

Uncertainty Meets Conformance: A Process Mining-based Multi-Objective Evaluation for Clinical Reasoning using LLMs

Seon Kim

Yonsei University
kso323@yonsei.ac.kr

Tae Hoon Kong*

Yonsei University
cochlear84@yonsei.ac.kr

Jeong-woo Lee

Yonsei University
jeongwoolee@yonsei.ac.kr

Jongchan Kim*

Yonsei University
jongchan.kim@yonsei.ac.kr

Abstract

We present a framework for evaluating LLM-based clinical reasoning using conformance metrics from process mining. Four local LLMs were tested on EMR data from 16 vertigo patients across four prompting methods, with five repetitions per combination, yielding six uncertainty metrics. Reasoning traces were converted into event logs and analyzed with Heuristic Miner to derive external conformance to other patients' reasoning patterns and internal conformance across repeated runs. Treating these as joint objectives, Pareto, Hypervolume, and TOPSIS analyses identified distinct evaluation scenarios, showing that conformance metrics can complement outcome-based uncertainty by revealing how LLMs reach their conclusions.

1 Introduction

Large language models for medical question answering and clinical decision support have been actively studied, but most evaluations emphasize output-level accuracy (Bedi et al., 2025), revealing little about how a model reaches its diagnosis, which is essential for clinical diagnosis. Uncertainty measures like Semantic entropy (Farquhar et al., 2024) capture output variation but provide no information about the reasoning process itself. Reasoning traces offer additional insight, though they should be treated as visible outputs rather than complete records of internal computation (Turpin et al., 2023).

Process mining discovers various patterns from event sequences and assesses trace conformance to discovered patterns (Van der Aalst et al., 2012). We encode clinical LLM reasoning as event sequences and apply process mining to assess conformance to reference patterns and stability across repeated inferences. We combine these with uncertainty in a multi-objective analysis using Pareto, Hypervolume, and TOPSIS to examine whether this process-

level view reveals information beyond outcomes or uncertainty alone.

2 Related Work

Recent works on LLM evaluation has focused on uncertainty quantification. Semantic entropy assesses how much the meaning of repeated answers disagrees, while sample consistency (Savage et al., 2025) checks how often the model repeats the same answer.

Beyond this answer-level view, other work has begun to look more closely at the reasoning itself. Da et al. (2025) represented model explanations as logical graphs to see how much their structure varies, Perez et al. (2025) built a benchmark to test how well LLMs handle statistical uncertainty quantification tasks, de Villiers et al. (2025) compared LLM judgments under uncertain evidence against normative approaches, and Deng et al. (2026) looked at how LLMs deal with uncertainty during multi-step decision making.

Process mining has been applied to healthcare workflows to identify clinical pathways (Xu et al., 2020). Recently, LLM reasoning steps were classified with reasoning type and effect on the final LLM output (Berti et al., 2025). Building on this work, we combine output-level uncertainty with process-based uncertainty via external and internal conformance of diagnostic reasoning.

3 Methodology

3.1 Framework Overview

The overall framework is illustrated in Figure 1. Four cases were randomly selected from each of the four diagnostic categories, yielding 16 cases in total. Four local LLMs, 16 patients with vertigo, and four prompting methods were used in combination, with five repetitions per combination, making 1,280 diagnostic runs, each producing a

