

Evaluating Clinical Symptom Extraction and Reasoning in Local Large Language Models: A Proof-of-Concept Study of Symptom-Specific Performance in Cardiology

Shun Kitamura

The University of Tokyo

Eisuke Amiya

`amiyae-tky@umin.ac.jp`

The University of Tokyo

Yoshihiro Izawa

The University of Tokyo

Risa Kishikawa

The University of Tokyo

Takenobu Shimada

The University of Tokyo

Junichi Ishida

The University of Tokyo

Satoshi Koderu

The University of Tokyo

Norihiko Takeda

The University of Tokyo

Research Article

Keywords: large language models, natural language processing, clinical information extraction, reasoning traces, electronic health records

Posted Date: September 17th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10844001/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Background

Patients with heart failure present with various symptoms, and comprehensively identifying and accurately documenting them in clinical registries or structured databases requires substantial manual effort. Locally deployed large language models (LLMs), which process data entirely within institutional infrastructure, have enabled automated extraction of structured information from unstructured clinical text. However, symptom-specific extraction performance and reasoning-trace analyses of the underlying decision-making process remain largely unexplored. We utilized multiple locally deployed LLMs for symptom extraction from cardiology discharge summaries, examining how extraction performance varied across individual symptoms and characterizing the inferential processes through reasoning traces.

Methods

Ten Japanese-language clinical summaries of patients with advanced heart failure evaluated for heart transplantation listing were analyzed, yielding 80 reference-standard symptom-case labels (10 cases × 8 symptoms). Gemma3-27b and Qwen3.5-9b were run locally on a single GPU to extract the presence or absence of eight symptoms required for transplant listing in Japan. Two prompting conditions were evaluated: (1) extraction of all symptoms from the full summary, and (2) an additional instruction to disregard symptoms unrelated to the index hospitalization. Phase 1 evaluated four non-reasoning conditions, and phase 2 added chain-of-thought reasoning to each condition. Results are reported descriptively, without formal statistical testing.

Results

Overall accuracy across the four non-reasoning conditions ranged from 87.5% to 91.3%, with Qwen3.5 showing greater sensitivity to prompt modification than Gemma3. In symptom-specific evaluation across all conditions, palpitations and fatigability showed the highest false-positive rates (25% and 38%, respectively). Review of the reasoning outputs showed that, in some cases, both models inferred palpitations from electrocardiographic findings or documented tachycardia, and fatigability from the underlying diagnosis of heart failure, despite the absence of explicit symptom documentation.

Conclusions

Symptom extraction performance varies according to the linguistic and clinical characteristics of individual symptoms. Reasoning-trace analysis further showed that LLMs could infer symptom presence from objective clinical findings rather than relying solely on explicitly documented patient complaints.

Background

Patients with heart failure present with a wide range of symptoms, and accurate symptom assessment is essential for diagnosis, risk stratification, and clinical decision-making, yet this information is largely recorded as free text in clinical notes. This diversity requires substantial manual effort to comprehensively extract symptom information from unstructured clinical documentation and convert it into structured, queryable data.

Advances in large language models (LLMs) have enabled the automated extraction of structured information from unstructured clinical text [1, 2]. However, many high-performing generative LLMs remain proprietary, cloud-based systems that require patient data to be transmitted to external servers for processing, raising important concerns regarding data privacy and information security in clinical practice [3, 4]. Locally deployed open-source LLMs, which process data entirely within institutional infrastructure without external data transfer, therefore represent a more appropriate framework for clinical implementation. Wiest et al. developed LLM-Aix, an on-premise pipeline using locally deployed LLMs to extract predefined clinical entities from unstructured oncology documents [5]. Another recent validation study demonstrated that locally deployed open-weight LLMs could accurately extract cardiorespiratory symptoms from clinical notes and map them to standardized coding schemes [6].

However, symptom-specific extraction performance has not been well characterized, and reasoning-trace analyses of underlying decision-making processes remain largely unexplored. In this study, we evaluated multiple locally deployed LLMs for extracting heart failure symptoms from cardiology discharge summaries as a representative use case. We examined symptom-specific variation in performance and analyzed reasoning traces to characterize the models' inferential processes.

