Medical Summarization in Practice: Design, Deployment, and Analysis of a Clinical Summarization System for a German Hospital

Moiz Rauf
myScribe Gmbh, Germany
m.rauf@myscribe.de

Sean Papay
Fundamentals of Natural Language Processing
University of Bamberg, Germany
sean.papay@uni-bamberg.de

Abstract

Over the course of hospital treatment, a large number of electronic health records (EHRs) are created for a patient, detailing aspects of care history such as lab results, physician notes, and treatments administered. At the conclusion of treatment, this collection of EHRs must be summarized into a discharge summary, describing the course of care clearly and cohesively. In this paper, we present the design and development of a clinical summarization system integrated into a live German hospital workflow to help with the generation of discharge summaries. We first describe the system, its components, and its context of use within a hospital, before performing a number of experiments to gain insights into how best to use and evaluate our system. We investigate summarization performance across multiple input encoding strategies, compare expert judgments against automatic evaluation of summaries, and analyze the consistency of model summaries across multiple text generations. This work not only serves as a case study to demonstrate the feasibility of LLM integration into healthcare infrastructure, but also provides actionable insights into the use and evaluation of such systems.

1 Introduction

The increasing digitization of healthcare has led to the accumulation of massive volumes of temporally distributed patient data in electronic health records (EHRs). These records, while central to clinical decision-making, are often inconsistently structured, lengthy, and cognitively demanding to navigate (Yadav et al., 2018; Hossain et al., 2023).

One of the most important artifacts generated from the EHR is the "*discharge summary*", a concise narrative that details inpatient treatment and helps in outpatient care. However, manually writing detailed hospital course summaries remains a time-intensive bottleneck for doctors (Arndt et al., 2017; Overhage and McCallie Jr, 2020; Alissa et al., 2022). This contributes to increased clinician workload, de-motivation in doctors, career-switching and possible burnout (Haycock et al., 2014; Shanafelt et al., 2015; Robertson et al., 2017), highlighting the need for automated tools to assist doctors in summary generation.

Recent advancements in large language models (LLMs) have demonstrated encouraging generalization capabilities for clinical summarization tasks (Keszthelyi et al., 2023; Bednarczyk et al., 2025). These models exhibit potential in handling heterogeneous EHR inputs and producing coherent summaries across various medical domains. Nevertheless, a number of challenges hinder progress towards wider-scale application of such technologies in practice. First, the summarization task is inherently complex, as models must encompass both domain- and language-specific knowledge, handle long-range dependencies, and perform accurate temporal reasoning. Second, adaptability of clinical summarization systems for most languages is limited due to the majority of the existing research focusing on English-language corpora, impacting the applicability of much of this research outside of English-speaking countries (Pal et al., 2023; Heilmeyer et al., 2024). Third, privacy concerns regarding patient data place strong constraints on the availability and use of task-specific training data. Finally, due to the sensitive nature of medical decision-making systems, many failure modes are simply intolerable, raising the threshold for acceptability much higher than for general-domain tasks.

In this work, we present and analyze an LLM-based clinical summarization system as currently deployed in a German hospital. We make contributions of two kinds: Firstly, we present our deployed model as a case study for the application of large language models to clinical summarization in practical hospital settings. We discuss the constraints placed upon our model's design by its deployment

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 5: Industry Track, pages 455–466

context, and how these constraints informed the specific design decisions taken. Secondly, we perform an experimental analysis of this system to better understand its effectiveness. Concretely, we make the following experimental contributions:

- We compare four data encoding schemes and evaluate their impact on summary quality across four expert-annotated axes: readability, completeness, logical clarity, and medical precision.
- We explore the relationship between automated evaluation metrics and human evaluation.
- We investigate the semantic consistency of our system when generating multiple summaries from the same data.

Through this work, we hope to provide other practitioners with actionable insights into the design, deployment, and analysis of similar medical summarization systems in practice.

2 Related Work

LLM for Clinical Summarization Prior works have investigated the automatic generation of discharge summaries and clinical narratives using machine learning & neural models (Shing et al., 2021; Hartman and Campion, 2022; Hartman et al., 2023). However, recent advancements in LLMs have showcased the viability of generating discharge summaries, hospital course narratives, and progress notes, with improvements in language quality and abstraction (Keszthelyi et al., 2023). These model have been adopted through prompt optimizations (Chuang et al., 2024; Socrates et al., 2024; Ganzinger et al., 2025), chain-of-thought prompting (Tang et al., 2024), fine-tuning, or domain-specific adapter training (Van Veen et al., 2024; Heilmeyer et al., 2024) in a *Direct Generation* setting. Other works have explored *RAG-based* approaches to handle long patient documents (Saba et al., 2024; Myers et al., 2025; Lopez et al., 2025). Finally, Kruse et al. (2025) attempted to capture temporal dependencies in clinical text in string format for summarization.

Prompt Content Formatting A common problem faced in the zero-shot use of LLMs is *prompt brittleness* (He et al., 2024; Liu et al., 2025; Ceron et al., 2024), wherein slight variations in prompt format can significantly affect the output quality of LLMs. This is particularly relevant to the current work, where the input format must represent the rich structures present in FHIR patient records. Existing approaches have sought to either elicit LLM responses in a manner independent of any particular prompt format (e.g. Ngweta et al., 2025), or alternatively to optimize prompt format for the task at hand (e.g. Lu et al., 2022; Oh et al., 2025).

As we expect input formatting to be critical to LLMs' ability to reason about the structures in patient records, our work takes the latter approach, seeking the best input representation for FHIR data. We develop and systematically evaluate different encoding formats for real patient records.