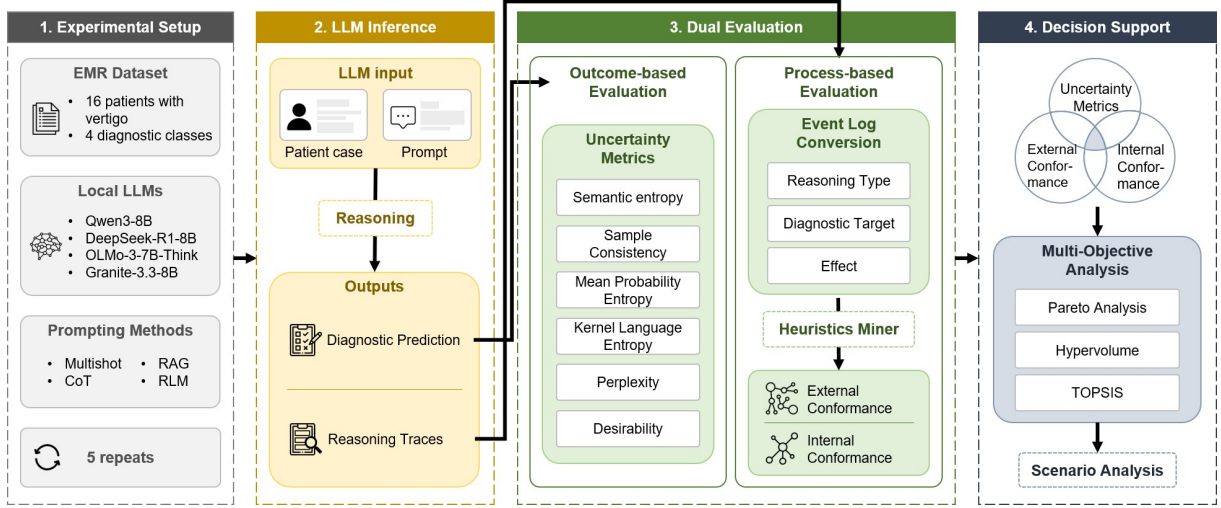


Figure 1: A Process Mining-based Multi-Objective Evaluation Framework for Clinical Reasoning using LLMs

diagnosis and a reasoning trace. Repeated predictions yielded uncertainty metrics, while reasoning traces were converted into event logs for process mining. Using event logs, external conformance, which measures agreement with reasoning patterns from patients sharing the same diagnosis, and internal conformance which estimates consistency across repeated runs of the same case were calculated. These three criteria were jointly examined via Pareto, Hypervolume, and TOPSIS analyses, followed by scenario analysis.

An individual run was the unit of LLM inference, whereas a patient-model-prompt combination was the unit of multi-objective analysis. The five repeated runs for each combination were used to derive a predicted diagnosis and uncertainty-related metrics. External and internal conformance measurements were averaged over five repeated runs to obtain combination-level process measures.

3.2 Uncertainty Measurement

Six metrics were calculated from five repeated predictions. Semantic entropy captured uncertainty in the diagnostic-label distribution, and sample consistency measured agreement with the most frequent diagnosis. Mean probability entropy and probability perplexity captured class-probability uncertainty on linear and exponential scales, respectively, while kernel language entropy assessed diversity in reasoning traces based on semantic similarity. Desirability combined certainty with correctness, assigning positive certainty to correct predictions and negative certainty to incorrect ones (Wang et al., 2025), calculated as follows:

$$D = \begin{cases} C, & \hat{y} = y, \\ -C, & \hat{y} \neq y. \end{cases} \quad (1)$$

where $C \in [0, 1]$ denotes the certainty derived from the probability perplexity of the mean class-probability distribution across repeated runs (five repetitions in this experiment), $\hat{y}$ denotes the diagnosis predicted by the model, and y denotes the true diagnosis. Accordingly, positivity and negativity of D indicate correct and incorrect predictions, respectively, while $|D|$ represents the degree of certainty. Probability perplexity and certainty were calculated as follows:

$$PP = \exp \left(- \sum_{k=1}^K \bar{p}_k \log \bar{p}_k \right). \quad (2)$$

$$C = \frac{K - PP}{K - 1}. \quad (3)$$

where $\bar{p}_k$ is the average probability for diagnostic category k across repeated runs, and $K = 4$ is the total number of diagnostic categories. A larger PP reflects greater uncertainty, while C reverses and normalizes PP so that a larger value represents greater certainty. As the sign of Desirability is determined using the true diagnosis, it is a retrospective, correctness-aware measure rather than an outcome-independent uncertainty measure. Accordingly, Desirability was analyzed jointly with external and internal conformance to characterize how correctness-aware uncertainty relates to the observed reasoning process. The conformance measures provided process-level information, including whether an incorrect diagnosis was produced

through a stable or reference-aligned reasoning pattern.

3.3 Process Mining Analysis

Reasoning traces were converted into event logs recording reasoning type, diagnostic target, and effect on that target. Following [Berti et al. \(2025\)](#), Gemma-3-12B was used as the judge model to label individual reasoning steps, with the predicted diagnosis recorded as a terminal event. Taxonomy and label definitions are described in Appendix A.