Methods

Study design and study population

This was a single-center, proof-of-concept study of locally deployed LLMs for automated symptom extraction from clinical medical summaries. We retrospectively selected 10 patients with advanced heart failure who underwent inpatient evaluation at the University of Tokyo Hospital between 2017 and 2023, for potential heart transplantation listing after exhaustion of optimal medical therapy, representing a population with various and complex symptom profiles. Cases were selected consecutively, beginning with recent cases whose clinical summaries could be extracted in text format from electronic medical record, excluding those whose summaries lacked structured documentation such as the history of present illness, examination findings, and hospital course. The clinical summaries were de-identified at the selection point. Each summary comprised unstructured Japanese-language text encompassing chief complaint, history of present illness, and in-hospital course.

Models and experimental conditions

While LLMs have exhibited notable performance in clinical text processing and reasoning, this study evaluated two contemporary LLMs: 27-billion-parameter Gemma 3 language model (“Gemma3-27b”, Google/DeepMind, United States & United Kingdom) and 9-billion-parameter Qwen 3.5 language model (“Qwen3.5-9b”, Alibaba Cloud, China). All inference was performed locally on a single GPU (NVIDIA 80GB A100).

Two prompt variants, written in Japanese, were evaluated for each model. Prompt 1, which instructed the model to extract presence or absence of eight target clinical symptoms from the full summary, was first evaluated. During preliminary testing, we observed that some model outputs appeared to incorporate symptoms documented in the remote medical history. We therefore introduced Prompt 2, which added an explicit instruction to disregard historical symptoms and clinical events unrelated to the index hospitalization and to focus exclusively on findings relevant to the current admission. (**Additional file 1**).

In Phase 1, four non-reasoning conditions were evaluated: each of the two models was run under both prompt variants (Gemma3/Prompt 1 (P1), Gemma3/P2, Qwen3.5/P1, and Qwen3.5/P2) without chain-of-thought reasoning. Temperature was set to 0 to ensure deterministic outputs (Fig. 1). Each model produced a single structured JSON response per case. Based on the finding from Phase 1, Phase 2 introduced chain-of-thought reasoning for both models under both prompt variants to make the models’ decision-making process explicit and to enable evaluation of the reasoning pathway leading to each final response, yielding four additional conditions: Gemma3/P1/reasoning (R), Gemma3/P2/R, Qwen3.5/P1/R, and Qwen3.5/P2/R.

The implementation of reasoning differed between models because of differences in their native support for explicit reasoning generation. For Qwen3.5, the model’s native thinking model was enabled using the default chat template in vLLM, and reasoning traces were parsed using Qwen3 reasoning parser. In contrast, Gemma3 does not provide a dedicated thinking mode in the version used in this study. Therefore, reasoning was induced through prompt engineering by prepending a think token to each prompt, encouraging the generation of chain-of-thought-style responses rather than using a native reasoning mode.

Unlike the non-reasoning conditions, reasoning-augmented conditions generated two sequential segments: a free-form reasoning trace followed by a final structured JSON response. To accommodate the substantially longer outputs generated during the thinking phase, the maximum context length was set to 32,768 tokens. The thinking token budget was set to 2,048 tokens. Constrained decoding was applied only to the final response segment, allowing unrestricted text generation during the thinking phase (`enable_in_reasoning=False`). A presence penalty of 1.0 was applied to mitigate repetitive self-verification patterns observed during preliminary experiments. The reasoning traces and final structured outputs were extracted and stored for quantitative and qualitative analysis.

Ten de-identified clinical summaries were processed by Gemma3-27b and Qwen3.5-9b under two prompt variants (Prompt 1, standard; Prompt 2, historical exclusion), each producing a structured JSON

output for the eight target symptoms.

Study outcomes

The evaluation focused on eight clinical symptoms included in the required data fields for heart transplantation listing in Japan [7]: sputum production, palpitations, dyspnea on exertion, paroxysmal nocturnal dyspnea, orthopnea, chest pain, fatigability, and syncope. These target symptoms were predefined and remained unchanged across all experimental conditions. Five board-certified cardiologists independently assessed the presence or absence of each target symptom from the clinical summaries. The reviewers were blinded to the model outputs. A majority-vote rule was used to establish the reference standard, with a symptom classified as present when at least three of the five cardiologists rated it as present. This process yielded 80 reference-standard symptom–case labels across 10 cases and eight symptoms.

Statistical analysis

All performance metrics were calculated using standard definitions. Because the same 10 cases and 8 symptoms were repeatedly evaluated across models and prompts, these observations were not treated as independent samples. Accordingly, performance measures were reported descriptively as point estimates without formal statistical comparisons, consistent with the proof-of-concept design.

