived independently, mobility level, baseline exercise tolerance)?

Audit Questions for Comprehensive Past Med History:
1) Does the discharge summary state the patient's past medical history (e.g. previous diagnoses or chronic
     conditions)?

2) Does the discharge summary state the patient's past surgical history? Answer "N/A" if there is an explicit
     statement that the patient has no prior surgeries.

3) Is a pre-admission medication list documented?

4) If a pre-admission medication list is documented, does it include doses and frequencies (not just drug names)?
     Answer "N/A" if no pre-admission medication list is present.


Audit Questions for Goals-of-care documentation:
1) Is there any documentation of goals of care, advance directives, code status, or advance care planning (e.g.
     serious illness conversations, advance medical directives)?

--------------------

Output Format (STRICT - valid JSON only):

{
  "S": { "1": {...}, "2": {...} },
  "C": { "1": {...},
        "2": { "answer": "Yes/No/Unclear/N/A", ... },
        "3": {...},
        "4": { "answer": "Yes/No/Unclear/N/A", ... } },
  "G": { "1": {...} }
}

-------------------------
Discharge Summary:
{discharge_summary}
```

B.3. Prompt 3: Recorded Medication Changes + Expected Follow-up

```
You are a clinical documentation auditor who works on medication changes and follow-up instructions.

You will be given a discharge summary. Your task is to answer the following audit questions based ONLY on the
     information present in the discharge summary.

Note:
- You are working with a de-identified dataset, information maybe explicitly stated but the details of it maybe
     blank (e.g. contact information)
- Give justification clearly when dealing with information which has blanks or dashes

Rules:
- Do NOT infer or assume information.
- Answers must be strictly one of: "Yes", "No", "Unclear", or "N/A".
- Use "Unclear" ONLY if partial or ambiguous information is present.
- If the information is completely absent, answer "No".
- Use "N/A" ONLY when the question's precondition does not apply (e.g., a conditional question whose triggering
     condition is not met).
- Evidence must be a direct quote or exact phrase(s) from the discharge summary.
- Justification must briefly explain why the evidence supports the selected answer.
- Do NOT add any content outside the specified JSON structure.
```

```
Audit Questions for Record of Medication Changes:
1) Is a discharge medication list documented?

2) If a discharge medication list is documented, does it include the purpose or indication for each medication?
     Answer "N/A" if no discharge medication list is present.

3) If a discharge medication list is documented, does it include dose, route, and/or frequency information? Answer
     "N/A" if no discharge medication list is present.

4) Are any medication changes (new medications started, medications stopped, or dose adjustments) clearly
     documented?

5) For medication changes that are documented, is the specific clinical rationale for each change provided? Answer
     "N/A" if no medication changes are documented.

6) For medications stopped during the stay, is there a clear plan for whether or when they should be restarted?
     Answer "N/A" if no medications were stopped.

Audit Questions for Expected Follow-up instructions:
1) Are follow up instructions or appointments included in the discharge summary?

2) Are there clear instructions regarding which outstanding investigations or pending results need to be reviewed
     or traced in the outpatient setting?

3) Is the contact information for the Primary Care Provider (PCP) listed in the summary, even if de-identified or
     blank?

--------------------

Output Format (STRICT - valid JSON only):

{
  "R": { "1": {...},
        "2": { "answer": "Yes/No/Unclear/N/A", ... },
        "3": { "answer": "Yes/No/Unclear/N/A", ... },
        "4": {...},
        "5": { "answer": "Yes/No/Unclear/N/A", ... },
        "6": { "answer": "Yes/No/Unclear/N/A", ... } },
  "E": { "1": {...}, "2": {...}, "3": {...} }
}

-------------------------
Discharge Summary:
{discharge_summary}
```

B.4. Prompt 4: History & Examinations

```
You are a clinical documentation auditor who works on history of presenting complaint and physical examination
     findings.

You will be given a discharge summary. Your task is to answer the following audit questions based ONLY on the
     information present in the discharge summary.

Note:
- You are working with a de-identified dataset, information maybe explicitly stated but the details of it maybe
     blank (e.g. contact information)
- Give justification clearly when dealing with information which has blanks or dashes

Rules:
- Do NOT infer or assume information.
- Answers must be strictly one of: "Yes", "No", or "Unclear".
- Use "Unclear" ONLY if partial or ambiguous information is present.
- If the information is completely absent, answer "No".
- Evidence must be a direct quote or exact phrase(s) from the discharge summary.
- Justification must briefly explain why the evidence supports the selected answer.
- Do NOT add any content outside the specified JSON structure.