Heuristics Miner ([Weijters et al., 2006](#)) was used to discover main reasoning patterns, as it identifies dependency relations from event frequencies while filtering infrequent behavior, making it suitable for noisy logs. Conformance was calculated using token-based replay ([Van der Aalst et al., 2012](#)), which compares an observed trace against a reference model by counting missing or extra steps needed to align them. External conformance measured similarity to patterns from correct responses of other patients sharing the same diagnosis, model, and prompting method, while internal conformance measured similarity among the five repeated runs of the same condition. In this study, Heuristics Miner was implemented in PM4Py using the default parameter settings.

3.4 Scenario Analysis

For each LLM, the six uncertainty metrics were evaluated separately. Each metric served as the uncertainty objective alongside external and internal conformance, producing six three-objective evaluations per model. Pareto-set accuracy, Hypervolume, and TOPSIS-based results were calculated for each evaluation. Based on this comparison, Desirability was selected for the subsequent model-specific results, Pareto visualizations, reasoning scenarios, and case study. The other five metrics were included in the metric comparison but were not carried forward to subsequent analyses. These metrics will be incorporated into an extended version of this work.

For the multi-objective analysis, patient-model-prompt combinations were compared using uncertainty metrics, external conformance, and internal conformance. Pareto analysis identified combinations that were not dominated across all three criteria. Hypervolume ([Zitzler and Thiele, 1999](#)) measured the quality of this set, with larger values indicating more favorable uncertainty and conformance. TOPSIS ([Hwang and Yoon, 2012](#))

ranked individual combinations according to their closeness to the ideal condition.

More specifically, Pareto dominance was evaluated using the calculated metric values without additional normalization and according to the specified optimization direction of each objective. In this paper, Desirability was maximized, together with external and internal conformance. In an extended version of this work, the same evaluation framework will be applied to the remaining uncertainty metrics, with sample consistency maximized and semantic entropy, mean probability entropy, probability perplexity, and kernel language entropy minimized.

For Hypervolume and TOPSIS, each objective was converted within each LLM to a scale in $[0, 1]$ using min-max normalization, with the scale reversed for objectives to be minimized. Thus, Desirability was used on its original scale for Pareto analysis and normalized for Hypervolume and TOPSIS. If all observed values of an objective are equal, its normalized value is set to 0.5. Hypervolume was calculated relative to the reference point $(0, 0, 0)$.

TOPSIS assigned an equal weight of $1/3$ to each objective because there was no empirical or clinical basis for prioritizing one objective over another. The TOPSIS score was calculated as $d_i^- / (d_i^+ + d_i^-)$, where d_i^+ and d_i^- are the Euclidean distances from the weighted ideal point $(1/3, 1/3, 1/3)$ and anti-ideal point $(0, 0, 0)$, respectively. Combinations with higher scores received better ranks.

For scenario analysis, Desirability, external conformance, and internal conformance were classified as high or low using their model-specific median values. The binary combinations of the three measures produced eight reasoning scenarios (S1–S8).

3.5 Clinical review of reasoning traces

To complement the computational conformance analysis, the reasoning traces of the representative cases included in the case study were independently reviewed by an otologist with expertise in vestibular disorders. The review focused on whether the reasoning pathway was clinically plausible, whether relevant clinical findings were appropriately interpreted, and whether the progression from the available clinical information to the diagnostic conclusion was broadly consistent with clinically recognizable diagnostic reasoning pathways. The assessment focused on the clinical coherence of the reasoning process rather than agreement with the final diagnostic label alone.

Patient / Model / Prompt	True → Predicted	D	E	I	TOPSIS	Rank	Scenario
0024 / Qwen3 / RAG	BPPV → BPPV	0.793	0.731	1.000	0.876	1	S1
0011 / DeepSeek / RLM	BPPV → BPPV	0.346	0.886	0.991	0.818	1	S1
0045 / OLMo / Multishot	MENIERE → MENIERE	1.000	0.877	0.934	0.774	5	S1
0043 / Granite / CoT	OTHERS → OTHERS	0.534	0.854	0.974	0.833	1	S1
0028 / DeepSeek / RLM	OTHERS → BPPV	-0.394	0.866	1.000	0.672	13	S5
0008 / OLMo / Multishot	VN → OTHERS	-0.492	0.774	0.955	0.585	30	S5

Table 1: Representative cases selected using Desirability(D), external conformance(E), and internal conformance(I).