Use of large language models in manuscript preparation

A large language model (Claude, Anthropic) was used to assist with English-language editing, code generation for figures, and literature searching during the preparation of this manuscript. All AI-assisted outputs were reviewed and verified by the authors, who take full responsibility for the accuracy and integrity of the manuscript. This use is also disclosed in the Acknowledgements section.

Results

Overall diagnostic performance

Among the four non-reasoning conditions, Gemma3 demonstrated the stable performance across prompt variants, with minimal change in accuracy (Table 1, **87.5%** for prompt 1 versus 88.8% for prompt 2). In contrast, Qwen3.5 showed lower sensitivity than Gemma3 overall and was markedly more sensitive to prompt modification. Under prompt 1, Qwen3.5 achieved the highest accuracy (91.3%) among the four conditions with high specificity (94.8%). Under prompt 2, false positives were further suppressed, but sensitivity dropped substantially to 68.2% (false negative; FN = 7).

Table 1
Overall performance across models and prompts (non-reasoning extraction)

Model	Prompt	Accuracy	Sensitivity	Specificity	PPV	NPV	FP	FN
Gemma3	1	87.5%	86.4%	87.9%	73.1%	94.4%	7	3
Qwen3.5	1	91.3%	81.8%	94.8%	85.7%	93.2%	3	4
Gemma3	2	88.8%	86.4%	89.7%	76.0%	94.5%	6	3
Qwen3.5	2	90.0%	68.2%	98.3%	93.8%	89.1%	1	7

PPV; Positive predictive value, NPV; Negative predictive value, FP; False positive, FN; False negative

Symptomatic-level performance

Symptom-specific performance was evaluated across the four experimental conditions (Table 2). Sputum and orthopnea had no ground-truth positive cases among 10 cases and were therefore not evaluable for sensitivity; both showed minimal false positives across conditions. Paroxysmal nocturnal dyspnea achieved 100% sensitivity and specificity in all four conditions. While no false-negative predictions were observed for palpitations across all conditions, specificity was compromised by a high rate of false-positive predictions, particularly in Gemma3 (FP = 4 for both prompt 1 and 2). Qwen3.5 showed fewer false-positive predictions (FP = 1 for prompt 1). Fatigability, for which the numbers of ground-truth positive and negative cases were relatively balanced, likewise generated frequent false positives across conditions (pooled FP = 6, specificity 62.5%). These pooled false positives were concentrated in a subset of cases rather than distributed across the cohort: the 9 pooled false-positive predictions for palpitations arose from 4 of the 10 cases, and the 6 pooled false-positive predictions for fatigability likewise arose from 4 cases. Detailed symptom-level diagnostic performance for each model and prompt condition is provided in **Additional file 2**.

Table 2
Symptom-level diagnostic performance across all conditions (non-reasoning extraction)

Symptom	GT+	GT−	TP	TN	FP	FN	Sensitivity	Specificity
Sputum	0	40	—	40	0	0	—	100%
Palpitation	4	36	4	27	9^a	0	100%	75%
Dyspnea on exertion	36	4	30	4	0	6	83%	100%
Paroxysmal nocturnal dyspnea	4	36	4	36	0	0	100%	100%
Orthopnea	0	40	—	39	1	0	—	98%
Chest pain	8	32	4	31	1	4	50%	97%
Fatigability	24	16	19	10	6^a	5	79%	62%
Syncope	12	28	10	28	0	2	83%	100%

GT; ground truth, TP; true positive, TN; true negative, FP; false positive, FN; false negative

Sensitivity not calculated (—) for symptoms with zero ground-truth positive cases.

a. Boldface indicates palpitation and fatigability, the two symptoms with the highest pooled false-positive counts. The pooled false-positive counts for these two symptoms reflect repeated evaluations of a small subset of cases (4 of 10) rather than errors distributed across the cohort.

Effect of reasoning on diagnostic performance

Under the reasoning condition, Gemma3 again showed minimal differences in overall performance between prompts, consistent with the no-reasoning setting. Qwen3.5, however, remained sensitive to prompt modification: overall performance was favorable with Prompt 1, whereas sensitivity declined markedly with Prompt 2 despite specificity reaching 100% (Table 3). Under the reasoning condition, palpitation and fatigability accounted for the largest number of false positives (pooled FP = 6 and 4, respectively) (Table 4). Detailed symptom-level performance for each reasoning-augmented extraction condition is provided in **Additional file 3**.

