`DistillNote`: LLM-based clinical note summaries improve heart failure diagnosis

Anonymous ACL submission

Abstract

Large language models (LLMs) offer unprecedented opportunities to generate concise summaries of patient information and alleviate the burden of clinical documentation that overwhelms healthcare providers. We present `Distillnote`,[1] a framework for LLM-based clinical note summarization, and generate over 64,000 admission note summaries through three techniques: (1) One-step, direct summarization, and a divide-and-conquer approach involving (2) Structured summarization focused on independent clinical insights, and (3) Distilled summarization that further condenses the Structured summaries. We test how *useful* are the summaries by using them to predict heart failure compared to a model trained on the original notes. Distilled summaries achieve 79% text compression and up to 18.2% improvement in AUPRC compared to an LLM trained on the full notes. We also evaluate the *quality* of the generated summaries in an LLM-as-judge evaluation as well as through blinded pairwise comparisons with clinicians. Evaluations indicate that one-step summaries are favoured by clinicians according to relevance and clinical actionability, while distilled summaries offer optimal efficiency (avg. 6.9× compression-to-performance ratio) and significantly reduce hallucinations. We release our summaries on PhysioNet to encourage future research.

1 Introduction

Electronic health records (EHRs) contain diverse patient health information, including free-text clinical notes. However, notes are often lengthy and filled with medical abbreviations and jargon, making important information difficult to extract (Shing et al., 2021; Liang et al., 2019). Using clinical note summaries, i.e., a distillation of the most important clinical insights without the noise, may lead to improved performance on clinical tasks, but manually creating them is impractical (Chuang et al., 2023).

Large language models (LLMs) have shown increasing promise for automating clinical note summarization, with recent studies indicating they perform at least as well as, and often better than, human experts (Veen et al., 2023; Schoonbeek et al., 2024). *One-step summarization* is a standard method where text is summarized in one pass (with one prompt). However, this approach often struggles with long medical documents, potentially missing critical information and introducing inaccuracies. An alternative is to apply a *divide-and-conquer strategy*, which aims to break down complex tasks into more manageable subtasks. This has proven effective in reducing LLM output inconsistencies by up to 90% in hallucination detection in news summaries (Cui et al., 2024), mitigating intermediate errors in tasks like arithmetic and fake news detection (Zhang et al., 2024), and summarizing long academic articles (Gidiotis and Tsoumakas, 2020).

To this end, we introduce `Distillnote` (Fig. 1), the first divide-and-conquer framework for clinical note summarization that preserves diagnostic signals while *significantly* reducing text volume. Our contributions include: (1) the application of a hierarchical summarization approach to clinical texts, whereby an LLM first summarizes key medical elements independently before further distilling them into even shorter summaries; (2) the first comprehensive analysis of how LLM-generated clinical note summaries impact a downstream clinical task (heart failure); (3) a dual-axis quality assessment, combining an LLM-as-judge methodology and manual validation by clinicians; and (4) a public dataset of 64,000+ clinical summaries to advance research in clinical NLP.

[1] To ensure reproducibility, all code will be available at `https://github.com/username/distillnote` upon acceptance.

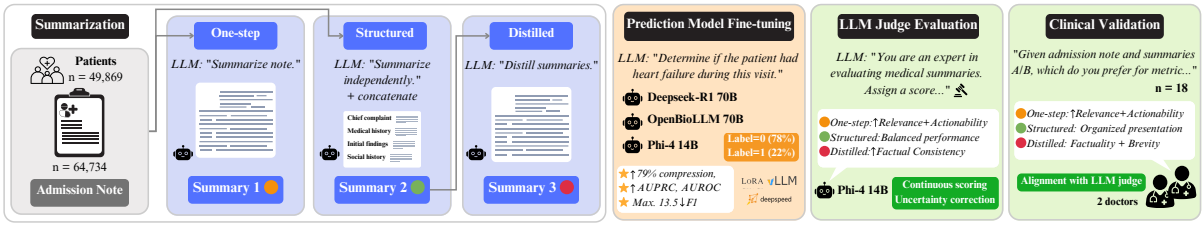

Figure 1: Overview of `Distillnote`. We generate admission note summaries, predict heart failure, and evaluate outputs using both LLM-as-judge and clinicians.

2 Methodology

Let $\{N_i\}_{i=1}^{M}$ denote the set of M *admission notes*. Each admission note is represented as $N_i = \{t_1, t_2, \cdots, t_N\}$, where t_j is the j-th token and N is the total number of tokens in note N_i.

Dataset Our prediction task require clinical notes written at hospital admission.[2] To avoid leakage from upcoming post-admission sections, we truncate discharge summaries from MIMIC-IV (Johnson et al., 2023b,a) into *admission notes* as per Röhr et al. (2024). Admission sections include *chief complaint*, *history of present illness*, *medical history*, *admission medications*, *allergies*, *physical exam*, *family history*, and *social history*. Our final dataset consists of 64,734 admission notes from 49,869 patients (details in A.3).

2.1 LLM-based summarization

One-step summarization A **one-step summary** $S_i^{\text{one-step}}$is produced by, given an admission note N_i, directly generating the summary in one shot:

$$\mathcal{S}_i^{\text{one-step}} = \text{LLM}([q_{\text{one-step}}; N_i]), \quad (1)$$

where $q_{\text{one-step}}$ is a summarization prompt, and [] denotes concatenation.

Divide-and-conquer summarization We define a set of four prompts $\{P_k\}_{k=1}^{4}$, each targeting one clinical insight at admission: *chief complaint*, *medical history*, *exam findings*, and *social/family background*. For each note N_i, the prompts $\{P_k\}_{k=1}^{4}$ are processed independently by an LLM:

$$t_{i,k} = \text{LLM}([P_k; N_i]), \quad (2)$$

where $t_{i,k}$ represents the LLM-generated summary for clinical insight k given an admission note N_i.

We generate a **structured summary** $\mathcal{S}_i^{\text{struct}}$ by concatenating all four intermediate summaries for note N_i as follows:[3]

$$\mathcal{S}_i^{\text{struct}} = [t_{i,1}; t_{i,2}; t_{i,3}; t_{i,4}]. \quad (3)$$

Next, we produce a **distilled summary** S_i^{distill} for note N_i that further condenses all the important insights in the patient's clinical trajectory. Note that we only use the structured summaries $\mathcal{S}_i^{\text{struct}}$ generated in the previous step as input:

$$S_i^{\text{distill}} = LLM([q_{\text{distill}}; \mathcal{S}_i^{\text{struct}}]), \quad (4)$$

where q_{distill} is the final summarization query.

