Cue Uptake Is Not Quantitative Execution: Symbolic Guardrails for Prompted LLMs in Sparse Clinical Time-Series Prediction

Anonymous Authors
Anonymous Institution
anonymized

Abstract

Prompted large language models (LLMs) can name clinically meaningful cues while producing numerical predictions, but cue recognition need not control quantitative action. We study this gap in sparse clinical time series, where variable identity, time anchors, trends, ranges, and thresholds define explicit symbolic constraints. Using 12,000 MIMIC-IV ICU admissions, we evaluate DeepSeek-chat and GPT-4o on masked point imputation and audit outputs with CRES, a five-dimensional behavioural verifier. In 240-task directional counterfactual benchmarks per model, cue mention is high while correct signed execution is near zero; a 60-task o4-mini reasoning-model probe preserves the gap, and paired McNemar tests show it is not sampling noise. CRES receives bounded 60-case dual-rater support, strongest for objective physiological checks and weaker for text-derived dimensions. Routing is therefore treated as a suggestive guardrail pattern, not a decisive performance claim. Verbal symbolic uptake is an unsafe proxy for numerical control, even under explicit clinical-symbolic prompting.

1 Introduction

LLMs are increasingly used for tasks that mix continuous values with symbolic structure. Clinical time series are a useful stress test: an ICU value is a measurement tied to a variable, time point, local trajectory, physiological range, and often a critical threshold. If an LLM writes that lactate is rising or sodium is bounded, a neuro-symbolic system should ask whether that cue actually constrains the generated number.

We treat generated predictions and rationales as artifacts to verify against explicit symbolic constraints, not as evidence of latent reasoning. The central result is that verbal symbolic recognition and quantitative execution are separable behaviours. In LLM-agent workflows, rationales, plans, and state descriptions can become inputs to downstream decisions; treating them as faithful control signals is therefore risky.

Our contributions are a paired directional counterfactual audit separating cue uptake from signed execution; CRES, a behavioural symbolic verifier over evidence, trend, time, physiology, and task alignment; and two bounded guardrail patterns, validation-derived routing and explanation-feature stacking. The contribution is diagnostic: explicit symbols are used to audit and route outputs, while fluent cue mentions are tested rather than assumed to certify quantitative action.

2 Related Work and Positioning

Imputation methods such as BRITS and SAITS learn temporal dependencies from partial sequences [Cao *et al.*, 2018; Du *et al.*, 2023], while LLM-adapted time-series models test whether pretrained language models can represent temporal structure [Jin *et al.*, 2024]. Our question is different: when a prompted LLM returns both a value and a rationale, do the symbolic cues in the rationale actually constrain the number it outputs?

Medical LLM studies show that correctness, confidence, explanation quality, and calibration are distinct [Xiong *et al.*, 2024; Savage *et al.*, 2025; Chen *et al.*, 2025], and chain-of-thought explanations can be unfaithful [Turpin *et al.*, 2023]. We extend this concern to sparse clinical time-series imputation by pairing each original prediction with a directionally signed intervention and testing whether cue mention becomes numerical movement. Counterfactual and causal-reasoning audits likewise warn that prompt sensitivity is not correct directional execution [Kiciman *et al.*, 2024]. Here, the symbolic layer is output-facing: variable identity, time anchors, local trends, physiological ranges, and thresholds are used to audit and route LLM predictions rather than to infer hidden reasoning [Garcez and Lamb, 2023].

3 Symbolic Verification Framework

Figure 1 summarizes the framework. The learned component produces a point estimate, interval, and rationale; the symbolic component checks time anchors, variable identity, local trend, ranges, and thresholds. These are behavioural checks over the generated artifact, not white-box claims about hidden mechanisms.

Data and task. We use MIMIC-IV v3.1 adult ICU admissions [Johnson *et al.*, 2023]: 12,000 admissions, 997,793 hourly aligned measurements, and 12 variables spanning dense vital signs and sparse labs. Grid-level missingness is 91.23% (80.6% for heart rate to 97.2% for lactate). The task

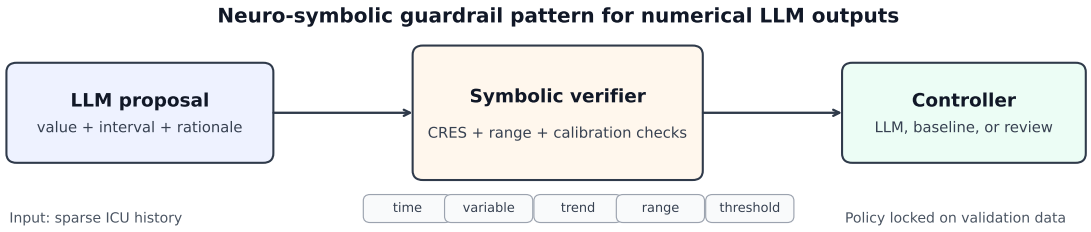

Figure 1: Neuro-symbolic guardrail pattern: an LLM proposes a value and rationale; a symbolic verifier checks time, variable, trend, range, threshold, CRES, and calibration signals; a validation-locked controller routes to LLM, baseline, or review.