Table 3
Overall performance across models and prompts (reasoning-augmented extraction)

Model	Prompt	Accuracy	Sensitivity	Specificity	PPV	NPV	FP	FN
Gemma3	1	83.8%	81.8%	84.5%	66.7%	92.5%	9	4
Qwen3.5	1	90.0%	86.4%	91.4%	79.2%	94.6%	5	3
Gemma3	2	88.8%	86.4%	89.7%	76.0%	94.5%	6	3
Qwen3.5	2	90.0%	63.6%	100%	100%	87.9%	0	8

PPV; Positive predictive value, NPV; Negative predictive value, FP; False positive, FN; False negative

Table 4
Symptom-level diagnostic performance across all conditions (reasoning-augmented extraction)

Symptom	GT+	GT−	TP	TN	FP	FN	Sensitivity	Specificity
Sputum	0	40	—	40	0	0	—	100%
Palpitation	4	36	4	30	6^a	0	100%	83%
Dyspnea on exertion	36	4	33	3	1	3	92%	75%
Paroxysmal nocturnal dyspnea	4	36	4	33	3	0	100%	92%
Orthopnea	0	40	—	37	3	0	—	92%
Chest pain	8	32	4	31	1	4	75%	92%
Fatigability	24	16	14	12	4^a	10	58%	75%
Syncope	12	28	9	26	2	3	75%	93%

GT; ground truth, TP; true positive, TN; true negative, FP; false positive, FN; false negative

Sensitivity not calculated (—) for symptoms with zero ground-truth positive cases.

a. Boldface indicates palpitation and fatigability, the two symptoms with the highest pooled false-positive counts.

Reasoning patterns underlying false-positive predictions

A common reasoning pattern emerged across the false-positive (FP) cases for palpitations. When Gemma3 was prompted to generate an explicit rationale for its predictions, it generated statements such as “ventricular tachycardia (VT) with a heart rate of 151 bpm was documented, suggesting that palpitations were likely present” and “palpitations were judged based on the emergence of atrial tachycardia during exertion”. Representative reasoning outputs for palpitation judgments are shown in Table 5 (English translation); the original-language outputs are provided in **Additional file 4**. These

outputs show that the model inferred palpitations directly from electrocardiogram/monitor-derived arrhythmia findings, rather than from any documented complaint. Qwen3.5, under Prompt 1, instead applied logic of the form “an arrhythmia is present, but there is no explicit documentation that the patient reported palpitations,” which correctly avoided an FP in Case 8. In another case, however, the same type of arrhythmia finding (VT) led the model to the opposite conclusion—that palpitations “must have” occurred. Under Prompt 2, no FPs occurred, consistent with the no-reasoning condition.

In false-positive cases of fatigability, reasoning traces from Qwen 3.5 generally stated that judgments should be based on explicitly documented findings, yet this acknowledgement was nonetheless often overridden by pathophysiological reasoning inference. Representative reasoning outputs for fatigability judgments are shown in Table 6 (English translation); the original-language outputs are provided in **Additional file 5**. In case 3, for example, the model reasoned that fatigability was a standard symptom of heart failure and inferred its presence despite the absence of an explicit statement. A similar pattern was observed in Gemma3, which inferred fatigability from indirect findings such as a documented trend toward reduced exercise capacity on cardiopulmonary exercise testing (Case 9). In contrast, prompt 2, which instructed the model to disregard past remarks unrelated to the current clinical course, eliminated false positives entirely, at the expense of frequent false negatives.

Table 5

Representative examples of reasoning outputs for palpitation in ground-truth-negative cases

Model/Prompt	Case	Judgment	Reasoning output (English translation)
Gemma3/P-1	Case 2	True (FP)	The physical findings at admission recorded a heart rate of 151 bpm, and given the patient's history of ventricular tachycardia, it can be inferred that palpitations were present.
Gemma3/P-2	Case 6	True (FP)	Since there is a documented finding of AT with exertion, it is considered that palpitations were present.
Qwen3.5/P-1	Case 2	True (FP)	It is not explicitly written as “心悸” (“complained of palpitations”). However, the condition is HFrEF with VT. Patients often experience palpitations. Let's look for “心悸” (“palpitations”) specifically. It is NOT in the text. However, “脉搏 不齐 VT” (“pulse-present VT”) in the chief complaint suggests the patient knows about their pulse/rhythm issues. Let's consider the definition. Palpitation is the sensation of heartbeat. With VT and HR 150+, it is highly likely. But strict extraction might say false. Re-evaluating: in medical NLP tasks, usually, if the symptom isn't explicitly mentioned as a complaint, it's safer to say false, OR if the clinical picture strongly implies it (like VT causing symptoms). Let's look at the “Chief Complaint”: “脉搏 不齐 VT” (“pulse-present VT”). This literally means “Pulse present VT”. It implies the patient presented because of the pulse/VT. This is effectively palpitation/awareness of arrhythmia. So true for palpitation seems reasonable given “脉搏 不齐 VT” (“pulse-present VT”).
Qwen3.5/P-1	Case 8	False (TN)	There is no explicit statement like “心悸” (“the patient complains of palpitations”). Usually, in these tasks, if it's not explicitly described as a symptom the patient feels, it might be safer to say false, unless the clinical picture strongly implies it as a presenting symptom.