3 Clinical Summaries in Practice

Clinical data in hospitals is accumulated over time by multiple professionals and stored across heterogeneous systems. Physicians must identify the pieces of information relevant to the current admission, synthesize them into a coherent narrative, and filter out unrelated or routine content (Adams et al., 2021; Hartman et al., 2023). A representative FHIR-based record illustrating this structure is provided in Appendix A.

Physician Expectations. Clinicians expect discharge summaries to be concise, accurate, and clinically meaningful. Unlike extractive summaries, discharge letters require selecting salient events, establishing temporal and causal relations, and presenting them in an interpretable narrative. Models must therefore integrate diverse events, use correct terminology, and avoid hallucinations by grounding descriptions in the actual patient record.

Deployment Context. Our study is part of an ongoing pilot deployment in a German hospital. The LLM runs fully on-premise within the hospital's secure environment under GDPR[1] constraints. The available hardware and privacy requirements rule out cloud-based or closed-source systems and excluded fine-tuning at this stage, making prompt design and input encoding especially critical. Generated summaries are not used directly; instead, they are shown to clinicians, who integrate, edit, or rewrite the content for inclusion in official discharge documentation.

4 System Description

Figure 1 provides a schematic overview of our summarization system, and how it is used within the context of a hospital to assist doctors with the generation of discharge summaries. This section details the components of our system, the structure of the

[1] https://gdpr-info.eu/

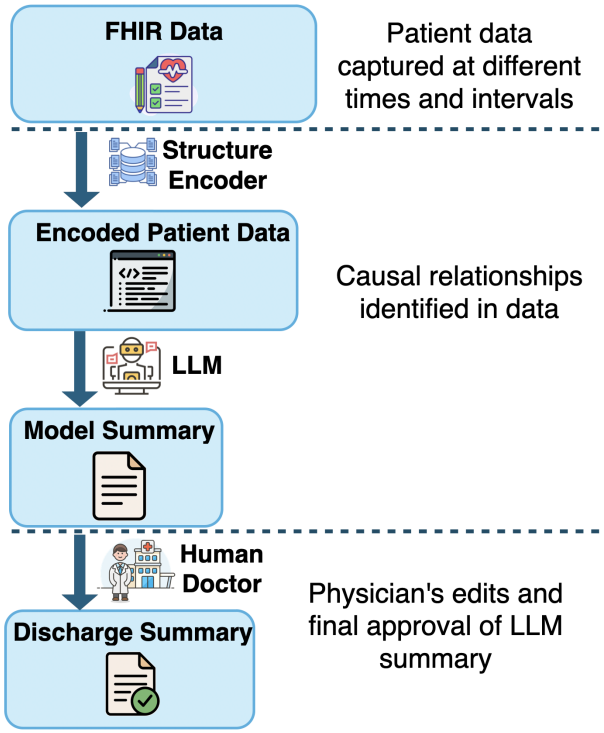

Figure 1: Illustration the summarization system pipeline. The process consumes structured FHIR data, encodes that as LLM input, and produces summary candidates for the Doctors.

patient data which it is tasked with summarizing, and its context of use within a German hospital.

4.1 Task Definition

Given a patient record $R = \{e_1, e_2, \ldots, e_T\}$, where each event $e_t = (c_t, v_t, d_t)$ is triple of category c_t, value v_t, and timestamp d_t, we would like to assist doctors in producing a final German-language discharge summary. To do this, our system automatically produces a *model summary* $S = \text{LLM}(P \oplus \text{Enc}(R))$, where P is a natural-language prompt describing the medical summarization task, and $\text{Enc}(\cdot)$ encodes the record into a sequence format readable and interpretable by the language model LLM. This model summary is then provided to a doctor, who may edit and verify its contents to be used as an official discharge summary.

As discussed in Section 3 privacy concerns require that no patient data leave the hospital premises, meaning that this language model must be physically hosted within the hospital. As such, based on work from (Sivarajkumar et al., 2024), we develop a zero-shot inference system, where a fixed natural-language prompt and the encoded patient record are provided directly to an out-of-the-box Mistral Small 3.2. This allows us to perform local inference on modest hardware without any initial need for pre-training with patient data.

The remainder of this section details the process for generating S, which constitutes our main methodological contribution. Pseudo-code for this approach is detailed in Appendix B.1.

4.2 Input Structuring: Temporal-Hierarchical Hybrid

As mentioned above, without any task-specific fine-tuning, the model relies entirely on how the input is structured to infer temporal order and clinical relevance. Our system combines two approaches to structure its input:

- **Temporal Windowing**: The timeline is chunked into fixed-width intervals according to time stamp.
- **Hierarchical Event Grouping**: Within each window, events are grouped into clinical categories (vitals, labs, meds, procedures, diagnoses, notes).

The next section describes how this structuring is made concrete through the specific input formatting strategies used for the model. Visual example of the input structuring is detailed in Appendix B.2

4.3 Encoding Format

Building on the temporal–hierarchical structuring described above, we investigate four distinct strategies for encoding patient histories into LLM prompts:

Flat Text: naive string serialization without structure.

JSON: object-based grouping of dates and categories.

Markdown: layout-optimized headers and bullets.

XML Tags: synthetic XML-style labeled spans.

Our encoding designs preserve temporal coherence and hierarchical grouping, reflecting both clinical and model-friendly structuring principles. Difference between encoding schemes are detailed in Appendix B.3

4.4 Summarization Model

As text-generation model, our system uses Mistral Small 3.2, a multilingual decoder-only transformer language model with 128K token context. For zero-shot inference, we sample texts from our model using parameters detailed in Appendix B.4.

5 Experiments

In this section, we present a number of experiments we carried out in order to better understand our

summarization system and its performance. In particular, these experiments investigate three research questions:

RQ1: How does data encoding strategy affect model output?