Overview We compare three sets of summaries for each note N_i. $S_i^{\text{one-step}}$is a standard summary generated with an LLM in one pass; S_i^{struct} is a structured summary where clinical insights are summarized independently and then appended, and $S_i^{\text{distill}}$is the distilled summary that integrates the structured summaries into a single, final summary.

Please refer to A.4 and A.5 for details on the prompts used for summarization and sampling parameters, respectively.

2.2 Summary evaluation

LLM-as-judge Due to the large volume of generated summaries, complete human evaluation is unfeasible. Thus, we follow recent summarization research and implement an LLM-as-judge framework with Phi-4 to assess the quality of the best-performing summaries (Bavaresco et al., 2024). We implement continuous scoring and take inspiration from Liu et al. (2023) and use the probability the LLM assigns to its scores to account for uncertainty in the final output (see A.8 for details).

Our evaluation is based on three 1–5 scale metrics: **Relevance**, **Factual Fabrication** (both

[2] We use admission notes as they represent the earliest clinical documentation in the patient visit. This enables anticipating complications during the current hospitalization.

[3] We add section headers for each clinical insight k in $\mathcal{S}_i^{\text{struct}}$, but omit these in Eq. (3) to avoid clutter.

adapted from Krolik et al. 2024), and **Clinical Actionability**, which we propose in this work. *Relevance* measures inclusion of essential medical details, *fabrication* checks for hallucinations, and *clinical actionability* assesses support for decision-making. LLM judge prompts (Appendices A.12, A.13, A.14) include the original note, generated summary, metric definition, and few-shot examples. We performed one/two-way ANOVA with post-hoc Tukey's HSD tests and Cohen's d calculations to analyze result differences.

Clinician validation Two board-certified clinicians participated in a blinded pairwise comparison study, evaluating 18 admission notes and their summaries. Clinicians were presented with the original note followed by two unlabeled summaries (A and B) in randomized order. For each pair, they indicated their preference across the same three metrics as the LLM judge. Binomial tests were used for each metric to determine if preferences across cases were statistically significant ($p < 0.01$) with Bonferroni correction for multiple comparisons.

2.3 Heart failure (HF) prediction

We fine-tune LLMs for HF prediction using the labels from Chen et al. (2024a). Patients appear in only one split (train/val/test) and we use stratification maintaining class distributions, i.e., 78% negative vs. 22% positive (see A.7 for details). Each model was trained on four input types: $S_i^{\text{one-step}}$, S_i^{struct}, S_i^{distill}, and N_i (the original admission notes). Given the imbalanced nature of the task, AUROC and AUPRC are our main metrics (McDermott et al., 2024), also reporting F1-scores for completeness.

2.4 Large language models (LLMs)

We use three LLMs across our experiments when generating summaries (§2.1) and HF prediction (§2.3): `Deepseek-R1-Distill-Llama-70B`,[4] `OpenBioLLM-70B`,[5] i.e., a LLaMA-based model fine-tuned on medical data, and `Phi-4-14B`.[6] Each has demonstrated strong performance in complex reasoning tasks and improved alignment with human preferences (Abdin et al., 2024; Lee et al., 2025; Subramanian et al., 2025; Pal and Sankarasubbu, 2024; Grattafiori et al., 2024; DeepSeek-AI et al., 2025).

[4] https://huggingface.co/Deepseek-ai/Deepseek-R1-Distill-Llama-70B

[5] https://huggingface.co/aaditya/Llama3-OpenBioLLM-70B

[6] https://huggingface.co/microsoft/phi-4

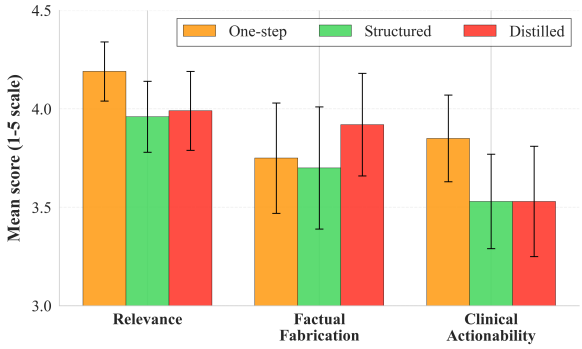

Figure 2: LLM-as-judge results for different summaries.

3 Results

3.1 LLM-as-judge summary evaluation

All summarization approaches achieved scores well within the 'adequate' to 'very good' ranges across all metrics (Figure 2), and statistical analysis indicated significant differences between strategies ($p \ll 0.01$, refer to A.9 for details). The One-step approach achieved the highest overall performance (3.93 ± 0.29), excelling in both relevance (4.19 ± 0.15) and clinical actionability (3.85 ± 0.22). The Distilled strategy demonstrated superior factual reliability (3.92 ± 0.26) with medium-to-large effect sizes compared to One-step ($d = 0.62$) and Structured ($d = 0.75$). The Structured approach had a more balanced performance but achieved the lowest overall scores (3.73 ± 0.30).

3.2 Clinician summary evaluation

Our pairwise comparisons revealed preferences without statistical significance ($\alpha = 0.01$). Clinicians favored One-step summaries for relevance and clinical actionability over both alternatives, with preference ratios of 8:4 and 7:5 compared to Structured summaries, and 9:3 versus Distilled summaries. These human preferences align with LLM evaluations, where One-step scored highest for both. For factual accuracy, Structured summaries were preferred over One-step (7:5), while Distilled summaries were preferred over Structured (7:5). A qualitative analysis of rater comments pointed that One-step summaries were "relevant," Structured summaries "organized," and Distilled summaries were described as "short" summaries that "suffice" for clear cases. Notably, one clinician reported difficulty in evaluating factual accu-

Strategy	S?	Model	AUROC	AUPRC	F1
Full note	✗	DS	0.820	0.423	0.766
	✗	OB	0.823	0.472	**0.770**
	✗	Phi-4	0.802	0.457	0.761
One-step	✓	DS	0.860	0.509	0.738
	✓	OB	0.860	0.540	0.708
	✓	Phi-4	0.850	0.526	0.715
Structured	✓	DS	0.859	0.507	0.727
	✓	OB	**0.873**	**0.576**	0.679
	✓	Phi-4	0.859	0.553	0.689
Distilled	✓	DS	0.857	0.521	0.712
	✓	OB	0.858	0.558	0.676
	✓	Phi-4	0.847	0.544	0.658

Table 1: Heart failure prediction performance. Full note baseline maintains highest F1 while summarization loses only 6.5-11.4% F1 (avg. per strategy w.r.t. best baseline). All approaches show improved AUROC/AUPRC scores over baseline, with Structured achieving highest discrimination. We highlight in red the three worst results per metric. **S?**: Summary? **DS**: Deepseek. **OB**: OpenBioLLM.