4 Experiments

4.1 Dataset

Source data consisted of EMRs from 1,263 patients presenting with dizziness at Wonju Severance Christian Hospital, South Korea. As this study is part of an ongoing work, only a subset of this pool has been processed. For this initial analysis, four patients were randomly selected from each of four diagnostic categories (BPPV, vestibular neuritis, Ménière’s disease, and other diagnoses), yielding a balanced set of 16 cases with structured clinical information including symptoms, examination findings, and medical history.

The study was approved by the Institutional Review Board of Wonju Severance Christian Hospital (Approval No. CR323027), with informed consent waived because of its retrospective design. Patient data were anonymized before analysis, and access was restricted to authorized study personnel. All LLM inference was performed locally without transmitting clinical data to external services, and no personally identifiable information was included in the model inputs, reasoning traces, or reported materials.

4.2 Models

Four open-source LLMs of comparable size (approximately 7–8B parameters) were evaluated: Qwen3-8B, DeepSeek-R1-Distill-Llama-8B (Q4_K_M), OLMo-3-7B-Think, and Granite-3.3-8B. Similar parameter counts were chosen to minimize scale-driven performance differences, enabling a fair comparison of reasoning behavior and prompting strategies.

4.3 Prompting Methods

We compared four prompting strategies. Multishot prompting provided 12 example cases (three per diagnostic category) with diagnoses before the target case. Chain-of-Thought (CoT) guided the model through a predefined clinical reasoning process based on (Newman-Toker and Edlow, 2015)’s

diagnostic workflow. Retrieval-Augmented Generation (RAG) incorporated information retrieved from a knowledge base built from the Bárány Society diagnostic criteria¹, which is used as the gold standard for vertigo diagnosis. Recursive Language Modeling (RLM, (Zhang et al., 2025)) guided the model through iterative four reasoning steps, recursively refining intermediate reasoning before the final prediction. Representative prompt structures we used in this study are provided in Appendix C.

5 Case study

Table 1 presents four desirable cases and two counterexamples using Desirability (D), external conformance (E), and internal conformance (I), equally weighted in TOPSIS. The four correct cases show these criteria jointly shaping rank rather than any single value. The two incorrect cases show that conformance alone does not guarantee correctness. Pareto sets captured varying trade-offs across models, where three-dimensional visualizations for all four LLMs are described in Appendix B.

Qualitative clinician review of the representative reasoning traces showed that the models generally followed clinically recognizable diagnostic pathways. The traces identified relevant clinical features, considered plausible diagnostic alternatives, and connected the available evidence to the diagnostic conclusion in a manner broadly consistent with clinical reasoning. No major systematic deviation from clinically plausible reasoning was identified among the reviewed representative cases.

6 Results

In Table 2, pareto accuracy refers to the proportion of correct predictions among Pareto-optimal cases, Hypervolume quantifies the criterion space covered by that Pareto-optimal set, TOPSIS top accuracy indicates whether the highest-ranked case was correct, and TOPSIS AUC measures how well the

¹<https://www.thebaranysociety.org/icvd-consensus-documents/>

Model	Pareto accuracy	Hypervolume	TOPSIS top accuracy	TOPSIS AUC
Qwen3-8B	100.0%	0.8768	100.0%	0.9798
DeepSeek-R1-Distill-Llama-8B	50.0%	0.8408	100.0%	0.8606
OLMo-3-7B-Think	80.0%	0.7686	100.0%	0.9933
Granite-3.3-8B	100.0%	0.8650	100.0%	0.9323

Table 2: Comparison of model performance using Pareto accuracy, hypervolume, and TOPSIS-based evaluation.

Scenario	D	E	I	Interpretation
S1	High	High	High	Correct diagnosis, reasoned consistently with similar patients, and reproducible across repeats
S2	High	High	Low	Correct diagnosis, reasoned consistently with similar patients, though the reasoning path varies across repeats
S3	High	Low	High	Correct diagnosis, reasoned through a distinct but reproducible pathway
S4	High	Low	Low	Correct diagnosis, without a consistent or reference-aligned reasoning pathway
S5	Low	High	High	Incorrect diagnosis, reasoned consistently with similar patients and reproducible across repeats
S6	Low	High	Low	Incorrect diagnosis, partially resembling similar patients’ reasoning, but not reproducible
S7	Low	Low	High	Incorrect diagnosis, reasoned through a distinct and reproducible pathway
S8	Low	Low	Low	Incorrect diagnosis, without a consistent or reference-aligned reasoning pathway

Table 3: Interpretation of the eight reasoning scenarios based on desirability(D), external conformance(E), and internal conformance(I).

resulting ranking separated correct from incorrect predictions.