RQ2: How do common automatic evaluation metrics for text generation relate to expert judgments of summaries?

RQ3: How consistent are generated summaries?

5.1 Dataset

For all experiments, we require a dataset of patient histories to summarize. Similar to previous studies (e.g. Heilmeyer et al., 2024; Ganzinger et al., 2025), we selected a small set of 31 real patients from Internal Medicine department of the Pilot Hospital for this exploratory study. The dataset consisted of patient records to be used as model input and actual discharge letters which served as Gold Standard. The patients were further categorized into three types (*short, medium, long*) based on their stay, Appendix C lists some descriptive statistics of these patient histories.

5.2 RQ1: Effect of Encoding Strategy

To investigate this question, we conduct a comparative evaluation across four encoding strategies: unstructured Flat-Text (baseline), JSON, Markdown, and XML-based across patients with varying hospital stay durations. The central question driving this experiment was: *Does the encoding strategy used to represent structured patient data affect the quality and reliability of LLM-generated clinical summaries?* Answering this question is of particular importance to the domain of medical summarization, where rich structures, both latent and explicit, exist in both model input and output, and encoding scheme might significantly affect models' ability to capture those structures. For each encoding variant, structured patient records were converted into the corresponding format and presented to the model as input, along with a natural language prompt describing the medical summarization task.

Evaluation The generated summaries were independently rated by three clinicians across four axes: *Logical Clarity (Clr.),Completeness (Comp.), Medical Precision (Prec.), Readability (Read.)* using a 5-point Likert scale (1 = poor, 5 = excellent):[2] See Appendix D for detailed definition. We also consider the average of these values as a simple aggregate quality measure for summaries.

[2] In actuality, clinicians were asked to rate these axes according to the German academic grading scale, wherein 1 represents excellent and 5 represents poor. In order to facilitate understanding, the numbers reported in this paper are 6 minus the grade assigned by clinicians.

Stay	Enc.	Clr.	Comp.	Prec.	Read.	Avg.
Short	Flat Text	2.407	2.431	2.906	2.285	2.504
	Json	(2.484)	(2.546)	2.983*	(2.477)	(2.619)
	Markdown	(2.599)	2.893*	3.329**	(2.554)	2.841*
	XML	(2.022)	(2.200)	(2.483)	1.670*	(2.091)
Medium	Flat Text	2.694	2.802	3.416	2.704	2.904
	Json	(2.916)	3.163*	3.666**	(3.093)	3.210*
	Markdown	(2.750)	(2.913)	(3.388)	(2.787)	(2.960)
	XML	(2.805)	(2.969)	(3.499)	(2.398)	(2.918)
Long	Flat Text	2.067	2.084	2.689	1.800	2.153
	Json	2.949**	3.011*	3.748**	3.271**	3.212**
	Markdown	2.714*	2.966*	3.689**	2.976**	3.09**
	XML	(2.243)	(2.378)	(2.983)	(1.976)	(2.388)
All	Flat Text	2.458	2.533	3.090	2.370	2.612
	Json	2.774*	2.925**	3.444*	2.927**	3.017**
	Markdown	(2.686)	2.915**	3.434*	2.75*	2.951**
	XML	(2.42)	(2.47)	(3.052)	2.066	2.533

Table 1: Average evaluation metrics for each encoding scheme across short-, medium-, and long-stay patients. We report the statistical significance of our encoding schemes relative to the baseline Flat-Text format, as measured by a mixed-effects model. Significance: $*p < 0.05$, $**p < 0.01$, $()p \geq 0.05$.

To estimate the contribution of each encoding scheme to expert-judged summary quality, we employed a linear mixed-effects model:

$$Y_{ijk} = \beta_0 + \beta_1 \cdot \text{Enc}_j + u_i + \epsilon_{ijk} \quad (1)$$

Here, Y_{ijk} denotes the score assigned by annotator i to encoding scheme j on summary instance k. The fixed effect β_1 captures the contribution of each encoding scheme, while u_i accounts for annotator-specific biases. We assume $u_i \sim \mathcal{N}(0, \sigma_u^2)$; $\epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2)$.

Results Table 1 summarizes the results of our analysis of expert judgments for all metrics and patient groups. For short-stay patients, Markdown yielded the strongest overall performance, outperforming JSON and XML-based formats across most axes. The XML-based representation produced no improvements and significantly reduced readability and precision ($p < 0.01$). For medium-stay patients, JSON outperformed all other encodings, producing positive coefficients across clarity, completeness, precision, readability, and the overall score. Markdown remained competitive but did not consistently exceed the baseline. For long-stay patients, JSON demonstrated the most robust gains, with statistically significant improvements across nearly all axes ($p < 0.01$ for precision, readability, and average quality). Markdown also showed consistently positive effects but remained slightly weaker than JSON. The XML-based format again

failed to show significant benefit, with coefficients remaining small and non-significant across all metrics.

Discussion and Implications This experiment highlights the *brittleness* of prompt formatting: seemingly minor structural changes result in substantial performance variation. In terms of overall performance, JSON formatting is highly effective, presumably as its explicit hierarchical organization that aligns well with the model's ability to track temporal and categorical relationships across longer narratives. Nonetheless, for short stays, where such explicit structure may not be as important, the lightweight structural cues provided by Markdown appear sufficient to support the model in producing clear and readable summaries. While our bespoke XML-based encoding scheme also explicitly marks structure in syntax, any performance gains from this seem to be overshadowed by the model's difficulty in parsing XML set not familiar from the language model's pre-training. These findings motivate us to employ patient-stay-specific approaches for our system as opposed to a unified encoding strategy.