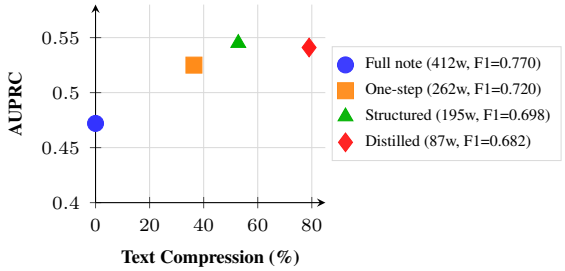

Figure 3: AUPRC scores versus text compression. All summaries show improved discrimination (+11.2-15.5%) compared to full notes, with Structured achieving highest AUPRC despite 53% compression.

racy due to the need to constantly cross-reference source documents and summaries, and that many summaries failed to put critical decision-making information upfront, corroborating the lower LLM clinical actionability scores.

3.3 Heart failure prediction

Predicting HF with summaries yields improvements across all models, and summarization consistently improves AUROC (3.8–5.0%) and AUPRC (11.2–14.6%) with modest F1-score reductions (6.5–11.4%) relative to the rate of text compression achieved (36.4–79.0%) (Table 1, Figure 3). The Structured approach with OpenBioLLM achieves the highest AUPRC and AUROC. Deepseek achieves the best efficiency ratios across strategies (10.5×, 9.4×, and 8.7× for Distilled, Structured, and One-step respectively), retaining 92.5-95.8% of baseline F1 while achieving 36-79% compression.

4 Discussion

LLM-generated note summaries maintain robust HF prediction performance while enhancing ranking metrics despite substantial length reductions (up to 79%), indicating models achieve sharper discrimination between positive and negative cases, essential for patient care decisions. This confirms LLMs can filter clinical documentation noise without compromising predictive utility. Furthermore, trade-offs exist between distillation strategies: One-step are better in relevance, Structured summaries yield the highest discrimination metrics, and Distilled summaries' superior factual reliability validates that our divide-and-conquer approach effectively helps address hallucination, a fundamental limitation in clinical LLMs, albeit at higher computational cost (4×). The alignment between clinician preferences and LLM evaluations for relevance and clinical actionability indicate automated metrics can approximate human judgment. However, consistently lower actionability scores indicate that beyond prompt engineering, structural constraints may be necessary to enforce upfront presentation of key clinical information, and clinicians' difficulty with factual verification highlights the need for better hallucination detection approaches.

5 Conclusion

We introduced the `DistillNote` framework for LLM-based clinical note summarization and investigate its downstream applicability on heart failure, an imbalanced clinical classification task. We compared using summaries generated with three summarization strategies with LLMs from three different model families on HF prediction. Summary quality was assessed both with an LLM judge and clinicians. Finally, we release our best summaries on Physionet. With this work, we show the value of underused textual data in EHRs and encourage further research whereby compressed summaries are used in a range of other patient predictions.

Future exploration may involve evaluating our framework on other note types (e.g., progress and discharge notes) and downstream clinical tasks (e.g., mortality prediction) and exploring other LLM backbones for generation and evaluation. We also believe that integrating other EHR modalities to ground textual claims to be a promising future direction. Furthermore, utilizing agent-based mechanisms to further mitigate hallucinations may further strengthen summary reliability.

5.1 Limitations

While DistillNote demonstrates promising results, it still has many limitations. First, our evaluations primarily focus on heart failure prediction with one EHR dataset, MIMIC-IV, potentially limiting generalizability to other clinical conditions and cohorts. Second, despite performing clinician evaluation, the sample size was relatively small (n = 18) compared to the scale of our dataset (n $\simeq$ 64k). Third, LLM-based evaluation, while comprehensive, may still inherit biases from the underlying model. Finally, our divide-and-conquer approach, while compressing the summaries further into a reduced number of tokens, also increases computational costs compared to one-step summarization due to the use of multiple prompts, which may impact accessibility.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. 2024. Phi-4 technical report. *Preprint*, arxiv:2412.08905 [cs].

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, André F. T. Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K. Surikuchi, Ece Takmaz, and Alberto Testoni. 2024. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. *Preprint*, arxiv:2406.18403 [cs].

Pei Chen, Jiming Sun, Yan Chu, Yue Jiang, Luyu Yuan, Jian Liang, Chun Jiang, Xiangyu Jiang, Hao Shen, Fulin Xu, and Zhenyu Huang. 2024a. Predicting in-hospital mortality in patients with heart failure combined with atrial fibrillation using stacking ensemble model: an analysis of the medical information mart for intensive care iv (mimic-iv). *BMC Medical Informatics and Decision Making*, 24(1):402.

Xiao Chen, Wei Zhou, Rashina Hoda, Andy Li, Chris Bain, and Peter Poon. 2024b. Exploring the opportunities of large language models for summarizing palliative care consultations: A pilot comparative study. *Digital Health*, 10:20552076241293932. Publisher: SAGE Publications Ltd.

Akash Choudhuri, Philip Polgreen, Alberto Segre, and Bijaya Adhikari. 2025. Summarizing clinical notes using llms for icu bounceback and length-of-stay prediction. *medRxiv preprint*. This article is a preprint and has not been peer-reviewed.

Yu-Neng Chuang, Ruixiang Tang, Xiaoqian Jiang, and Xia Hu. 2023. SPeC: A soft prompt-based calibration on performance variability of large language model in clinical notes summarization. *Preprint*, arxiv:2303.13035 [cs].

Wendi Cui, Zhuohang Li, Damien Lopez, Kamalika Das, Bradley A. Malin, Sricharan Kumar, and Jiaxin Zhang. 2024. Divide-conquer-reasoning for consistency evaluation and automatic improvement of large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 334–361. Association for Computational Linguistics.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, and et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint*.

David Fraile Navarro, Enrico Coiera, Thomas W. Hambly, Zoe Triplett, Nahyan Asif, Anindya Susanto, Anamika Chowdhury, Amaya Azcoaga Lorenzo, Mark Dras, and Shlomo Berkovsky. 2025. Expert evaluation of large language models for clinical dialogue summarization. *Scientific Reports*, 15(1):1195. Publisher: Nature Publishing Group.

Alexios Gidiotis and Grigorios Tsoumakas. 2020. A divide-and-conquer approach to the summarization of long documents. *Preprint*, arxiv:2004.06190 [cs].

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, and et al. 2024. The llama 3 herd of models. *arXiv preprint*.