Audit Questions for History of presenting complaint and
physical examination findings:
1) Does the discharge summary document the reason for the patient's admission?

2) Does the discharge summary mention the admission date?

3) Does the discharge summary document the source of referral or mode of admission (e.g. self-referral, emergency
     department, transfer from another facility)?
```

```
4) Does the discharge summary document vital signs or clinical parameters on presentation?

5) Does the discharge summary document targeted physical examination findings on presentation?

6) Is the presenting symptom characterized with any detail (e.g. nature, onset, duration, progression,
     alleviating/exacerbating factors)?

7) Are associated symptoms or significant negatives (especially to rule out red-flag symptoms) documented?

8) Is relevant surgical history, drug history, or family history documented where pertinent to the presenting
     complaint (e.g. risk factors affecting pretest probability or differential diagnosis)?

--------------------

Output Format (STRICT - valid JSON only):

{ "H": { "1": {...}, "2": {...}, "3": {...}, "4": {...},
       "5": {...}, "6": {...}, "7": {...}, "8": {...} } }

-------------------------
Discharge Summary:
{discharge_summary}
```

B.5. Prompt 5: Assessment & Clinical Course

```
You are a clinical documentation auditor who works on assessment and clinical course.

You will be given a discharge summary. Your task is to answer the following audit questions based ONLY on the
     information present in the discharge summary.

Note:
- You are working with a de-identified dataset, information maybe explicitly stated but the details of it maybe
     blank (e.g. contact information)
- Give justification clearly when dealing with information which has blanks or dashes

Rules:
- Do NOT infer or assume information.
- Answers must be strictly one of: "Yes", "No", "Unclear", or "N/A".
- Use "Unclear" ONLY if partial or ambiguous information is present.
- If the information is completely absent, answer "No".
- Use "N/A" ONLY when the question's precondition does not apply (e.g., a conditional question whose triggering
     condition is not met).
- Evidence must be a direct quote or exact phrase(s) from the discharge summary.
- Justification must briefly explain why the evidence supports the selected answer.
- Do NOT add any content outside the specified JSON structure.


Audit Questions for Assessment & Clinical Course:
1) Are medical diagnoses given in the summary (actual medical diagnosis, not just symptoms)?

2) Is the severity or complication level of the main diagnoses clearly described (e.g., KDIGO stage for AKI)?

3) Where appropriate, does the summary include a brief one-sentence problem representation explaining the key
     features that support the diagnosis?

4) Are clinical investigations listed (i.e blood tests, lab tests, imaging, diagnostic procedures)?

5) Is there a concise description of the patient's hospital course or clinical trajectory during admission?

6) Does the summary describe the management plan for each main problem, including conservative measures,
     pharmacologic treatments, and any procedures or surgeries?

7) Is the response to treatment documented for each major problem (e.g., resolution of symptoms, improvement in
     oxygen requirement, trending of creatinine)?

8) If recommended investigations or treatments were withheld or stopped, is the reason documented (e.g., patient
     preference, goals of care, futility, risk greater than benefit)? Answer "N/A" if no investigations or
     treatments appear to have been withheld or stopped.

--------------------

Output Format (STRICT - valid JSON only):

{ "A": { "1": {...}, "2": {...}, "3": {...}, "4": {...},
       "5": {...}, "6": {...}, "7": {...},
       "8": { "answer": "Yes/No/Unclear/N/A", ... } } }

-------------------------
Discharge Summary:
```

```
{discharge_summary}
```

B.6. Prompt 6: Discharge Information (Additional)

```
You are a clinical documentation auditor who works on general discharge documentation completeness.

You will be given a discharge summary. Your task is to answer the following audit questions based ONLY on the
     information present in the discharge summary.

Note:
- You are working with a de-identified dataset, information maybe explicitly stated but the details of it maybe
     blank (e.g. contact information)
- Give justification clearly when dealing with information which has blanks or dashes

Rules:
- Do NOT infer or assume information.
- Answers must be strictly one of: "Yes", "No", or "Unclear".
- Use "Unclear" ONLY if partial or ambiguous information is present.
- If the information is completely absent, answer "No".
- Evidence must be a direct quote or exact phrase(s) from the discharge summary.
- Justification must briefly explain why the evidence supports the selected answer.
- Do NOT add any content outside the specified JSON structure.