5.3 RQ2: Automatic Evaluation vs. Expert Judgments

Motivated by the previous studies (Deutsch et al., 2021; Casola et al., 2025), we explore the question *Can standard automatic evaluation metrics such as* ROUGE, BLEU, *and* BERTSCORE *reliably approximate clinician-rated summary quality in production settings?* In addition to manual annotation described earlier, we assessed generated summaries from each encoding scheme against actual discharge letters using commonly used syntactic and semantic metrics ROUGE and BLEU for surface-level fluency (Lin, 2004; Post, 2018), and BERTSCORE for semantic fidelity in medical texts (Zhang* et al., 2020). Furthermore, in order to test if these metrics are good proxies for human annotation, we fit a linear regression model to predict the average human score using these metrics as input features.

Encoding	ROUGE	BLEU	BERTScore
Flat Text	0.487	12.4	0.863
JSON	0.526	15.6	0.875
Markdown	0.515	14.6	0.871
XML	0.492	12.0	0.864

Table 2: Average automatic evaluation scores (ROUGE, BLEU, and BERTSCORE) for each encoding scheme. JSON format achieves the highest performance across all three metrics.

Metric	Coef.	Std. Err.	*p*-value
Intercept	-3.650**	1.388	0.009
ROUGE-Lsum	(-0.237)	0.207	0.255
ROUGE-L	(0.204)	0.196	0.300
ROUGE-1	1.017**	0.222	<0.001
ROUGE-2	(0.146)	0.144	0.311
BERTSCORE	3.606**	1.036	0.001
BLEU	(0.137)	0.145	0.349

Table 3: Regression Coefficients for Predicting Human Rating. Significance: *$p < 0.05$, **$p < 0.01$, ()$p > 0.05$

While these metrics are inherently task-specific and sensitive to dataset characteristics, their combined use may provide a broader perspective on summarization quality and facilitate comparison with prior work (e.g. Xu et al., 2024).

Results and Implications As shown in Table 2, the patterns closely mirror those observed in the expert evaluations, JSON-encoded inputs consistently outperformed other schemes across most metrics, particularly ROUGE and BERTSCORE, indicating better content preservation and semantic alignment with reference summaries. Markdown offered moderate improvements over the unstructured Flat-Text format, likely benefiting from light structural cues. In contrast, XML-based encoding performed comparably or worse than the Flat-Text baseline, potentially due to its deviation from token patterns observed during model pretraining. Finally, BLEU scores were relatively low across all formats, which is expected in an abstractive summarization task where exact n-gram overlap is rare. This result is in agreement with existing works (Peyrard, 2019; Ernst et al., 2023).

Table 3 shows the model coefficients for each metric, showing that BERTSCORE ($\beta = 3.606$, $p = 0.001$) and ROUGE-1 ($\beta = 1.017$, $p <$

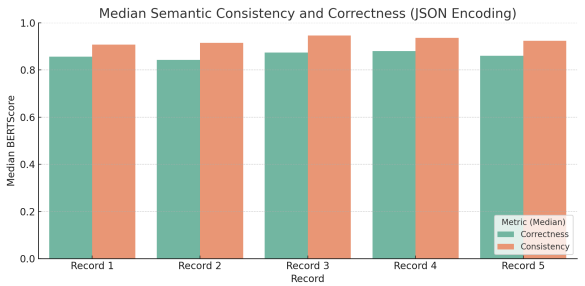

Figure 2: Semantic Consistency (SC_{median}) and Semantic Correctness (SS_{median}) for each patient record using the JSON prompt format.

Category	JSON Encoding	Flat Text Encoding
Admission Context	✓Clearly stated at the beginning	✓Mentioned at the start
Logical Clarity	✓Focused on current admission, symptoms, and key diagnostics	✗Included irrelevant past diagnoses and routine lab data
Medical Accuracy	✓Relevant findings included ✓Colonoscopy, ultrasound findings well integrated ✓Antibiotics, fluid therapy included ✗used irrelevant linking ward notes	✗Speculative or incorrect recommendations ✓Endoscopy, imaging findings summarized ✓Included relevant therapy; notes improvement ✗Misrepresented care with confusing notes
Follow-up Info	✗Not completely clarified	✗Unstructured with mixed observations and irrelevant history not grounded in data ✗Hallucinated follow-up steps
Language Quality	✗Minor phrasing issues	✓Consistent, grammatically correct but verbose

Table 4: Comparison of Clinician Feedback on JSON vs Flat-Text Encoded Summaries.

0.001) are statistically significant predictors of clinician-assigned ratings. However, the model's moderate explanatory power ($R^2 \approx 0.33$) suggests that, while automatic metrics can capture some of the signal in expert annotations, they leave a substantial portion of the variance unaccounted for.

These results indicate usefulness of these metrics as engineering tools for system development, providing a lightweight proxy signal that enables rapid experimentation, model iteration, and regression testing before committing to more resource-intensive expert evaluations.

5.4 RQ3: Summary Consistency

As observed in previous experiments, the JSON encoding format consistently outperformed other prompt formats across multiple human-evaluated dimensions and automated metrics. To further validate its robustness in a real-world setting, we examine whether summaries generated using the JSON content formatting exhibit semantic stability across repeated runs and maintain high clinical alignment with reference summaries.

Method In line with Atil et al. (2024), we assess stability of the JSON-encoding by generating 10 summaries for 5 patient records using the same generation parameters. As defined by (Carandang et al., 2025), we calculate Semantic Consistency (SC) by computing pairwise BERTSCORE across all generations, whereas for Semantic Correctness (SS) generated outputs are compared against gold standard summaries.

As shown in Figure 2, JSON-formatted prompts lead to consistently high semantic stability ($SC_{median} > 0.90$), indicating minimal variability across generations. Semantic correctness (SS_{median}) values also remain strong ($\sim$0.85–0.88), confirming our model also retains alignment with clinically accurate content. These results provide evidence for the robustness and output reliability of EHR summarization systems, at least when using a strong input encoding strategy.