S. Hegselmann, S. Shen, F. Gierse, M. Agrawal, D. Sontag, and X. Jiang. 2024a. Medical expert annotations of unsupported facts in doctor-written and llm-generated patient summaries (version 1.0.0).

S. Hegselmann, S. Z. Shen, F. Gierse, M. Agrawal, D. Sontag, and X. Jiang. 2024b. A data-centric approach to generate faithful and high quality patient summaries with large language models. *arXiv preprint arXiv:2402.15422*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*. Version 2, revised October 16, 2021.

Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2023a. MIMIC-IV-note: Deidentified free-text clinical notes.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Liwei H. Lehman, Leo A. Celi, and Roger G. Mark.

2023b. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1.

HyoJe Jung, Yunha Kim, Heejung Choi, Hyeram Seo, Minkyoung Kim, JiYe Han, Gaeun Kee, Seohyun Park, Soyoung Ko, Byeolhee Kim, Suyeon Kim, Tae Joon Jun, and Young-Hak Kim. 2025. Enhancing clinical efficiency through llm: Discharge note generation for cardiac patients. *Healthcare AI Research*. This study evaluates the use of Mistral-7B for automating discharge note generation for cardiac patients.

Jack Krolik, Herprit Mahal, Feroz Ahmad, Gaurav Trivedi, and Bahador Saket. 2024. Towards leveraging large language models for automated medical q&a evaluation. *Preprint*, arxiv:2409.01941 [cs].

Sunbowen Lee, Junting Zhou, Chang Ao, Kaige Li, Xinrun Du, Sirui He, Jiaheng Liu, Min Yang, Zhoufutu Wen, and Shiwen Ni. 2025. Distillation quantification for large language models. *Preprint*, arxiv:2501.12619 [cs].

Jili Li, Siru Liu, Yundi Hu, Lingfeng Zhu, Yujia Mao, and Jialin Liu. 2022. Predicting mortality in intensive care unit patients with heart failure using an interpretable machine learning model: Retrospective cohort study. *Journal of Medical Internet Research*, 24(8):e38082. Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada.

Jun Li, Yiwu Sun, Jie Ren, Yifan Wu, and Zhaoyi He. 2025. Machine learning for in-hospital mortality prediction in critically ill patients with acute heart failure: A retrospective analysis based on the MIMIC-IV database. *Journal of Cardiothoracic and Vascular Anesthesia*, 39(3):666–674.

Yizhan Li, Sifan Wu, Christopher Smith, Thomas Lo, and Bang Liu. 2024. Improving clinical note generation from complex doctor-patient conversation. *arXiv preprint arXiv:2408.14568*.

Jennifer Liang, Ching-Huei Tsou, and Ananya Poddar. 2019. A novel system for extractive clinical note summarization using EHR data. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 46–54, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Matthew B. McDermott, Haoran Zhang, Lasse Hyldig Hansen, Giovanni Angelotti, and Jack Gallifant. 2024. A closer look at auroc and auprc under class imbalance. In *Advances in Neural Information Processing Systems*, volume 37, pages 44102–44163. Curran Associates, Inc.

'Connor Kyle D. O, Yu Yamamoto, Sounok Sen, Marc D. Samsky, F. Perry Wilson, Nihar Desai, Tariq Ahmad, and Michael A. Fuery. 2023. Risk prediction for heart failure patients admitted to the intensive care unit. *JACC: Heart Failure*, 11(6):727–728. Publisher: American College of Cardiology Foundation.

Ankit Pal and Malaikannan Sankarasubbu. 2024. Openbiollms: Advancing open-source large language models for healthcare and life sciences. `https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B`.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 3505–3506. ACM.

Tom Röhr, Alexei Figueroa, Jens-Michalis Papaioannou, Conor Fallon, Keno Bressem, Wolfgang Nejdl, and Alexander Löser. 2024. Revisiting clinical outcome prediction for MIMIC-IV. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 208–217. Association for Computational Linguistics.

Rosanne Schoonbeek, Jessica Workum, Stephanie C.E. Schuit, Job Doornberg, Tom P. Van Der Laan, and Charlotte M.H.H.T. Bootsma-Robroeks. 2024. Completeness, correctness and conciseness of physician-written versus large language model generated patient summaries integrated in electronic health records.

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

Shreyas Subramanian, Vikram Elango, and Mecit Gungor. 2025. Small language models (SLMs) can still pack a punch: A survey. *Preprint*, arxiv:2501.05465 [cs].

D. Van Veen, C. Van Uden, L. Blankemeier, J.B. Delbrouck, A. Aali, C. Bluethgen, A. Pareek, M. Polacin, E.P. Reis, A. Seehofnerová, N. Rohatgi, P. Hosamani, W. Collins, N. Ahuja, C.P. Langlotz, J. Hom, S. Gatidis, J. Pauly, and A.S. Chaudhari. 2023. Clinical text summarization: Adapting large language models can outperform human experts. *Research Square [Preprint]*. Update in: Nat Med. 2024 Apr;30(4):1134-1142. doi: 10.1038/s41591-024-02855-5. PMID: 37961377; PMCID: PMC10635391.

Yizhou Zhang, Lun Du, Defu Cao, Qiang Fu, and Yan Liu. 2024. Guiding large language models with divide-and-conquer program for discerning problem solving. *Preprint*, arxiv:2402.05359 [cs]. Version: 1.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

A Appendix

A.1 Acknowledgements

A.2 Related work

Other researchers have explored clinical summarization using LLMs. Veen et al. (2023) detailed the superiority of LLM-generated summaries, but their approach did not compare multiple summarization strategies, neither employed LLM judges. Similarly, Choudhuri et al. (2025); Jung et al. (2025) concentrated on summarizing progress and discharge notes in one step with LLMs, respectively. Fraile Navarro et al. (2025); Li et al. (2024); Chen et al. (2024b) focused on producing note summaries but from medical conversations. Chuang et al. (2023) proposed a model-agnostic calibration method to enhance stability in LLM-based clinical summarization, while Hegselmann et al. (2024a,b) introduced strategies to reduce hallucinations. However, in these approaches, the exploration of multiple methodologies and evaluations for note summaries are lacking, as well clinical outcome prediction based on notes.

A.3 Admission note details

Admission notes provide initial patient assessments in the hospital, including chief complaints, medical history, and initial treatment plans. In our study, we focus on heart failure patients admitted to the intensive care unit (ICU), aligning with recent studies (Li et al., 2022, 2025; O et al., 2023).

As per Röhr et al. (2024), we truncate discharge notes into admission notes, keeping only sections available at admission time. The reader can find the statistics of each section in Table 2.