CRES dim.	Symbolic constraint	Operational check
Evidence	cited observation exists	variable/evidence hit
Trend	stated direction matches trajectory	direction consistency
Temporal	rationale uses relevant window	anchor hit
Physiology	value respects plausible range	range/threshold check
Diagnostic	rationale matches target task	task-variable alignment

Table 1: CRES operationalizes symbolic grounding as output-level checks.

masks an observed value after patient-level splitting and asks the model to impute it from prior target values, cross-variable context, metadata, ranges, and thresholds. Masked tasks are sampled from the held-out evaluation pool after preprocessing rather than selected from model outputs. This keeps the audit tied to a fixed clinical-symbolic task distribution rather than to post hoc failure selection.

Models and baselines. Prompted models are DeepSeek-chat and GPT-4o at temperature 0.3, plus a smaller o4-mini reasoning probe. Baselines include true-time and index interpolation, last observation, local medians, BRITS, SAITS, and other neural baselines. We use them to bound claims rather than build a leaderboard.

CRES as symbolic verifier. Table 1 formalizes CRES as a set of symbolic output constraints. Each dimension maps a clinical symbol to an output-level check; CRES is a behavioural proxy, not latent reasoning ground truth. We do not treat it as monolithic: hard checks–physiological range, evidence existence, and task alignment–support guardrail decisions, while softer text-derived trend and temporal checks are audit signals. The distinction is operational. Hard checks compare generated artifacts to fixed objects in the prompt or to prespecified clinical bounds, so disagreement is usually about extraction or annotation. Soft checks require judging whether free text uses the right local trajectory or time window, so lower agreement is expected and these dimensions are not allowed to drive routing by themselves. A blinded 60-case dual-rater package uses ordinal 0/1/2 ratings. Ranges, thresholds, and cue markers are fixed before model evaluation.

Directional counterfactuals. For each task, the original prompt gives y; a paired prompt injects a fixed variable-specific directional cue and gives y'. With expected direction $d \in \{-1, +1\}$, correct execution requires $\mathrm{sign}(y' - y) = d$ beyond tolerance $\max(0.1, 0.05w_v)$. Cue uptake and negative-control stability are measured separately. The cue templates are fixed before model evaluation. Each intervention keeps the same clinical context and changes only the directional statement; the matched negative control adds a neutral variable-relevant statement without an expected signed movement. The cues are mechanism statements rather than numeric answers, e.g., atrial fibrillation for higher heart rate, shock for lower MAP, SIADH for lower sodium, kidney injury for higher creatinine/BUN, bleeding for lower haemoglobin/haematocrit, infection for higher WBC, and coagulopathy for lower platelets; they never state the masked target value. The expected sign comes from the locked cue table, not from inspecting model predictions. The main benchmark uses 240 tasks per chat model, balanced as 20 per variable; o4-mini uses 60 tasks, five per variable, as a cost-bounded reasoning-model probe. A retrospective 60-item author audit checks the full intervention unit–context, injected cue, expected sign, and prompt wording–for cue plausibility, context compatibility, directional validity, naturalness, and absence of direct target leakage.

Parsing and statistical tests. The parser extracts the structured point estimate, interval, evidence points, and free-text rationale from each response. Counterfactual labels are computed from the paired numerical outputs and from locked cue markers; a parser audit re-reads raw responses to test whether extraction errors could explain the gap. For the main cue-execution claim, each task contributes two paired binary indicators: cue mentioned and correct signed execution. We use exact McNemar tests on discordant pairs, so the reported p-values test whether cue-only cases dominate sign-only cases within the same prompts. For MAE comparisons, all models and baselines are evaluated on the same held-out masked points, and bootstrap intervals preserve paired outputs when resampling. Anonymized supplementary materials will include full prompts, the locked cue table, variable ranges and thresholds, task-selection protocol, parser-audit records, and audit sheets.

Guardrail comparison. Layer 1 uses validation-derived variable routing plus hard physiological checks, selecting an LLM or temporal baseline. Layer 2 stacks baseline predictions, LLM predictions, uncertainty features, and CRES-derived explanation features in a ridge-style meta-model.

Algorithm 1 Symbolic Guardrail Routing

Require: context x, LLM value y, rationale r, baseline b
Parse value, interval, evidence, trend, and rationale fields.
Compute hard checks and CRES audit signals.
if y violates a physiological range or critical threshold **then**
  route to baseline or review.
else if the validation-locked route for the target variable favours the LLM **then**
  route to y.
else
  route to temporal baseline b.
end if
Treat native intervals as qualitative unless calibrated.

Component	Locked design
Data	MIMIC-IV v3.1; 12k ICU; patient split
Models	DeepSeek-chat, GPT-4o; o4-mini probe
Counterfactuals	240/model + 60 o4-mini; directional + negative control
Cue audit	60/60 context-usable; no leakage; 41.7% opposing local trend
Parser audit	DeepSeek mismatch 8.2%; o4-mini 0/180
CRES validation	60 cases; 2 raters; hard checks for control, soft checks for audit
Routing	Variable-level validation route; held-out paired bootstrap
Supplement	Prompts, cue table, ranges, parser records, audit sheets

Table 2: Audit and reproducibility capsule.

Routing is validation-locked before held-out evaluation; soft CRES dimensions are audit signals, not routing controls. The route is chosen at the variable level on validation data and then frozen for the held-out set. Thus routing evaluates a fixed validation-derived controller, not an adaptive per-example selector.

4 Results

Figure 2 summarizes the claim-bearing results. CRES has bounded validation support, not gold-standard clinical validity. Mean weighted kappa is 0.292, but objective dimensions are stronger: Evidence 0.401, Physiological 0.482, and Diagnostic 96.7% exact agreement with kappa 0 from a ceiling effect. Trend and temporal-text dimensions are weaker, so CRES is used mainly as an audit proxy.

The directional benchmark exposes the key failure mode. DeepSeek mentions the cue in 78.8% of tasks but executes the correct sign in 2.5%; GPT-4o shows 75.0% versus 0.8%. Exact McNemar discordance is 183 versus 0 for DeepSeek ($p \approx 1.6 \times 10^{-55}$) and 178 versus 0 for GPT-4o ($p \approx 5.2 \times 10^{-54}$), with negative-control stability 77.5% and 79.2%. The o4-mini reasoning probe preserves the separation: cue mention 75.0%, sign execution 1.7%, negative-control stability 93.3%, and discordance 44 versus 0 ($p \approx 1.1 \times 10^{-13}$). Parser audits do not change the conclusion: DeepSeek mismatch is 8.2%; o4-mini mismatch is 0/180. The cue audits accept 60/60 sampled interventions as plausible, directionally valid, non-leaking, and context-usable with no screened explicit contradictions. However, 41.7% contain an opposing local trend, so correct sign is interpreted as all-else-equal symbolic-intervention compliance, not as proof that the cue clinically dominates all patient evidence. Together, the negative controls, parser audit, and cue audit narrow the most direct artifact explanations. The residual claim is intentionally behavioural: under fixed clinical-symbolic prompts, models can verbalize an intervention without quantitatively following it.

Claim boundaries. Local temporal structure remains strong: true-time interpolation reaches MAE 5.60, and the held-out index-linear baseline is 6.07 over 604 points. Native 95% LLM intervals cover only 51.4%. Held-out routing is a validation-locked control pattern, not a decisive improvement claim: routed MAE is 5.52, but the bootstrap delta against index-linear is -0.55 with 95% CI [-1.78, 0.11]. Explanation-feature stacking is negative evidence for using rationales as direct numerical correction, with MAE about 6.77. Because true-time interpolation nearly matches routing, the defensible conclusion is that symbolic checks can bound LLM use and prevent raw LLM dependence, while the primary empirical contribution remains the signed counterfactual audit. The failed stacking result is also informative: text-derived rationale features can describe cues without carrying reliable numerical control information. This is why they are treated as audit evidence rather than as trusted correction features.

Mechanism audits support this interpretation. Across 604 points, LLM outputs correlate with last-observation carry-forward (0.820 DeepSeek, 0.839 GPT-4o) and regression-to-normal proxies, yet remain worse than the simple rules they resemble. Among abnormal truths, predictions are normal-valued in 30.8% and 32.8% of cases. A targeted sodium audit finds one GPT-4o physiological violation: 0.8 mEq/L for true sodium 141 mEq/L, motivating hard range checks.

5 From Audit to Guardrail Policy

The audit suggests a conservative deployment pattern for numerical LLM outputs. LLM rationales should be treated as hypotheses to verify, not faithful traces of computation. Algorithm 1 converts the audit into a reusable control pattern: parse the proposal, check symbolic constraints, calibrate where possible, and route to an LLM, baseline, or review. Its value is not a new imputation architecture but a separation between generation, verification, and control. In this setting, symbols are most useful as externally specified constraints that make LLM behaviour inspectable. They should not be inferred from fluent rationales alone, because the same rationale channel can acknowledge a cue while leaving the numerical action unchanged or moving in the wrong direction.

The resulting policy keeps the LLM behind symbolic checks rather than treating it as a standalone imputer. A reliable workflow should retain local baselines, audit explicit symbols, calibrate uncertainty, reject physiological violations, and route LLM predictions only where validation evidence supports complementarity. Before rationales influence action, symbolic cue uptake should be measured separately from quantitative execution.

Limitations. The experiments are behavioural, single-dataset, and based on synthetic masking rather than the full ICU measurement process. Routing is suggestive because its bootstrap interval crosses zero. CRES has bounded dual-rater support; the cue-naturalness audit is retrospective au-

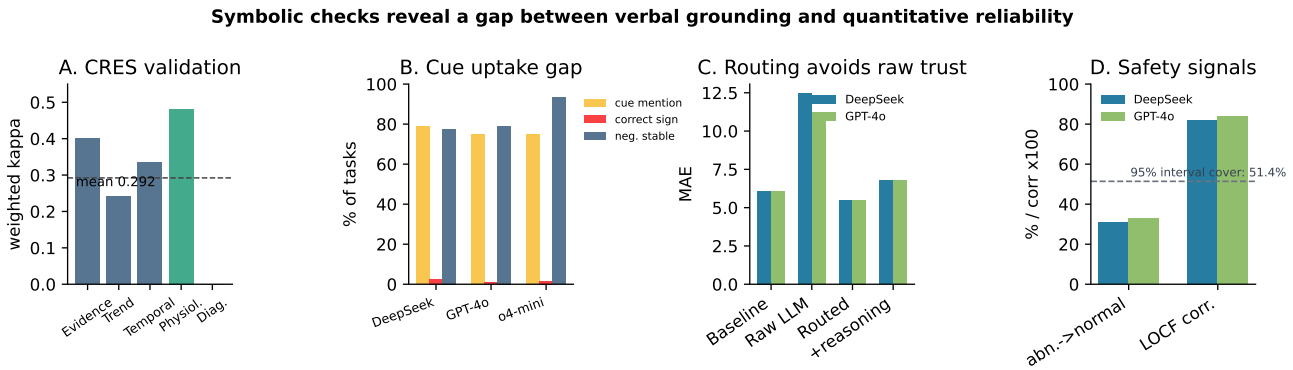

Figure 2: Evidence dashboard. A: bounded CRES validation. B: cue-execution gap, including the 60-task o4-mini probe. C: bounded routing versus raw LLM output and explanation-feature stacking. D: interval under-coverage, regression-to-normal, and LOCF correlation signals.

thor review, not prospective clinician validation. The context-conditioned cue audit screens for explicit contradiction but does not replace clinician adjudication of whether a cue should dominate competing bedside evidence. The o4-mini run is a 60-task probe, and broader reasoning-model coverage remains future work. The central cue-execution result is still measured directly with paired counterfactual perturbations. Future work should replace retrospective review with prospective clinician adjudication, test additional datasets and reasoning models, and evaluate end-to-end clinical workflows.

Ethics Statement

The study uses de-identified MIMIC-IV data under PhysioNet access and involves no patient contact. The results do not support autonomous clinical deployment.

References

[Cao *et al.*, 2018] Wei Cao, Dong Wang, Jian Li, Hao Zhou, Lei Li, and Yitan Li. Brits: Bidirectional recurrent imputation for time series. In *Advances in Neural Information Processing Systems*, 2018.

[Chen *et al.*, 2025] Hanjie Chen, Zhouxiang Fang, Yash Singla, and Mark Dredze. Benchmarking large language models on answering and explaining challenging medical questions. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2025. arXiv:2402.18060.

[Du *et al.*, 2023] Wenjie Du, David Cote, and Yan Liu. Saits: Self-attention-based imputation for time series. *Expert Systems with Applications*, 219:119619, 2023.

[Garcez and Lamb, 2023] Artur d'Avila Garcez and Luis C. Lamb. Neurosymbolic ai: The 3rd wave. *Artificial Intelligence Review*, 56:12387–12406, 2023.

[Jin *et al.*, 2024] Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y. Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. Time-llm: Time series forecasting by reprogramming large language models. In *International Conference on Learning Representations*, 2024.

[Johnson *et al.*, 2023] Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023.

[Kiciman *et al.*, 2024] Emre Kiciman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality. *Transactions on Machine Learning Research*, 2024. arXiv:2305.00050.

[Savage *et al.*, 2025] Thomas Savage, John Wang, Robert Gallo, Abdessalem Boukil, Vishwesh Patel, Seyed Amir Ahmad Safavi-Naini, Ali Soroush, and Jonathan H. Chen. Large language model uncertainty proxies: discrimination and calibration for medical diagnosis and treatment. *Journal of the American Medical Informatics Association*, 32(1):139–149, 2025.

[Turpin *et al.*, 2023] Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36, 2023.

[Xiong *et al.*, 2024] Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in large language models. In *International Conference on Learning Representations*, 2024.