5.5 Qualitative Evaluation by Clinicians

To assess the practical implications of our system design, a clinician was tasked with conducting a qualitative analysis of summaries generated from JSON and Flat-Text formats. The feedback was categorized into different type classes reflecting doctors' criteria. All comments were originally provided in German and subsequently translated into English for ease of presentation.

The feedback detailed in Table 4 shows the difference between summaries generated by different encoding schemes. It highlights that representing input in JSON format better enables the model to capture clinical relationships and temporal dependencies as compared to simple Flat-Text formulation. However, identifying events relevant to current medical discourse remain a challenge. Similarly, the system often failed to formulate correct follow-up steps; this information was not provided in the FHIR source data, often resulting in hallucinated text.

6 Conclusion

In this work, we presented a case study of a deployed EHR summarization system with design decisions grounded in NLP-driven analysis and empirical insights. We also presented experiments investigating our system's performance under input variation, comparisons of manual and automatic evaluation, and the consistency of our system's output. These experiments shed valuable and actionable insights into how best to use and evaluate our model for real-world medical summarization. We expect the findings presented in this work to directly guide the development of future iterations of our system, and hope that this study and others like it can help to build a better understanding of what

is needed for high-quality medical summarization.

Limitations

Despite demonstrating the feasibility of an on-premise LLM-based summarization system and providing detailed analyses on input structuring and encoding strategies, several limitations remain.

Our experiments use a single open-source model (Mistral Small 3.2), selected due to on-premise hardware, privacy, and compliance constraints. Although alternative models (e.g., GPT-OSS etc) could be deployed similarly, model comparison was not the aim of this pilot; the study instead focuses on how input structuring and encoding choices affect summarization quality under realistic operational conditions. Our system operates strictly in a zero-shot configuration due to deployment and runtime constraints; while this aligns with on-premise requirements, it prevents adaptation to local documentation practices and may limit performance compared to supervised or RLHF-based approaches. Techniques such as retrieval augmentation, synthetic fine-tuning, or controlled adaptation were not explored in this pilot but constitute promising directions for future work.

The analysis is further limited to data from a single German hospital, whose documentation style and workflows may not generalize to other institutions. While this hinders broad applicability of our results, the proposed pre-processing and encoding pipeline provides a transferable foundation for evaluating the approach in additional clinical settings.

Similarly the volume of summaries and evaluations process performed by practicing clinicians was necessarily limited by clinical workloads. This may narrow the range of perspectives represented. The mapping from the German grading system to a five-point scale may also introduce interpretive variability, and differences in individual rating styles could influence the results. While automatic metrics show only moderate correspondence with clinician judgments, and the regression analyses explain limited variance ($R^2 \approx 0.33$), indicating that these measures, although useful for iterative development, do not capture the full range of clinical quality considerations.

Finally, the study did not investigate the system under atypical or degraded EHR conditions, such as very high event density, inconsistent timestamps, or partially missing data. This highlights the need for further robustness testing in broader deployment settings.

Acknowledgments

We thank the clinicians and medical domain experts who contributed to the annotation and evaluation of clinical summaries. In particular, we would like to thank Dr. Ira Stoll, Dr. Georg Brosinsky and Dr. Kira Knauer, whose clinical expertise and detailed feedback were instrumental in defining the evaluation criteria and interpreting the results. We also gratefully acknowledge the partner hospital for providing access to de-identified patient data and for supporting the on-premise deployment and evaluation of the system within a real clinical workflow under GDPR constraints.

Ethical Considerations

All data processing took place within the hospital's secured infrastructure, and all records were de-identified in accordance with GDPR requirements; no information was transmitted to external services or cloud environments. However, even properly de-identified clinical narratives may contain residual contextual signals that could increase re-identification risk if mishandled. Appropriate safeguards, auditing mechanisms, and access controls remain essential.

The system is designed strictly as a support tool, clinicians retain full responsibility for verifying, editing, and approving every generated summary. Allowing automated summaries to enter the clinical record without oversight could introduce risks through hallucinations, omissions, or ambiguous statements, and our deployment therefore followed a strict "human-in-the-loop" model.

Furthermore, differences in documentation practices across demographics, departments, and institutions may also lead to uneven model performance. As the dataset used in this study comes from a single hospital, broader fairness and bias assessments were not possible and should be prioritized in future work. Similarly, as automated summarization becomes integrated into clinical workflows, attention must be paid to its impact on clinical labor. While such systems can reduce administrative workload, poorly designed automation may shift cognitive burden onto clinicians or induce over-reliance on generated content. Responsible deployment requires ensuring that the system augments rather than replaces expert judgment,

and that transparency, accountability, and public trust are maintained as LLMs become more widely adopted within healthcare.

Finally, under the EU AI ActAct[3], clinical summarization systems qualify as high-risk applications, requiring traceability, logging, interpretability, and human oversight. Our pipeline aligns with these requirements by maintaining on-premise compute, transparent prompt structures, and strict clinician verification. However, additional governance mechanisms error reporting, monitoring dashboards, retraining workflows would be necessary for large-scale deployment.

References

Griffin Adams, Emily Alsentzer, Mert Ketenci, Jason Zucker, and Noémie Elhadad. 2021. What's in a summary? laying the groundwork for advances in hospital-course summarization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4794–4811, Online. Association for Computational Linguistics.

Rana Alissa, Jennifer A Hipp, and Kendall Webb. 2022. Saving time for patient care by optimizing physician note templates: a pilot study. *Frontiers in Digital Health*, 3:772356.