Audit Questions for Additional:
1) Is the date of discharge documented?

2) Is the specialty of the doctor that discharged the patient included in the summary?

3) Is the discharge disposition documented (e.g. discharged home, rehab, skilled nursing facility, step-down care)?

4) Is the type of discharge documented (e.g. normal, against medical advice, abscondment)?

5) Is the condition of the patient at discharge described (e.g. stable, improved, critical)?

6) Is hospital contact information listed for patient perusal, even if de-identified or blank?

7) Is information about the discharge summary writer included, even if de-identified?

8) Is the attending physician or discharging provider identified in the summary, even if de-identified?

--------------------

Output Format (STRICT - valid JSON only):

{ "Additional": {
   "1": {...}, "2": {...}, "3": {...}, "4": {...},
   "5": {...}, "6": {...}, "7": {...}, "8": {...} } }

-------------------------
Discharge Summary:
{discharge_summary}
```

C. Distribution of Clinician `Unclear` Labels

Across the 50 annotated discharge summaries and 46 audit questions, the clinician applied the `Unclear` label 38 times (1.7% of all labels). Table 4 reports the per-question distribution. The label is heavily concentrated: three questions account for 18 of the 38 `Unclear` labels, and 14 of the 46 questions received no `Unclear` labels in any summary.

Several patterns emerge. The single most ambiguous question is pre-hospitalization functional status (S2), which is rarely documented as a discrete field in MIMIC-IV summaries. Medication-related ambiguity in the Recorded Medication Changes (R) component (questions R2, R3, R4, R5, R6 collectively contribute 9 `Unclear` labels), reflecting that even when medications are listed, the surrounding details required by the audit (purpose, dose/route/frequency, change rationale, restart plans) are inconsistently provided. The long tail of single-hit questions across the remaining components indicates that `Unclear` is otherwise rare.

Table 5 reports how three representative models, the two top performers by overall κ (Gemini 3 Flash, Sonnet 4.5) and the locally deployed Qwen 2.5-7B, labeled the 38 questions on which the clinician selected `Unclear`. First, Gemini and Sonnet almost never reproduce the clinician's `Unclear` label (1 and 2 cases respectively), instead forcing a definitive Yes/No on genuinely ambiguous documentation. Critically, the two models resolve this ambiguity in *opposite directions*

Table 4. Per-question distribution of clinician `Unclear` labels across 50 summaries. Questions not appearing in the table received zero `Unclear` labels.

ID	Question (abbreviated)	Count
S2	Pre-hospitalization functional status	11
R6	Restart plan for stopped medications	4
S1	Social history documented	3
C2	Past surgical history documented	3
R5	Clinical rationale for medication changes	2
Add4	Type of discharge documented	2
I2	Allergens and reaction types documented	2
A5	Hospital course / clinical trajectory described	1
Add3	Discharge disposition documented	1
R3	Discharge med list: dose/route/frequency	1
R2	Discharge med list: purpose/indication	1
Add5	Patient's condition at discharge described	1
E3	PCP contact information listed	1
E2	Outstanding investigations flagged for follow-up	1
R4	Medication changes clearly documented	1
H6	Presenting symptom characterized with detail	1
A2	Severity/complication of main diagnoses	1
A8	Reason documented for withheld/stopped treatment	1
	Total	**38**

on the same 38 pairs, Gemini selects Yes 20 times while Sonnet selects No 20 times. That two top-tier models disagree systematically on clinician-flagged ambiguous cases is itself evidence that the underlying documentation is genuinely ambiguous. Second, Qwen 2.5-7B hedges substantially more readily (10 `Unclear` and 5 `N/A` out of 38), suggesting that smaller locally-deployed models may exhibit the opposite failure mode from the API tier. Further analysis is needed on these findings as follow-up work.

Table 5. Model label distributions on the 38 (summary, question) pairs where the clinician labeled `Unclear`. Gemini and Sonnet rarely produce `Unclear`, splitting their forced choices in opposite directions on the same pairs; Qwen hedges far more readily.

Model	Yes	No	Unclear	N/A
Gemini 3 Flash	20	17	1	0
Sonnet 4.5	15	20	2	0
Qwen 2.5-7B[†]	10	13	10	5

[†]Locally deployed.