Performance was generally strong and consistent across models, with TOPSIS AUC ranging from 0.86 to 0.99. Qwen3-8B and Granite-3.3-8B achieved 100% Pareto accuracy, whereas DeepSeek-R1-Distill-Llama-8B showed the lowest accuracy (50%), suggesting that in this model, high external conformance or internal conformance could locate incorrect predictions within the Pareto-optimal set.

For interpretability, Desirability, external conformance, and internal conformance were each split into high and low levels using model-specific median thresholds, accounting for differences in metric distributions across LLMs. Combining these levels yielded eight qualitative reasoning scenarios (Table 3).

7 Conclusion

This study introduces a process mining-based framework that combines predictive uncertainty with external and internal conformance of reasoning traces, enabling clinical LLM evaluation to move beyond diagnostic accuracy to how models actually reason toward their conclusions. By treating uncertainty, external conformance, and internal conformance as joint objectives, the framework captures a range of reasoning behaviors that neither accuracy nor uncertainty alone can explain. Case studies revealed that the framework can identify

stable but incorrect reasoning patterns which cannot be detected by accuracy or uncertainty metrics alone. This is extremely critical in clinical settings, where a confident, well-reasoned but wrong diagnosis can be more dangerous. We expect process mining-based conformance analysis to be a valuable addition to how these models are evaluated, giving clinicians and scientists a clearer view into how LLM-based diagnostic reasoning actually plays out, and helping these models find a safer place in clinical workflows.

Limitations

As this study is ongoing work, the main limitation at current step is the small sample size. This study included only 16 patients and four locally deployed LLMs, which limits the statistical power and generalizability of the findings. Although each diagnostic group was represented by four cases, the dataset may not adequately capture the clinical heterogeneity of vertigo presentations. As the original dataset consists of 1,263 patients, we are planning to include more number of patients in our future work.

A direction for future work is looking at how diagnostic difficulty differs across vertigo subtypes. Some conditions are simply harder to reason about than others, so higher uncertainty there may just reflect that difficulty rather than a weakness in the model. Future work should take this into account, considering each diagnosis's inherent reasoning difficulty alongside accuracy and conformance when interpreting the results.

Although we incorporated qualitative clinician review of representative reasoning traces, this assessment was limited to the cases included in the case study and was not intended as a comprehensive clinical validation of all generated reasoning traces. Future work should evaluate larger and more diverse case sets using multiple independent clinicians and predefined clinical reasoning criteria.

The conversion of reasoning traces into event logs relied on Gemma-3-12B as a single judge model. Errors or model-dependent biases in classifying reasoning types, diagnostic targets, and effects may therefore influence the discovered process models and conformance measurements. An otologist assessed whether the representative reasoning traces were clinically plausible as a whole. However, the otologist did not examine whether Gemma-3-12B correctly assigned the reasoning type, diagnostic target, and effect to each individual step. Future work will extend the evaluation to a broader range of LLM-based judge models and validate these step-level labels using independent annotations from multiple reviewers.

Acknowledgments

This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number : RS-2025-

25460099)