Brian G Arndt, John W Beasley, Michelle D Watkinson, Jonathan L Temte, Wen-Jan Tuan, Christine A Sinsky, and Valerie J Gilchrist. 2017. Tethered to the ehr: primary care physician workload assessment using ehr event log data and time-motion observations. *The Annals of Family Medicine*, 15(5):419–426.

Berk Atil, Alexa Chittams, Liseng Fu, Ferhan Ture, Lixinyu Xu, and Breck Baldwin. 2024. Llm stability: A detailed analysis with some surprises. *CoRR*, abs/2408.04667.

Lydie Bednarczyk, Daniel Reichenpfader, Christophe Gaudet-Blavignac, Amon Kenna Ette, Jamil Zaghir, Yuanyuan Zheng, Adel Bensahla, Mina Bjelogrlic, and Christian Lovis. 2025. Scientific evidence for clinical text summarization using large language models: Scoping review. *J Med Internet Res*, 27:e68998.

Kristine Ann M. Carandang, Jasper Meynard Arana, Ethan Robert Casin, Christopher Monterola, Daniel Stanley Tan, Jesus Felix B. Valenzuela, and Christian Alis. 2025. Are LLMs reliable? an exploration of the reliability of large language models in clinical note generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pages 1413–1422, Vienna, Austria. Association for Computational Linguistics.

[3]https://artificialintelligenceact.eu/

Silvia Casola, Yang Janet Liu, Siyao Peng, Oliver Kraus, Albert Gatt, and Barbara Plank. 2025. References matter: Investigating the impact of reference set variation on summarization evaluation. pages 274–291.

Tanise Ceron, Neele Falk, Ana Barić, Dmitry Nikolaev, and Sebastian Padó. 2024. Beyond prompt brittleness: Evaluating the reliability and consistency of political worldviews in LLMs. *Transactions of the Association for Computational Linguistics*, 12:1378–1400.

Yu-Neng Chuang, Ruixiang Tang, Xiaoqian ng, and Xia Hu. 2024. Spec: a soft prompt-based calibration on performance variability of large language model in clinical notes summarization. *Journal of biomedical informatics*, 151:104606.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2021. A statistical analysis of summarization evaluation metrics using resampling methods. *Transactions of the Association for Computational Linguistics*, 9:1132–1146.

Ori Ernst, Ori Shapira, Ido Dagan, and Ran Levy. 2023. Re-examining summarization evaluation across multiple quality criteria. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13829–13838, Singapore. Association for Computational Linguistics.

Matthias Ganzinger, Nicola Kunz, Pascal Fuchs, Cornelia K Lyu, Martin Loos, Martin Dugas, and Thomas M Pausch. 2025. Automated generation of discharge summaries: leveraging large language models with clinical data. *Scientific Reports*, 15(1):1–13.

V Hartman and T R Campion. 2022. A Day-to-Day approach for automating the hospital course section of the discharge summary. *AMIA Jt Summits Transl Sci Proc*, 2022:216–225.

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.

Michael Haycock, Laura Stuttaford, Oliver Ruscombe-King, Zoe Barker, Kathryn Callaghan, and Timothy Davis. 2014. Improving the percentage of electronic discharge summaries completed within 24 hours of discharge. *BMJ Quality Improvement Reports*, 3(1).

Jia He, Mukund Rungta, David Koleczek, Arshdeep Sekhon, Franklin X Wang, and Sadid Hasan. 2024. Does prompt formatting have any impact on llm performance? *arXiv preprint arXiv:2411.10541*.

Felix Heilmeyer, Daniel Böhringer, Thomas Reinhard, Sebastian Arens, Lisa Lyssenko, and Christian Haverkamp. 2024. Viability of open large language models for clinical documentation in german health care: Real-world model evaluation study. *JMIR Medical Informatics*, 12:e59617.

Elias Hossain, Rajib Rana, Niall Higgins, Jeffrey Soar, Prabal Datta Barua, Anthony R. Pisani, and Kathryn Turner. 2023. Natural language processing in electronic health records in relation to healthcare decision-making: A systematic review. *Computers in Biology and Medicine*, 155:106649.

Daniel Keszthelyi, Christophe Gaudet-Blavignac, Mina Bjelogrlic, Christian Lovis, and 1 others. 2023. Patient information summarization in clinical settings: scoping review. *JMIR Medical Informatics*, 11(1):e44639.

Maya Kruse, Shiyue Hu, Nicholas Derby, Yifu Wu, Samantha Stonbraker, Bingsheng Yao, Dakuo Wang, Elizabeth Goldberg, and Yanjun Gao. 2025. Zero-shot large language models for long clinical text summarization with temporal reasoning. *medRxiv*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yuanye Liu, Jiahang Xu, Li Lyna Zhang, Qi Chen, Xuan Feng, Yang Chen, Zhongxin Guo, Yuqing Yang, and Peng Cheng. 2025. Beyond prompt content: Enhancing llm performance via content-format integrated prompt optimization. *arXiv preprint arXiv:2502.04295*.

Ivan Lopez, Akshay Swaminathan, Karthik Vedula, Sanjana Narayanan, Fateme Nateghi Haredasht, Stephen P Ma, April S Liang, Steven Tate, Manoj Maddali, Robert Joseph Gallo, and 1 others. 2025. Clinical entity augmented retrieval for clinical information extraction. *npj Digital Medicine*, 8(1):45.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Skatje Myers, Timothy A Miller, Yanjun Gao, Matthew M Churpek, Anoop Mayampurath, Dmitriy Dligach, and Majid Afshar. 2025. Lessons learned on information retrieval in electronic health records: a comparison of embedding models and pooling strategies. *Journal of the American Medical Informatics Association*, 32(2):357–364.