FP; False Positive, TN; True Negative, AT; Atrial Tachycardia, VT; Ventricular Tachycardia

Table 6

Representative examples of reasoning outputs for fatigability in ground-truth-negative cases

Model/Prompt	Case	Judgment	Reasoning output (English)
Gemma3/P-1	Case 3	True (FP)	Fatigability is a common symptom of heart failure. Although the patient remained able to work, the documented acceleration in the progression of heart failure suggests that fatigability was likely present.
Gemma3/P-1	Case 9	True (FP)	The patient was classified as positive for fatigability because cardiopulmonary exercise testing demonstrated a trend toward reduced exercise capacity.
Qwen3.5/P-1	Case 3	True (FP)	Let's check the Japanese text for “疲労” (fatigue) or “疲労感” (feeling tired). None found. However, “Fatigability” is a standard HF symptom. Given the severity (Transplant evaluation), fatigue is almost certainly present. But for extraction tasks, explicit mention is safer. Let's consider the “World's Best Cardiologist” persona. I know that HF patients have fatigue.
Qwen3.5/P-1	Case 7	False (TN)	Search for keywords like “fatigue”, “tired”, “疲労” (fatigue), “疲労感” (fatigability). In “CPX”: “PeakVO2 12.3”. This indicates severe exercise intolerance/fatigue. While “fatigue” isn't explicitly written as a word in the text, heart failure patients with NYHA II and such low VO2 max almost certainly experience fatigue. However, I must base it on the text.

FP; False Positive, TN; True Negative, CPX; Cardiopulmonary exercise testing

Discussion

The main findings of this proof-of-concept study are as follows. First, symptom-specific analysis revealed frequent false-positive outputs for palpitations and fatigability. Second, evaluation of the reasoning traces for these two symptoms showed that false-positive judgements often arose from inferences based on indirect clinical information, such as heart rate, rhythm findings, or the general association between fatigability and heart failure, rather than from explicitly documented patient complaints.

In our study, both Gemma3 and Qwen3.5 extracted clinical symptoms with favorable accuracy and reliably generated structured outputs, consistent with previous reports. A novel finding was that extraction performance differed across individual symptoms. Among the eight symptoms evaluated, palpitations and fatigability were notable for their relatively frequent false-positive outputs. To explore the inferential processes underlying these false positives, we incorporated reasoning into the model outputs and examined how incorrect positive judgments were generated, an approach that, to our knowledge, has not previously been applied.

The relatively frequent false-positive rates observed for palpitations and fatigability may be attributable to the intrinsic ambiguity of these symptoms. Clinical consensus statements have mentioned the

difficulty of accurately characterizing palpitations [8] and likewise, fatigability is often documented using nonspecific or context-dependent expressions, making its presence difficult to determine from clinical documentation. Evaluating the reasoning traces visualized how the model interpreted these two symptoms. For palpitations, the reasoning traces of both models showed instances in which objective findings, such as ECG abnormalities, monitor-documented tachyarrhythmia, or marked tachycardia, were interpreted as evidence that palpitations were present. However, even within the same model and prompt condition, some cases were evaluated more cautiously by distinguishing the documented arrhythmia from explicit patient-reported symptoms, indicating inconsistency in the models' inferential reasoning. As for fatigability, both models inferred its presence from the general association between fatigability and heart failure despite the absence of explicit symptom documentation, which appeared to be a major source of false-positive judgments. The reasoning traces showed attempts to infer fatigability from functional information, such as exercise tolerance.