The dataset we used to generate the note summaries is MIMIC-IV, which consists of de-identified EHR information, and is available under a credentialed data use agreement via the Physionet portal. Our summaries are released under the PhysioNet Credentialed Health Data license and distributed solely for research purposes.

Section	Avg word/section	Std
Chief Complaint	7.83	34.21
Present Illness	193.55	130.63
Medical History	53.38	82.76
Medication Administered	64.50	63.39
Allergies	6.32	5.85
Physical Exam	75.08	54.82
Family History	13.40	14.52
Social History	0.02	3.41

Table 2: Statistics by admission note sections.

A.4 Summarization prompts

Each note $N \in \mathcal{N}$ is processed independently four times, with each concatenated to a different summarization prompt strategy. In the four S_i^{struct}prompts, we use a one-shot methodology, providing one example of the desired output format to give the LLM clear guidance on the structure and content of the target information. This aims to reduce task ambiguity, preventing the lack of structure in a zero-shot approach. In the S_i^{distill}prompt, we instruct the LLM to synthesize the four previous S_i^{struct}summaries into one final summary. The $S_i^{\text{one-step}}$prompt follows a similar structure, but instead instructs the LLM to summarize the full original admission note at once. Below the reader can find the prompts utilized for generation across approaches.

A.4.1 'One-step' prompt

> Summarize the patient's admission note, focusing on clinically relevant aspects affecting diagnosis, treatment, and risk assessment. For the synthesis, consider only explicitly documented information to maintain accuracy. Patient admission note: **note**

A.4.2 'Structured' prompts

Chief complaint

> Summarize the patient's primary reason for admission in one concise sentence based strictly on the chief complaint and history of present illness. Outputting more than one sentence or adding remarks or notes is strictly forbidden. Extract only explicitly documented information to maintain accuracy. Example output: 'The patient, a female, presented with worsening shortness of breath and lower extremity edema over three days.' Patient admission note: **note**

Past medical history

Summarize the patient's past medical history in one concise sentence, including chronic conditions, major illnesses, hospitalizations, and surgeries. Only mention ongoing medications if recently changed and allergies if clinically relevant. Outputting more than one sentence or adding remarks or notes is strictly forbidden. Extract only explicitly documented information to maintain accuracy. Example output: 'The patient, a male, has a history of hypertension, type 2 diabetes, and a prior myocardial infarction with stent placement.' Patient admission note: **note**

Physical exam

Summarize key physical exam findings in one concise sentence, including only significant abnormalities and pertinent negatives. Only include vital signs if explicitly relevant. Outputting more than one sentence or adding remarks or notes is strictly forbidden. Extract only explicitly documented information to maintain accuracy. Example output: 'The patient, a female, is afebrile with a distended abdomen, shifting dullness, and trace lower extremity edema.' Patient admission note: **note**

Social/family history

Summarize relevant family history, social determinants, and lifestyle factors in one concise sentence, focusing only on hereditary risks, substance use, living situation, occupational exposures, and support systems. Outputting more than one sentence or adding remarks or notes is strictly forbidden. Extract only explicitly documented information to maintain accuracy. Example output: 'The patient, a male, lives alone, has a history of heavy alcohol use, and has a father with a history of early-onset cardiovascular disease.' Patient admission note: **note**

Strategy	Model	Avg word	Std	Min	Max	Compr. (%)
One-go	Phi-4	262.04	41.18	106	438	36.4%
Structured	Phi-4	194.65	39.69	74	482	52.8%
Distilled	Deepseek	86.51	20.02	25	248	79.0%
Baseline	N/A	412.12	196.74	18	2425	0.0%

Table 3: Statistics across summarization strategies and baseline case (original admission notes). 'Model' refers to the LLM used to generate each summary type. Compression is calculated relative to the average word count of the baseline.

A.4.3 'Distilled' prompt

Summarize the summaries extracted from the patient's admission note into a single cohesive admission summary. Focus on clinically relevant aspects affecting diagnosis, treatment, and risk assessment. For the synthesis, consider only explicitly documented information to maintain accuracy. Patient summaries: **structured summary**

A.5 Sampling parameters

For summary generation, all models used the same decoding configuration across prompt types: `temperature`: 0.0, `top_p`: 0.9, `repetition_penalty`: 1.2. The `max_tokens` parameter varied:

- Phi-4, OpenBioLLM: 300 for S_i^{struct}, 700 for $S_i^{\text{one-step}}$ and S_i^{distill}
- Deepseek-R1: 2000 for all prompt types

Outputs from Deepseek include internal reasoning markers between `<think>` tags preceding the final summary, requiring the generation of more tokens. Hence, we parsed and isolated the main summary content after generation.

A.6 Summary details

We conducted preliminary fine-tuning evaluations to identify the most promising LLM for each summarization approach based on the validation F1-score. Based on that, we selected Deepseek for generating S_i^{distill} summaries, and Phi4 for S_i^{struct} and $S_i^{\text{one-step}}$ summaries. The statistics of the generated summaries, in comparison with the original admission notes, are present in Table 3.

A.7 Fine-tuning

We fine-tuned models using LoRA (Hu et al., 2021) with LoRA+ enhancement for parameter efficiency. Training utilized DeepSpeed ZeRO-3 (Rasley et al.,

Table 4: Dataset split and class distribution.

Split	Total Samples	% pos	% neg
Train	46,702	21.83%	78.17%
Validation	8,217	22.08%	77.92%
Test	9,815	21.67%	78.33%
Total	64,734	**21.86%**	**78.14%**

2020) for memory optimization with bf16 precision in NVIDIA H100 GPUs (4 for Deepseek and OpenBioLLM, 2 for Phi-4). We trained LLMs for one epoch using a cosine learning rate schedule (initial lr=5e-5, warmup=10%) and evaluated models every 80 steps, selecting the best checkpoints based on validation performance. The fine-tuning of 1 model took around 1 day. The training implementation used LLaMA-Factory (Zheng et al., 2024) with the Alpaca prompt template. Dataset statistics are available in Table 4.

The prompt we utilized for heart failure prediction can be found below. We note that, when predicting for the baseline case, we change "*admission note summary*" to "*admission note*".

> You are an expert in clinical diagnosis. Determine whether the patient had heart failure during this visit based on their admission note summary. Output only a number between double brackets: [[0]] for No, [[1]] for Yes. Patient summary: **summary**

A.7.1 Summary results

Table 5 details the performance of the three $S_i^{\text{one-step}}$, S_i^{struct}, and S_i^{distill}summarization strategies by comparing their average word count, compression percentage, and the performance with respect to the F1, AUROC, and AUPRC of the best baseline (original note). The efficiency ratio, calculated as % compression divided by F1 drop, indicates the trade-off between text distillation and prediction quality. A higher ratio indicates a more efficient summary, i.e. a summary with a greater reduction in length and relatively less impact on performance.