Lilian Ngweta, Kiran Kate, Jason Tsay, and Yara Rizk. 2025. Towards LLMs robustness to changes in prompt format styles. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 529–537, Albuquerque, USA. Association for Computational Linguistics.

Jio Oh, Geon Heo, Seungjun Oh, Hyunjin Kim, JinYeong Bak, Jindong Wang, Xing Xie, and Steven Euijong Whang. 2025. Better think with tables: Tabular structures enhance llm comprehension for data-analytics requests. *Preprint*, arXiv:2412.17189.

J Marc Overhage and David McCallie Jr. 2020. Physician time spent using the electronic health record during outpatient encounters: a descriptive study. *Annals of internal medicine*, 172(3):169–174.

Koyena Pal, Seyed Ali Bahrainian, Laura Mercurio, and Carsten Eickhoff. 2023. Neural summarization of electronic health records. *CoRR*, abs/2305.15222.

Maxime Peyrard. 2019. Studying summarization evaluation metrics in the appropriate scoring range. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5093–5100.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Sandy L Robertson, Mark D Robinson, and Alfred Reid. 2017. Electronic health record effects on work-life balance and burnout within the i3 population collaborative. *Journal of graduate medical education*, 9(4):479–484.

Walid Saba, Suzanne Wendelken, and James Shanahan. 2024. Question-answering based summarization of electronic health records using retrieval augmented generation. *CoRR*, abs/2401.01469.

Tait D Shanafelt, Omar Hasan, Lotte N Dyrbye, Christine Sinsky, Daniel Satele, Jeff Sloan, and Colin P West. 2015. Changes in burnout and satisfaction with work-life balance in physicians and the general us working population between 2011 and 2014. In *Mayo clinic proceedings*, volume 90, pages 1600–1613. Elsevier.

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

Sonish Sivarajkumar, Mark Kelley, Alyssa Samolyk-Mazzanti, Shyam Visweswaran, and Yanshan Wang. 2024. An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: algorithm development and validation study. *JMIR Medical Informatics*, 12:e55318.

Vimig Socrates, Thomas Huang, Xuguang Ai, Soraya Fereydooni, Qingyu Chen, R Andrew Taylor, and David Chartash. 2024. Yale at "discharge me!": Evaluating constrained generation of discharge summaries with unstructured and structured information. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 724–730.

An Quang Tang, Xiuzhen Zhang, and Minh Ngoc Dinh. 2024. IgnitionInnovators at "discharge me!": Chain-of-thought instruction finetuning large language models for discharge summaries. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 731–739, Bangkok, Thailand. Association for Computational Linguistics.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, and 1 others. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature medicine*, 30(4):1134–1142.

Justin Xu, Zhihong Chen, Andrew Johnston, Louis Blankemeier, Maya Varma, Jason Hom, William J. Collins, Ankit Modi, Robert Lloyd, Benjamin Hopkins, Curtis Langlotz, and Jean-Benoit Delbrouck. 2024. Overview of the first shared task on clinical text generation: RRG24 and "discharge me!". pages 85–98.

Pranjul Yadav, Michael Steinbach, Vipin Kumar, and Gyorgy Simon. 2018. Mining electronic health records (ehrs): A survey. *ACM Comput. Surv.*, 50(6).

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

A Example FHIR-Style Record

The examples illustrates a raw clinical data, originally stored as FHIR resources. Hospitals save patient information in different information systems which is aggregated as a FHIR resource before being passed to the model for summarization. Each patient record contains multiple resource types (e.g., Observations, Medications, Procedures, Notes), each representing time-stamped clinical events recorded throughout the hospital stay.

```
Patient: P001
E1 (2024-05-10):
Obs: BP 140/85 (07:45)
HR 98 bpm (08:00)
Med: Amoxicillin 1g IV (09:00)

E2 (2024-05-11):
Obs: CRP 75 mg/L (10:15)
Proc: Abd-US (14:30)
result: cholecystitis
Note: RUQ pain + fever (16:00)
```

B Input Structuring and Encoding

B.1 Zero-Shot Inference Pipeline

The summarization system takes as input a patient record R in FHIR format, an encoding choice ENCTYPE specifying one of the formatting strategies described in Section4.3 a prompt template P, and the language model LLM. The procedure consists of five stages:

1. **Event Extraction:** All clinical events (e.g., vitals, labs, procedures, notes) and their timestamps are extracted from R to form an event set E.
2. **Temporal–Hierarchical Structuring (Section 4.2):** The event set E is partitioned into a sequence of fixed-width temporal windows W, and within each window events are grouped by clinical category, producing a structured representation $\mathcal{H}$.
3. **Encoding (Section 4.3):** The structured representation $\mathcal{H}$ is converted into an encoded text sequence X using the selected encoding scheme ENCTYPE.
4. **Prompt Construction:** The encoded patient record X is inserted into the natural-language prompt template P to form the model input $\tilde{P}$.
5. **Zero-Shot Inference:** The model LLM processes $\tilde{P}$ and generates a summary S without any task-specific fine-tuning.

B.2 Temporal–Hierarchical Structuring

As detailed in Section 4.2 our system first flattens these heterogeneous resources into a unified list of events, preserving only the information necessary for summarization: timestamps, clinical categories, and event values. For this paper we refer to the FHIR record example provided in Appendix A

Flattened Events

```
05-10-2024 07:45 vitals BP 140/85
05-10-2024 08:00 vitals HR 98
05-10-2024 09:00 meds   Amoxi 1g IV
05-11-2024 10:15 labs   CRP 75
05-11-2024 14:30 imag   US:cholecystitis
05-11-2024 16:00 notes  RUQ pain+fever
```

These Events are grouped into fixed-width temporal windows and, within each window, sorted according to broad clinical categories (such as vitals, labs, imaging, notes). This process exposes the temporal progression of the patient's condition while presenting the model with a clinically meaningful, human-aligned grouping of information.