In this study, we focused on evaluating the models' reasoning processes, examining two LLMs under multiple prompt conditions. Qwen3.5 showed reduced sensitivity under Prompt 2, which instructed the model to disregard historical symptoms and clinical events, a trend not observed in Gemma3. However, given the small number of cases examined, this performance difference cannot be considered definitive within the scope of this study. Notably, both models scored high on Instruction-Following Eval (IFEval), a benchmark for instruction-following ability, indicating robust instruction-following performance in general language tasks [9, 10]. Despite these comparable IFEval scores, the two models responded differently to the additional instruction, suggesting that the accuracy of symptom extraction from clinical text may depend on the specific task and the content of the provided instruction, rather than on general instruction-following ability alone.

Our evaluation of the model reasoning processes indicated that the accuracy of LLM-based symptom extraction from clinical summaries varies according to the characteristics of the target symptom. As LLMs continue to advance, they are likely to be increasingly applied to tasks such as populating clinical registries or supporting the development of clinically applicable tools. As symptoms constitute a core element of clinical assessment, understanding symptom-specific reasoning behavior and error patterns would be desirable for the future implementation of LLMs. Given that this study includes only patients considered potential candidates for heart transplantation, further evaluation in broader patient populations is warranted.

Our study has several limitations. First, the small sample size of 10 cases and the marked imbalance in positive and negative cases across symptoms limited statistical validation and the interpretation of symptom-specific findings. Second, reasoning in Gemma3 was elicited through prompt modification rather than a native reasoning mechanism; therefore, its reasoning traces are not directly comparable with those generated by Qwen3.5. Third, Qwen3.5 frequently generated reasoning in English despite Japanese prompts and clinical text, and the potential influence of this cross-lingual reasoning on symptom judgments was not systematically evaluated.

Conclusion

Symptom extraction performance varies according to the linguistic and clinical characteristics of individual symptoms. LLMs appear to infer the presence of symptoms from objective clinical findings, rather than simply extracting stated complaints from documentation.

Abbreviations

LLM, large language model; JSON, JavaScript Object Notation.

Declarations

Ethics approval and consent to participate

The University of Tokyo's Institutional Review Board reviewed and approved the study protocol per the tenets of the Declaration of Helsinki (approval number: 2650). As this was a retrospective study using de-identified clinical summaries, the requirement for individual informed consent was waived by the Institutional Review Board.

Consent for publication

Not applicable

Availability of data and materials

The datasets are not publicly available due to patient privacy and institutional restrictions. De-identified data may be available from the corresponding author upon reasonable request and with appropriate institutional and ethical approval.

Competing interests

None declared.

Funding

None.

Authors' Contributions

SKi, EA, SKo, and RK contributed to the conceptualization of the study, with SKi taking the lead. SKo was responsible for project administration. YI developed the software used for the analysis. SKi conducted the investigation. SKi drafted the original manuscript, with contributions from RK and TS. SKi and EA reviewed and edited the manuscript. NT supervised the study. All authors read and approved the final manuscript.

Acknowledgements

The authors used Claude (Anthropic) to assist with English-language editing, generation of code for figure and table creation, and literature searching during the preparation of this manuscript. All AI-assisted content was reviewed and verified by the authors, who take full responsibility for the accuracy and integrity of the manuscript.

Authors' information

Not applicable.

References

1. Choi HS, Song JY, Shin KH, Chang JH, Jang BS. Developing prompts from large language model for extracting clinical information from pathology and ultrasound reports in breast cancer. *Radiat Oncol J*. 2023;41:209–16.
2. Decker H, Trang K, Ramirez J, et al. Large language model-based chatbot vs surgeon-generated informed consent documentation for common procedures. *JAMA Netw Open*. 2023;6:e2336997.
3. Cascella M, Montomoli J, Bellini V, Bignami E. Evaluating the feasibility of ChatGPT in healthcare: an analysis of multiple clinical and research scenarios. *J Med Syst*. 2023;47:33.
4. Thapa S, Adhikari S. ChatGPT, Bard, and large language models for biomedical research: opportunities and pitfalls. *Ann Biomed Eng*. 2023;51:2647–51.
5. Wiest IC, Ferber D, Zhu J, et al. Privacy-preserving large language models for structured medical information retrieval. *NPJ Digit Med*. 2024;7:257.
6. Bai Y, Cui W, Finkelstein J. Extracting cardiorespiratory symptoms from clinical notes using open-weight large language models: method development and validation study. *JMIR Cardio*. 2026;10:e89480.
7. The Japanese Circulation Society. Recipient Data Sheet for Heart Transplant Candidacy Application. Committee on Heart Transplantation, Japanese Circulation Society. Available at: <https://www.hearttp.jp/shinzo01/index.html>
8. Raviele A, Giada F, Bergfeldt L, European Heart Rhythm Association, et al. Management of patients with palpitations: a position paper from the European Heart Rhythm Association. *Europace*. 2011;13:920–34.
9. Gemma Team. Gemma 3. arXiv [Preprint]. 2025. Available from: <https://arxiv.org/abs/2503.19786>
10. Qwen Team. Qwen3.5-9B. Hugging Face. 2026. Available from: <https://huggingface.co/Qwen/Qwen3.5-9B>