A.8 LLM-as-judge

In the LLM-as-judge step, we evaluate the quality of the generated summaries. We select Phi-4 as the judge model due to its demonstrated alignment with human judgment in evaluation tasks (Abdin et al., 2024) and strong reasoning capacity despite its medium size (14B parameters). This cost-effective performance is important given our need to evaluate over 64,000 summaries across 3 metrics and 3 summary strategies. For the methodology, like G-Eval (Liu et al., 2023), we account for model uncertainty by adjusting the obtained scores based on token probability distributions. This helps address model uncertainty and smoothes out extreme scores when the model shows lower confidence.

We use continuous scoring to allow for a more nuanced evaluation. Instead of directly extracting e.g "4.2" from model output, we analyze the underlying probability distribution across top-k possible digit tokens. For example, for a model generating a score with the first digit d_1 and second digit d_2, we extract probabilities $p(d_1)$ and $p(d_2)$, along with the probabilities of the top-k (k = 5) alternative numbers the model considered. We then calculate a weighted average of all possible score combinations, where the weight is the product of the digit probabilities. The final score is calculated as:

$$\text{Score} = \frac{\sum_{d_1 \in D_1} \sum_{d_2 \in D_2} p(d_1) \cdot p(d_2) \cdot \text{val}(d_1.d_2)}{\sum_{d_1 \in D_1} \sum_{d_2 \in D_2} p(d_1) \cdot p(d_2)} \quad (5)$$

where D_1 and D_2 are the sets of possible digits at each position and val() converts the digit concatenation to its numerical value.

The reader can find the distribution of original scores in Figure 4, and the adjusted scores per summarization approach in Table 6 and Figure 5. Figure 6 illustrates the adjustments made.

Strategy	Avg Words	Compression	↑ AUPRC	↑ AUROC	↓ F1
Baseline	412	—	—	—	—
One-step	262	36.4%	+11.2%	+4.1%	6.5%
Structured	195	52.8%	+15.5%	+5.0%	9.4%
Distilled	87	79.0%	+14.6%	+3.8%	11.4%

Table 5: Trade-off between performance metrics and text compression for different summarization strategies. Gains are calculated considering the average of the scores per strategy with respect to the best baseline model (F1=0.770, AUROC=0.823, AUPRC=0.472). Efficiency ratio = compression%/F1 drop%. One-step: 5.6×, Structured: 5.6×, Distilled: 6.9×.

Table 6: LLM-as-judge scores by summary approach and metric (mean ± SD).

Approach	Fabrication	Relevance	Actionability	Overall
One-step	3.75±0.28	**4.19±0.15**	**3.85±0.22**	**3.93±0.29**
Structured	3.70±0.31	3.96±0.18	3.53±0.24	3.73±0.30
Distilled	**3.92±0.26**	3.99±0.20	3.53±0.28	3.81±0.32

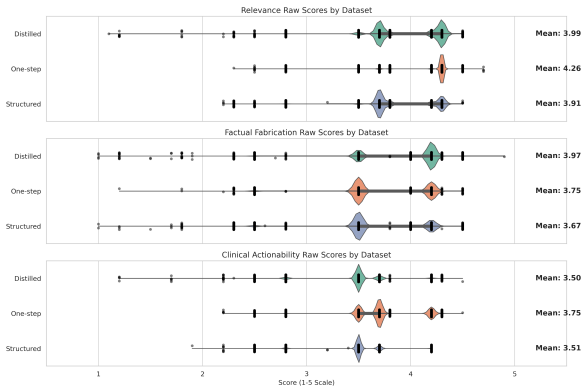

Figure 4: LLM-as-judge raw scores per summarization approach and metric.

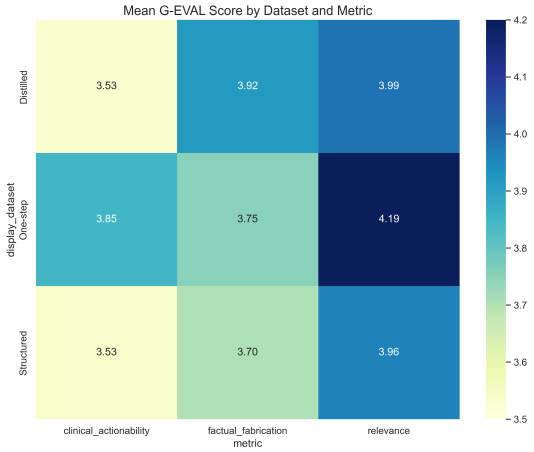

Figure 5: Heatmap of scores per summarization approach and metric.

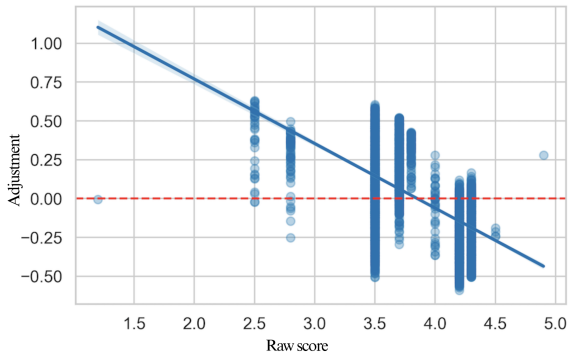

Figure 6: Score correction.

A.9 LLM-as-judge score analysis

We applied four statistical tests with Scipy and Pingouin to assess differences among summarization strategies. A one-way ANOVA (Table 7) tested for significant differences between the $S_i^{\text{one-step}}$, S_i^{struct}, and S_i^{distill}methods across the three metrics: relevance, factual fabrication, and clinical actionability. A two-way ANOVA (Table 8) examined interaction effects between method and metric. Post hoc Tukey HSD tests (Table 9) identified specific pairwise differences, and Cohen's d quantified effect sizes (Table 10) to indicate the magnitude of the differences.

Source	SS	DF	MS	F	p-value
Dataset	3820.81	2	1910.41	20529.63	<0.001
Residual	54212.89	582583	0.093	–	–

Table 7: One-way ANOVA testing for differences across summarization strategies.

Source	SS	DF	MS	F	p-value
Dataset	3820.78	2	1910.39	33623.26	<0.001
Metric	16843.69	2	8421.85	148226.31	<0.001
Interaction	4268.63	4	1067.16	18782.21	<0.001
Residual	33100.56	582577	0.057	–	–

Table 8: Two-way ANOVA testing main and interaction effects of summarization strategy and metric on scores.