References

- Suhana Bedi, Yutong Liu, Lucy Orr-Ewing, Dev Dash, Sanmi Koyejo, Alison Callahan, Jason A Fries, Michael Wornow, Akshay Swaminathan, Lisa Soleymani Lehmann, et al. 2025. Testing and evaluation of health care applications of large language models: a systematic review. *Jama*, 333(4):319–328.
- Alessandro Berti, Humam Kourani, Gyunam Park, and Wil MP Van Der Aalst. 2025. Configuring large reasoning models using process mining: A benchmark and a case study. In *International Conference on Business Process Management*, pages 412–424. Springer.
- Longchao Da, Xiaoou Liu, Jiaxin Dai, Lu Cheng, Yaqing Wang, and Hua Wei. 2025. Understanding the uncertainty of llm explanations: A perspective based on reasoning topology. *arXiv preprint arXiv:2502.17026*.
- JP de Villiers, Allan De Freitas, Anne-Laure Jousselme, L Kaplan, Erik Blasch, Claire Laudy, and PC Costa. 2025. Evaluation of llm reasoning under uncertainty: An atomic comparison to normative approaches. In *2025 28th International Conference on Information Fusion (FUSION)*, pages 1–8. IEEE.
- Ziwei Deng, Mian Deng, Chenjing Liang, Zeming Gao, Chennan Ma, Chenxing Lin, Haipeng Zhang, Songzhu Mei, Siqi Shen, and Cheng Wang. 2026. Planu: Large language model reasoning through planning under uncertainty. *Advances in Neural Information Processing Systems*, 38:142636–142673.
- Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.
- Ching-Lai Hwang and Kwangsun Yoon. 2012. *Multiple attribute decision making: methods and applications a state-of-the-art survey*. Springer Science & Business Media.
- David E Newman-Toker and Jonathan A Edlow. 2015. Titrate: a novel approach to diagnosing acute dizziness and vertigo. *Neurologic clinics*, 33(3):577.
- Dorianis M Perez, Bryan E Kaiser, and Ismael Boureima. 2025. Uncertainty quantification by large language models. *Machine Learning with Applications*, page 100773.
- Thomas Savage, John Wang, Robert Gallo, Abdessalem Boukil, Vishwesh Patel, Seyed Amir Ahmad Safavi-Naini, Ali Soroush, and Jonathan H Chen. 2025. Large language model uncertainty proxies: discrimination and calibration for medical diagnosis and treatment. *Journal of the American Medical Informatics Association*, 32(1):139–149.

- Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965.
- Wil Van der Aalst, Arya Adriansyah, and Boudewijn Van Dongen. 2012. Replaying history on process models for conformance checking and performance analysis. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(2):182–192.
- Yinong Oliver Wang, Nivedha Sivakumar, Falaah Arif Khan, Rin Metcalf Susa, Adam Golinski, Natalie Mackraz, Barry-John Theobald, Luca Zappella, and Nicholas Apostoloff. 2025. Is your model fairly certain? uncertainty-aware fairness evaluation for llms. *arXiv preprint arXiv:2505.23996*.
- Anton JMM Weijters, Wil MP van Der Aalst, and AK Alves De Medeiros. 2006. Process mining with the heuristicsminer algorithm.
- Haifeng Xu, Jianfei Pang, Xi Yang, Jinghui Yu, Xueming Li, and Dongsheng Zhao. 2020. Modeling clinical activities based on multi-perspective declarative process mining with openehr's characteristic. *BMC Medical Informatics and Decision Making*, 20(Suppl 14):303.
- Alex L Zhang, Tim Kraska, and Omar Khattab. 2025. Recursive language models. *arXiv preprint arXiv:2512.24601*.
- Eckart Zitzler and Lothar Thiele. 1999. Multiobjective evolutionary algorithms: a comparative case study and the strength pareto approach. *IEEE transactions on Evolutionary Computation*, 3(4):257–271.

A Reasoning Trace Structuring

This appendix describes the label definitions used to transform free-text reasoning traces into structured event sequences.

A.1 Reasoning Types

The original taxonomy included ten reasoning types: Pattern Recognition (PR), Deductive Reasoning (DR), Inductive Reasoning (IR), Abductive Reasoning (AR), Hypothesis Generation (HG), Validation (V), Backtracking (B), Ethical or Moral Reasoning (ER), Counterfactual Reasoning (CR), and Heuristic Reasoning (HR) (Berti et al., 2025). We added Metacognitive Reasoning (MR) to represent self-monitoring and planning statements that managed the reasoning process without directly providing diagnostic evidence.

In the clinical diagnostic context, PR identified patterns or relevant findings in the patient information; DR derived diagnostic implications from clinical evidence or established rules; IR generalized from observed findings or prior cases; AR inferred plausible explanations for the observed symptoms; HG generated candidate diagnoses; V examined whether the available evidence supported a diagnostic hypothesis; B revised or reconsidered an earlier inference; ER addressed ethical or patient-related considerations; CR considered how an alternative finding might change the diagnosis; and HR applied simplified clinical rules or practical decision strategies. MR represented the monitoring, organization, or planning of the reasoning process itself.