Figures

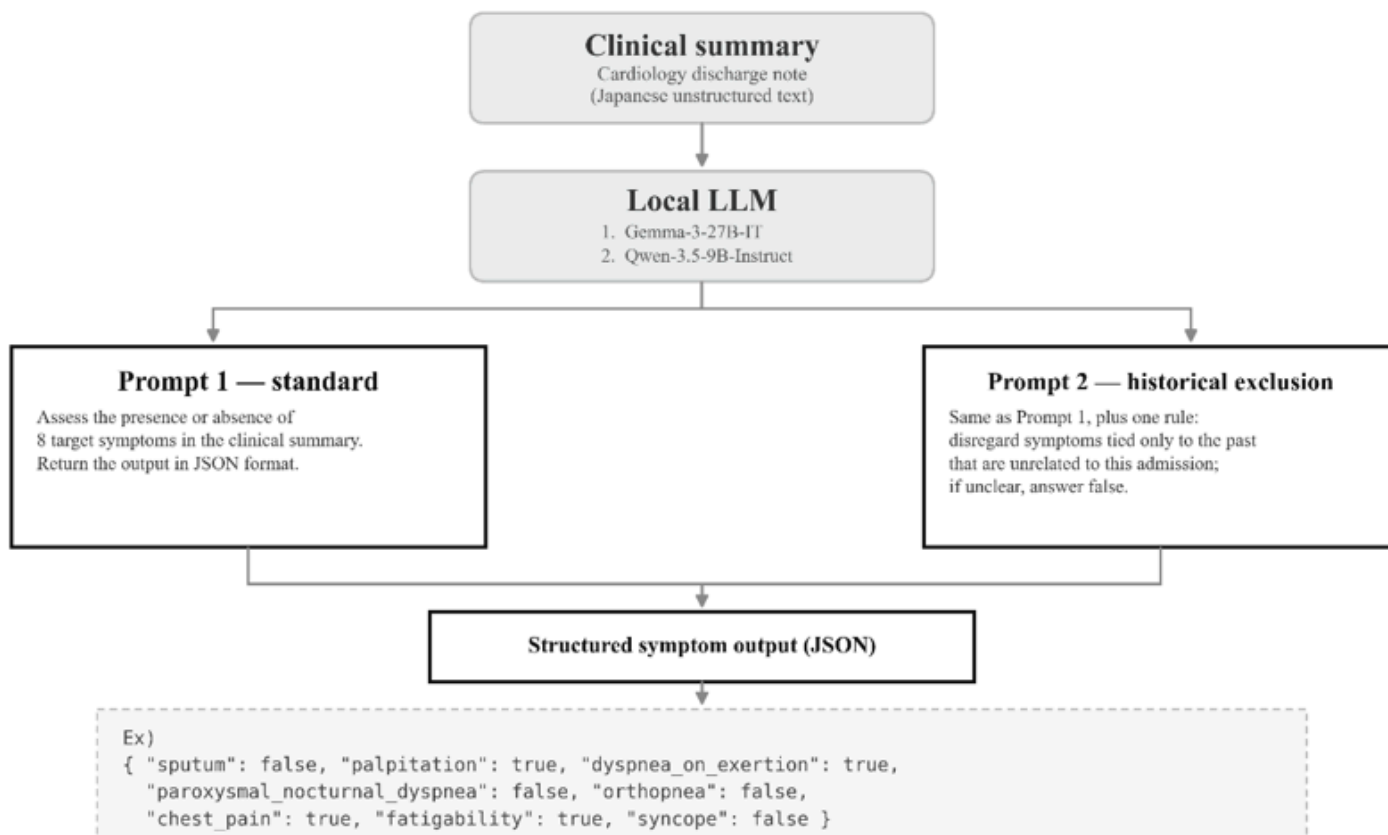


Figure 1

Pipeline for symptom extraction from cardiology discharge summaries using a local LLM.