A	B	Diff	SE	T	p	Hedges' g
Distilled	One-step	-0.120	0.00098	-122.69	<0.001	-0.39
Distilled	Structured	0.077	0.00098	78.31	<0.001	0.25
One-step	Structured	0.197	0.00098	201.00	<0.001	0.66

Table 9: Tukey HSD post hoc test comparing summarization strategies. All differences are statistically significant.

Comparison	Cohen's d
Distilled vs One-step	-0.39
Distilled vs Structured	0.25
One-step vs Structured	0.66

Table 10: Cohen's d overall effect sizes between summarization strategies, indicating the magnitude of performance differences.

A.10 Clinical validation

Doctors (n = 2) were recruited voluntarily. Table 11 shows their preferences between different summary methods for 18 admission notes. Values represent the number of times each method was preferred. P-values were calculated using two-sided binomial tests with Bonferroni correction.

Comparison	Metric	Preference Count			p-value	Significant
		OS	Struct	Dist		
OS vs. Struct	Relevance	8	4	–	0.388	No
	Fabrication	5	7	–	0.774	No
	Actionability	7	5	–	0.774	No
OS vs. Distill	Relevance	9	–	3	0.146	No
	Fabrication	6	–	6	1.000	No
	Actionability	9	–	3	0.146	No
Struct vs. Dist	Fabrication	–	5	7	0.774	No

Table 11: Clinician preferences between summaries. No significant differences were found across comparisons ($p > 0.05$). OS: One-step, Struct: Structured, Dist: Distilled.

A.11 Clinical summary evaluation form guidelines

Purpose: This study aims to evaluate the quality of admission note summaries generated with large language models (LLMs) using different approaches. As a clinical expert, your assessment will help determine which approach produces the best summaries across three metrics.

Task: We kindly request your assessment of 18 cases. For each case, you will be presented with:
1. An original admission note
2. Two different summaries (labeled A and B)
Your task is to compare the two summaries and indicate which one you prefer based on the metrics.

Evaluation Criteria:
- Clinical relevance: Which summary better captures and preserves the medically important information from the original note, maintaining appropriate focus on key clinical findings?
- Factual fabrication: Which summary contains fewer factual errors, made-up information, or hallucinations compared to the source note? Consider whether all statements are directly supported by the original note.
- Clinical actionability: Which summary would be more useful for clinical decision-making, providing clearer information for next steps, treatment planning, or handoffs at admission time?

Important Notes:
- The summaries are presented in random order and are unlabeled as to their source.
- There are no "right" answers - we are interested in your professional clinical judgment.
- Both summaries may have strengths and weaknesses - please select the one you consider superior.
- Try your best to choose either A or B. However, if you find the two summaries to be equal in quality based on a specific metric, mention the "tie" in the comment section.
- Please complete all 18 cases.

A.12 LLM-as-judge prompt template: Relevance

```
You are an expert in evaluating medical summaries. Your task is to assign a score for the following admission note summary based on the admission note, using the specified evaluation criteria.

IMPORTANT INSTRUCTIONS:
- This is specifically evaluating an ADMISSION NOTE SUMMARY - a concise summary of a patient's initial hospital admission record.
- Provide ONLY a single decimal number as your response, with NO explanation or additional text.
- The score MUST include a decimal point (e.g., 4.3, 3.7, 2.2). Do not use whole numbers.
- DO NOT add extra words or explanations after the score.

# INPUT (Original Admission Note)
<inputs>
[ADMISSION NOTE]
</inputs>

# OUTPUT (Admission Note Summary to be evaluated)
<output>
[SUMMARY TO BE EVALUATED]
</output>

# EVALUATION CRITERIA FOR ADMISSION NOTE SUMMARY
<evaluation_criteria>
Clinical Relevance (1.0-5.0): Measures how completely and accurately the admission note summary captures key medical facts from the original admission note. This evaluation is specifically for an admission note summary, which should capture the essential information documented during a patient's initial hospital admission.
</evaluation_criteria>

EVALUATION GUIDELINES FOR ADMISSION NOTE SUMMARIES:
- Focus on whether all CLINICALLY IMPORTANT information from the note is included
- Consider whether the information is appropriately prioritized based on clinical significance
- Assess if the summary captures the essential patient history, chief complaint, presentation, initial findings, and preliminary diagnoses
- Look for appropriate inclusion of vital signs, lab values, and test results documented at admission that influence care
- Consider whether medication information and allergies with clinical impact are included
- Do NOT overly penalize for omitting minor details that don't affect initial clinical care decisions

# SCORING RUBRIC
<scoring_rubric>
5.0: Perfect clinical relevance - captures all critical information with perfect prioritization
4.0-4.9: Very good clinical relevance - captures most critical information with good prioritization
3.0-3.9: Adequate clinical relevance - captures basic information with acceptable prioritization
2.0-2.9: Poor clinical relevance - misses important elements or prioritizes less relevant information
1.0-1.9: Very poor clinical relevance - misses most critical information
</scoring_rubric>

# CALIBRATED EXAMPLES OF ADMISSION NOTE SUMMARIES

## EXAMPLE 1: Score 5.0 (High Quality)
/synthetic_admission_note>
Example 1 (high)
</synthetic_admission_note>

<admission_note_summary>
Example 1 (high)
</admission_note_summary>

<rationale>
Example 1 (high)
</rationale>

## EXAMPLE 2: Score 3.5 (Medium Quality)
/synthetic_admission_note>
Example 2 (medium)
</synthetic_admission_note>

<admission_note_summary>
Example 2 (medium)
</admission_note_summary>
```

```
<rationale>
Example 2 (medium)
</rationale>

## EXAMPLE 3: Score 1.4 (Low Quality)
/synthetic_admission_note>
Example 3 (low)
</synthetic_admission_note>

<admission_note_summary>
Example 3 (low)
</admission_note_summary>

<rationale>
Example 3 (low)
</rationale>

# YOUR TURN - PROVIDE ONLY A SINGLE DECIMAL SCORE BELOW FOR THIS ADMISSION NOTE SUMMARY

Clinical Relevance:
```