A.2 Diagnostic Targets

Each reasoning step was assigned a diagnostic target indicating the condition addressed by that step. The target labels were BPPV for benign paroxysmal positional vertigo, VN for vestibular neuritis, MENIERE for Ménière's disease, and OTHERS for other diagnoses. GENERAL was assigned to reasoning concerning the overall diagnostic process without focusing on a single diagnosis, whereas UNCLEAR was used when the diagnostic target could not be determined from the text.

A.3 Reasoning Effects and Conclusion

The effect label represented the direction of a reasoning step with respect to its diagnostic target. Positive Effect (PE) indicated that the step supported the target diagnosis, Negative Effect (NE)

indicated that it argued against the target diagnosis, and Indifferent Effect (IND) indicated that its direction was neutral or could not be clearly determined. These labels described the role of each step in the diagnostic reasoning process and did not independently indicate whether the final diagnosis was correct. The activity label combined the reasoning type, diagnostic target, and effect, for example, *Deductive Reasoning: BPPV-NE*.

After the ordered reasoning steps, a terminal Conclusion event was added to record the model’s final predicted diagnosis. Unlike the intermediate events, the Conclusion event represented the outcome of the reasoning process rather than an additional reasoning type.

B Pareto Front

Figure 2 presents the three-dimensional objective spaces and model-specific Pareto fronts for the four LLMs. Each point represents an analyzable patient–model–prompt combination. Red circles indicate Pareto-optimal combinations, gray circles indicate dominated combinations, and blue diamonds indicate the ideal points. The three axes represent normalized Desirability, external conformance, and internal conformance.

The composition of the Pareto set varied across models. Qwen3-8B and Granite-3.3-8B both achieved a Pareto-set accuracy of 100%, indicating that all Pareto-optimal combinations identified for these models were correctly classified. OLMo-3-7B-Think achieved a Pareto-set accuracy of 80%, with most, but not all, Pareto-optimal combinations corresponding to correct diagnoses. DeepSeek-R1-Distill-Llama-8B had the lowest Pareto-set accuracy among the four models (50%), indicating an equal proportion of correct and incorrect combinations within its Pareto set.

For DeepSeek-R1-Distill-Llama-8B and OLMo-3-7B-Think, some combinations with relatively low Desirability remained Pareto optimal because no other combination performed at least as well across all three criteria while performing better on at least one. These combinations generally retained favorable external conformance or internal conformance despite their lower Desirability. The model-specific Pareto fronts therefore illustrate different trade-offs among Desirability, external conformance, and internal conformance rather than diagnostic correctness alone.

C Prompts

This appendix presents the prompt templates used in the experiments. Patient identifiers and true diagnostic labels were not included in the model inputs.

C.1 Base Prompt

The Multishot, CoT, RAG and RLM prompting methods shared the following classification instructions and output structure. The content inserted into the method-specific instruction field differed across prompting methods.

```
You are classifying an emergency
department dizziness/vertigo case.

Choose exactly one diagnosis class:
- BPPV
- VN
- MENIERE
- OTHERS

Use only the structured case variables.
"unknown" means not recorded; it does not
mean "no".
Do not use the true label, raw label,
assessment, patient id, or row index as
evidence.

<METHOD-SPECIFIC INSTRUCTIONS>

New case:
<STRUCTURED CASE INFORMATION>

<JSON OUTPUT SCHEMA>
```

C.2 Multishot Prompt

The Multishot prompt inserted 12 fixed labeled examples from the non-test set, with three examples from each diagnostic category. The examples were kept constant across models, target cases, and repeated runs.

```
Example 1
Case:
<STRUCTURED EXAMPLE CASE>
Answer: <DIAGNOSTIC LABEL>

...

Example 12
Case:
<STRUCTURED EXAMPLE CASE>
Answer: <DIAGNOSTIC LABEL>

Now classify the new case.
<STRUCTURED NEW CASE>
```

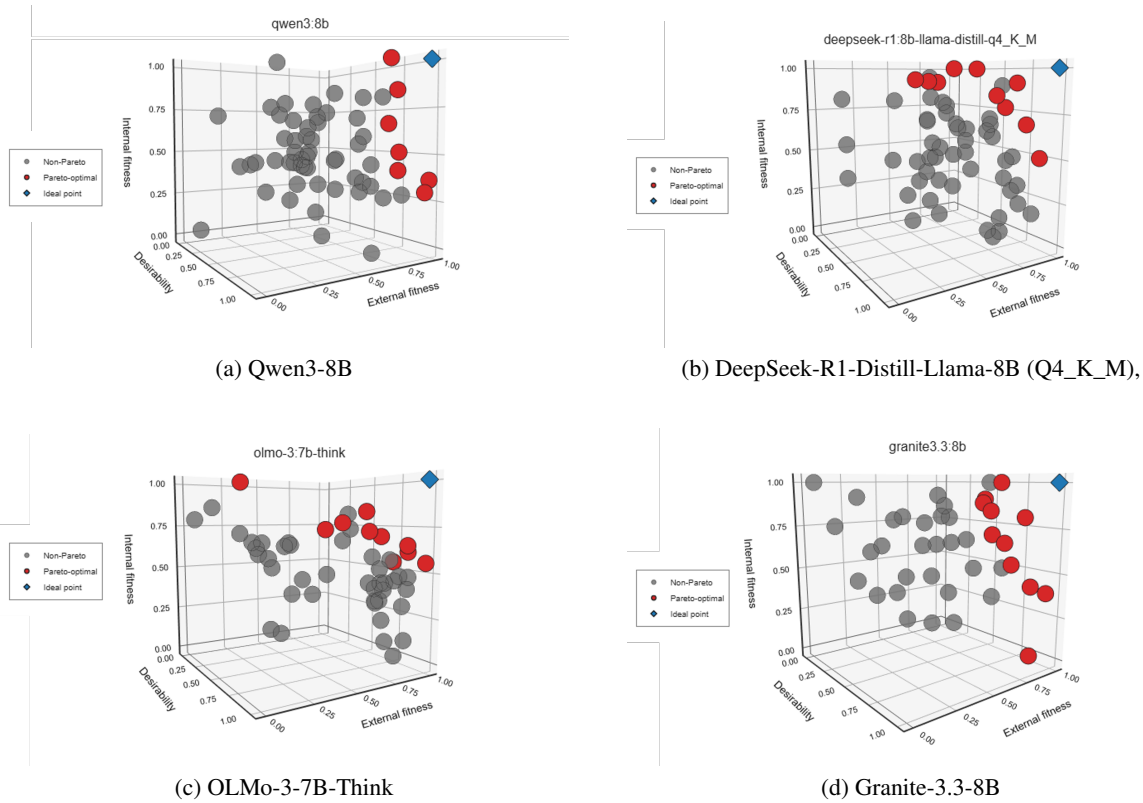


Figure 2: Three-dimensional objective spaces and model-specific Pareto fronts based on normalized Desirability, external conformance, and internal conformance: (a) Qwen3-8B, (b) DeepSeek-R1-Distill-Llama-8B (Q4_K_M), (c) OLMo-3-7B-Think, and (d) Granite-3.3-8B.

C.3 CoT Prompt

The chain-of-thought (CoT) prompt was adapted from the TiTrATE (Newman-Toker and Edlow, 2015) framework for evaluating patients with acute dizziness and vertigo. The following method-specific instructions were inserted into the shared prompt template:

Analyze the case using a domain-guided clinical reasoning framework based on the TiTrATE approach.

<INSTRUCTIONS based on TiTrATE>

Use the above sequence to reach the diagnosis. Keep the reported reasoning concise and evidence-based.

C.4 RAG Prompt

The retrieval-augmented generation (RAG) method retrieved the eight highest-ranked Barany Society diagnostic-criteria passages using BM25 and inserted them into the shared prompt template:

Use the retrieved diagnostic criteria below as supporting evidence. Compare the clinical findings in the case with the retrieved criteria and select the most appropriate diagnosis.

<RETRIEVED BARANY SOCIETY CRITERIA>

C.5 RLM Prompt

The RLM method used the same diagnostic classes, input restrictions, and JSON output schema as the shared prompt. However, it used a separate interaction structure in which the structured case was stored in the context variable and inspected through a restricted Python REPL. The model could iteratively analyze the case and request depth-one submodel analyses using `llm_query()`. To limit computational cost and maintain consistent inference conditions, the process was restricted to four iterations, a recursion depth of one, and four submodel calls.