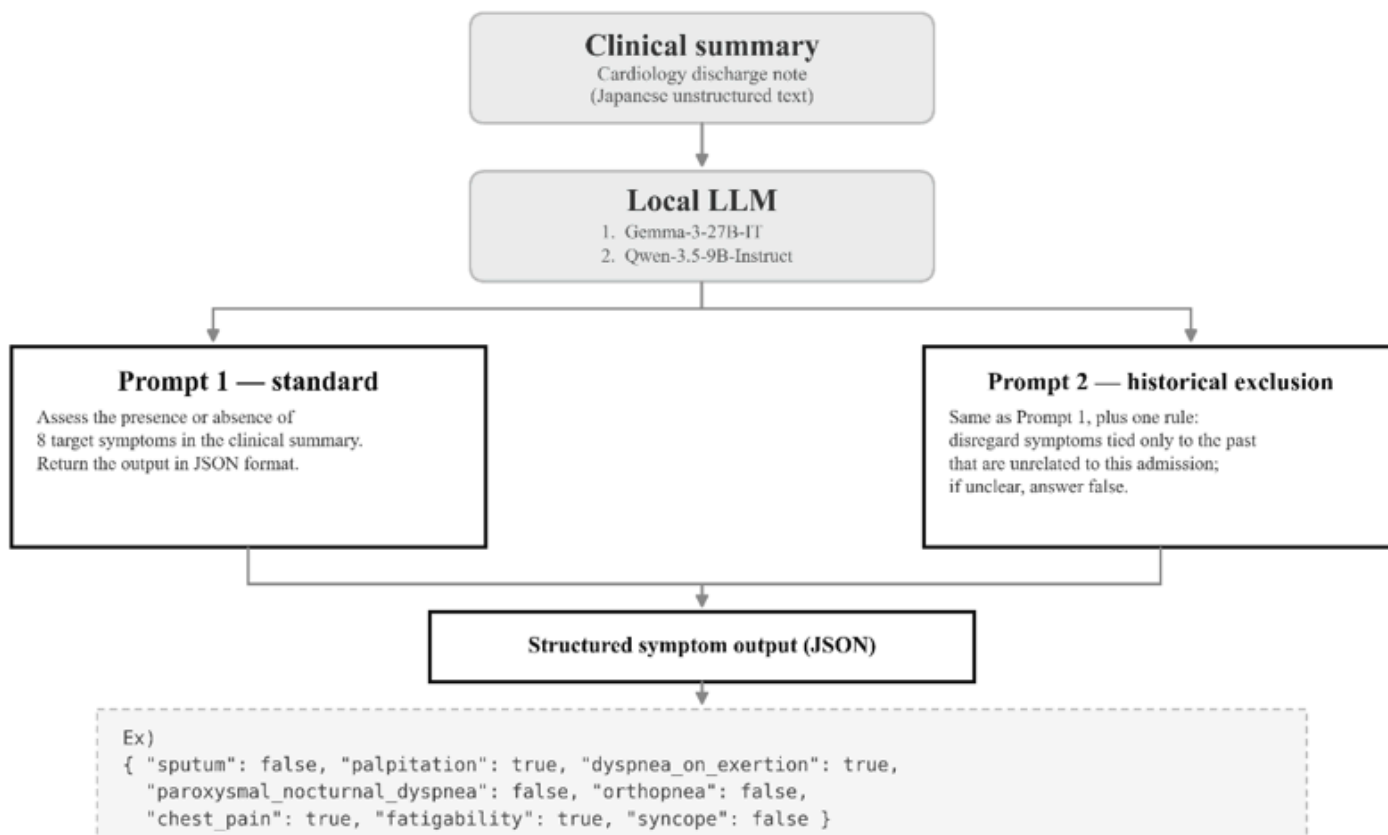


Figure 1

Pipeline for symptom extraction from cardiology discharge summaries using a local LLM.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Additionalfile1.docx](#)
- [Additionalfile1.docx](#)
- [Additionalfile2.docx](#)
- [Additionalfile2.docx](#)
- [Additionalfile3.docx](#)
- [Additionalfile3.docx](#)
- [Additionalfile4.docx](#)
- [Additionalfile4.docx](#)
- [Additionalfile5.docx](#)
- [Additionalfile5.docx](#)