Algorithm 1: Zero-Shot Clinical Summarization with Temporal–Hierarchical Encoding

Input: FHIR patient record R
Input: Encoding type ENCTYPE $\in$ {Flat-Text, JSON, Markdown, XML}
Input: Language model LLM
Input: Prompt template P
Output: Generated discharge summary S

$E \leftarrow$ extract all events (t_i, c_i, v_i) from R
$W \leftarrow$ divide E into fixed-size temporal windows
foreach W_k *in* W **do**
  $G_k \leftarrow$ group events in W_k by clinical category
end
$\mathcal{H} \leftarrow \{G_1, G_2, \ldots, G_n\}$
$X \leftarrow$ encode $\mathcal{H}$ using format ENCTYPE
$\tilde{P} \leftarrow$ insert encoded record X into prompt template P
$S \leftarrow \text{LLM}(\tilde{P})$
return S

Windows (t = 24h)

```
W1 (05-10-2024)
- Vitals: BP 140/85; HR 98
- Meds: Amoxi 1g IV

W2 (05-11-2024)
- Labs: CRP 75
- Imaging: US cholecystitis
- Notes: RUQ pain + fever
```

B.3 Encoding Formats

As mentioned in Section 4.3 the intermediate structure is transformed into four concrete encoding formats used during zero-shot inference: a lightweight Flat-Text format, a structured JSON object, a Markdown layout optimized for readability, and a XML-based representation using synthetic markers. These formats differ in the degree of structure they expose to the model, providing a controlled way to study how input representation affects summarization quality. The examples below show how the same underlying patient record is rendered through each of these stages.

Flat-Text

```
05-10-2024: Vitals BP140/85 HR98;
Med Amoxi1g.
05-11-2024: CRP75; US cholecystitis;
RUQ pain+fever.
```

JSON

```
{
 "05-10-2024":{
  "vitals":["BP140/85","HR98"],
  "meds":["Amoxi1g IV"]
 },
 "05-11-2024":{
  "labs":["CRP75"],
  "imaging":["US cholecystitis"],
  "notes":["RUQ pain+fever"]
 }
}
```

Markdown

```
## 05-10-2024
- Vitals: BP140/85; HR98
- Meds: Amoxi1g IV

## 05-11-2024
- Labs: CRP75
- Imaging: US cholecystitis
- Notes: RUQ pain+fever
```

XML

```
<event>
<date>"05-10-2024"<\date>
<vital>BP140/85; HR98</vital>
<meds>Amoxi1g IV</medds>
</event>

<event>
<date>"05-11-2024"<\date>
<labs>CRP75</labs>
<image>US cholecystitis</image>
<note>RUQ pain+fever</note>
</event>
```

B.4 Model Configuration

The following table 5 details the decoding hyperparameters used for generating clinical summaries with the `Mistral-Small-3.2-24B-Instruct-2506` [4] model. These settings were selected to ensure high factual fidelity and consistency, optimized for a production clinical summarization environment.

Parameter	Value
Temperature	0.15
Top-p	0.85
Max Tokens	2500
Min Tokens	100
Number of Completions (n)	1
Best-of	1

Table 5: Generation hyperparameters used for the Mistral-Small-24B-Instruct model.

C Descriptive Statistics of Patient Data

We analyzed patient records based on hospital stay duration. Records were grouped into three categories: *short stays* (fewer than 3 days), *medium stays* (3 to 7 days), and *long stays* (8 days or more).

[4] https://huggingface.co/mistralai/Mistral-Small-3.2-24B-Instruct-2506

This stratification helps to understand how the complexity and volume of medical documentation vary across different clinical scenarios.

Table 6 presents key statistics for each group, including the average length of stay, number of visits, procedures, radiology reports, diagnoses, clinical entries, and total token count. As expected, longer stays are associated with higher procedural complexity and richer documentation.

Category	Short	Medium	Long	Overall
# Records	11	13	7	31
Stay (days)	1.55	4.38	30.00	9.16
Visits	0.45	2.38	27.6	7.38
Procedures	2.1	1.76	7.2	3.1
Radiology	1.2	3.4	11.4	4.4
Diagnoses	1.2	2	1.5	2
Others	1.2	3.4	6.14	3.4
Word Count	305.1	924	3478.9	1281.5

Table 6: Summary statistics of patient records stratified by hospital stay duration.

D Annotation Criteria

To assess the clinical quality of generated summaries, we adopted a structured annotation schema comprising four key criteria: *completeness of critical information*, *medical accuracy and terminology*, *clarity of diagnostic and therapeutic logic*, and *readability*. These dimensions were selected to reflect both clinical validity and textual usability in real-world hospital settings. Given that our evaluation was conducted in a German clinical environment, annotators followed the standard German *Schulnotensystem* grading scale, where **1** indicates the best possible score (excellent) and **5** the worst (insufficient). This system was familiar to clinicians and facilitated consistent, context-appropriate assessment across summaries.

Criterion	Description & Guidelines
Completeness	Checks whether the summary includes all essential and relevant clinical details. Review for missing or incorrect facts and omissions that could affect patient care.
Medical Precision	Evaluates use of accurate terminology and whether clinical concepts are correctly represented (e.g., incorrect abbreviations or mislinked findings).
Clarity of Logic	Assesses clarity and logical structure of diagnostics and treatment. Checks for coherence in the treatment course and appropriate sequencing of medical information.
Readability	Focuses on fluency, professional tone, and grammatical quality. Highlights unclear or verbose sections, and identifies overly simplified language.

Table 7: Annotation Criteria for Evaluating Clinical Summaries. For this study the rating scale follows the following scale: 1 = perfect, 5 = poor