A.13 LLM-as-judge prompt template: Factual Fabrication

```
You are an expert in evaluating medical summaries. Your task is to assign a score for the following admission note
summary based on the admission note, using the specified evaluation criteria.

IMPORTANT INSTRUCTIONS:
- This is specifically evaluating an ADMISSION NOTE SUMMARY - a concise summary of a patient's initial hospital admission
record.
- Provide ONLY a single decimal number as your response, with NO explanation or additional text.
- The score MUST include a decimal point (e.g., 4.3, 3.7, 2.2). Do not use whole numbers.
- DO NOT add extra words or explanations after the score.

# INPUT (Original Admission Note)
<inputs>
[ADMISSION NOTE]
</inputs>

# OUTPUT (Admission Note Summary to be evaluated)
<output>
[SUMMARY TO BE EVALUATED]
</output>

# EVALUATION CRITERIA FOR ADMISSION NOTE SUMMARY
<evaluation_criteria>
Factual Fabrication (1.0-5.0): Measures ONLY whether the note summary introduces facts that are completely fabricated
or invented and cannot be traced to or reasonably inferred from the original admission note. This evaluation is
specifically for an admission note summary, which should capture the essential information documented during a
patient's initial hospital admission.
</evaluation_criteria>

EVALUATION GUIDELINES FOR ADMISSION NOTE SUMMARIES:
- Focus ONLY on identifying information that is purely invented with no basis in the note
- Do NOT penalize for:
  * Reasonable clinical interpretations or conclusions based on information in the note
  * Organization or categorization of information present in the note
  * General demographic descriptions
  * Implied severity or acuity that matches clinical findings documented in the note
  * Standard medical terminology used to describe conditions mentioned in the note
- DO penalize for:
  * Adding medical conditions not mentioned in the note
  * Inventing specific test results, vital signs, or measurements not in the note
  * Creating patient history elements with no basis in the note
  * Stating specific treatments were given when not mentioned in the note
  * Making definitive statements about prognosis or outcomes not supported by the note

# SCORING RUBRIC
<scoring_rubric>
5.0: No fabrication - every statement is directly supported by or can be reasonably inferred from the note.
4.0-4.9: Minimal fabrication - contains only 1-2 minor details that might be slight overextensions but
do not contradict the note.
3.0-3.9: Some fabrication - contains a few statements that have no basis in the note but do not significantly alter
the clinical picture.
2.0-2.9: Substantial fabrication - contains multiple statements that are entirely invented with no support in the note.
1.0-1.9: Pervasive fabrication - contains critical invented information that fundamentally misrepresents the
patient's condition as documented in the note.
</scoring_rubric>

# CALIBRATED EXAMPLES OF ADMISSION NOTE SUMMARIES

## EXAMPLE 1: Score 5.0 (High Quality)
/synthetic_admission_note>
Example 1 (high)
</synthetic_admission_note>

<admission_note_summary>
Example 1 (high)
</admission_note_summary>

<rationale>
Example 1 (high)
</rationale>
```

```
## EXAMPLE 2: Score 3.8 (Medium Quality)
/synthetic_admission_note>
Example 2 (medium)
</synthetic_admission_note>

<admission_note_summary>
Example 2 (medium)
</admission_note_summary>

<rationale>
Example 2 (medium)
</rationale>

## EXAMPLE 3: Score 1.2 (Low Quality)
/synthetic_admission_note>
Example 3 (low)
</synthetic_admission_note>

<admission_note_summary>
Example 3 (low)
</admission_note_summary>

<rationale>
Example 3 (low)
</rationale>

# YOUR TURN - PROVIDE ONLY A SINGLE DECIMAL SCORE BELOW FOR THIS ADMISSION NOTE SUMMARY

Factual Fabrication:
```

A.14 LLM-as-judge prompt template: Clinical Actionability

```
You are an expert in evaluating medical summaries. Your task is to assign a score for the following admission note summary based on the admission note, using the specified evaluation criteria.

IMPORTANT INSTRUCTIONS:
- This is specifically evaluating an ADMISSION NOTE SUMMARY - a concise summary of a patient's initial hospital admission record.
- Provide ONLY a single decimal number as your response, with NO explanation or additional text.
- The score MUST include a decimal point (e.g., 4.3, 3.7, 2.2). Do not use whole numbers.
- DO NOT add extra words or explanations after the score.

# INPUT (Original Admission Note)
<inputs>
[ADMISSION NOTE]
</inputs>

# OUTPUT (Admission Note Summary to be evaluated)
<output>
[SUMMARY TO BE EVALUATED]
</output>

# EVALUATION CRITERIA FOR ADMISSION NOTE SUMMARY
<evaluation_criteria>
Clinical Actionability (1.0-5.0): Measures how clearly, concisely, and effectively the admission note summary presents urgent or decision-critical information to support clinical decision-making at the time of hospital admission. This evaluation is specifically for an admission note summary, which should capture the essential information documented during a patient's initial hospital admission.
</evaluation_criteria>

EVALUATION GUIDELINES FOR ADMISSION NOTE SUMMARIES:
- Focus on how well the summary facilitates immediate clinical decisions at the time of admission
- Consider if critical information from the note is highlighted prominently
- Assess whether the organization helps prioritize initial clinical concerns
- Look for clear presentation of abnormal findings
- Consider if medication information, allergies, and contraindications are presented
- Evaluate whether the format supports rapid understanding of the patient's status
- Consider if next steps or needed interventions are clearly implied by the information presented

# SCORING RUBRIC
<scoring_rubric>
5.0: Perfect clinical actionability - optimally presents all decision-critical information
4.0-4.9: Very good clinical actionability - presents most decision-critical information with good organization
3.0-3.9: Adequate clinical actionability - presents important information but with suboptimal organization
2.0-2.9: Poor clinical actionability - presents some information but with poor prioritization
1.0-1.9: Very poor clinical actionability - insufficient information for clinical decision-making
</scoring_rubric>

# CALIBRATED EXAMPLES OF ADMISSION NOTE SUMMARIES

## EXAMPLE 1: Score 5.0 (High Quality)
/synthetic_admission_note>
Example 1 (high)
</synthetic_admission_note>

<admission_note_summary>
Example 1 (high)
</admission_note_summary>

<rationale>
Example 1 (high)
</rationale>

## EXAMPLE 2: Score 3.4 (Medium Quality)
/synthetic_admission_note>
Example 2 (medium)
</synthetic_admission_note>

<admission_note_summary>
Example 2 (medium)
</admission_note_summary>
```

<rationale>
Example 2 (medium)
</rationale>

EXAMPLE 3: Score 1.3 (Low Quality)
/synthetic_admission_note>
Example 3 (low)
</synthetic_admission_note>

<admission_note_summary>
Example 3 (low)
</admission_note_summary>

<rationale>
Example 3 (low)
</rationale>

YOUR TURN - PROVIDE ONLY A SINGLE DECIMAL SCORE BELOW FOR THIS ADMISSION NOTE SUMMARY

Clinical Actionability: