Preprints.org

Article

Not peer-reviewed version

Preventing Wrong-Patient, Wrong-Timepoint, and Stale-Chart Errors in Large Language Model Clinical Decision Support via Discourse-Stateful Verification

Ravi Shanker Thadishetti * and Aishwarya Nair

Posted Date: 26 November 2026

Keywords: biomedical large language models; clinical decision support systems; biomedical natural language processing; electronic health records; evidence-grounded verification; tool-augmented AI agents; clinical text mining; healthcare AI reliability

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC, OpenAlex.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org | NOT PEER-REVIEWED | Posted: 26 November 2026 https://doi.org/10.20944/preprints202611.2737.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Preventing Wrong-Patient, Wrong-Timepoint, and Stale-Chart Errors in Large Language Model Clinical Decision Support via Discourse-Stateful Verification

Ravi Shanker Thadishetti [1,*] and Aishwarya Nair [2]

[1] Department of Information Technology, University of the Cumberlands, Williamsburg, KY 40769, USA

[2] Syracuse University, Syracuse, NY 13244, USA

* Correspondence: rthadishetti57651@ucumberlands.edu

Abstract

Delegating arithmetic from large language models (LLMs) to clinical calculators prevents mental-math drift, yet tool calls remain unsafe when models bind relatives' vitals, confuse discharge with admission labs, misconvert molar units, or commit against electronic health record (EHR) notes updated mid-inference. We present MedCalc-Enforce, a runtime verification and transactional commit kernel for clinical calculator calls. Before execution, a paragraph-stateful discourse automaton tracks subject identity and encounter phase across sentences while blocking relative deltas; a 42-digit decimal canonicalizer converts molar and mass units across 16 clinical scores; and a SQLite write-ahead logging (WAL) protocol re-projects bound fields atomically under concurrent addenda. Across 620 held-out runs over five open-weight LLMs (1.5B–8B; 124 patient clusters) at zero API cost, MedCalc-Enforce lifts agreement from 76.9% (477/620) to 89.4% (554/620; Δ = +12.4%, p = 4.10×10^{-12}) and cuts unsafe releases by 89.7% from 17.3% (107/620) to 1.8% (11/620; Δ = −15.5%, p = 2.26×10^{-16}), with 93.6% of releases (99.1%–100.0% on 3.8B–8B models) from verified proposals. MedCalc-Enforce blocks 240/240 live multi-sentence traps and 24/24 unseen surface-form traps, sustaining 2,881.8 checks/s with zero stale commits or false addendum blocks.

Keywords: biomedical large language models; clinical decision support systems; biomedical natural language processing; electronic health records; evidence-grounded verification; tool-augmented AI agents; clinical text mining; healthcare AI reliability

1. Introduction

Although frontier large language models (LLMs) achieve strong scores on static medical licensing examinations [1], realistic clinical workflows expose persistent vulnerabilities in laboratory interpretation, guideline adherence, and sensitivity to the order and framing of electronic health record (EHR) narratives [2]. Delegating arithmetic to external software tools has therefore become standard practice when deploying LLMs in quantitative clinical tasks [3–5]. Rather than asking a neural decoder to evaluate fractional exponents in the 2021 CKD-EPI glomerular filtration equation [6] or square roots in Mosteller body surface area [7], an agent emits a structured JSON tool call and lets deterministic code compute the number. Grammar-constrained decoding [8] strengthens this pattern further by guaranteeing at token-sampling time that every emitted payload parses against a declared JSON schema.

Yet syntactic validity and arithmetic delegation solve only half of the clinical safety problem. In real hospital notes, numerical failures routinely originate upstream of the formula—inside argument binding—and downstream of the formula—inside concurrent state publication:

1. **Cross-sentence subject and temporal misattribution:** Narrative notes frequently group measurements under a framing sentence before listing bare values in subsequent sentences (e.g., 'Family history: The following measurements belong to his brother. Height 182 cm. Weight 94

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 26 November 2026 https://doi.org/10.20944/preprints202611.2737.v1

kg. Patient refused exam.'). Schema, physiological range, and literal quote gates all pass on 182 cm and 94 kg, and clause-local filters fail once the subject or timepoint cue sits in an earlier sentence.

2. **Silent unit-system and delta-quantity drift:** Clinical charts mix mass concentrations (mg/dL) with molar concentrations (mmol/L, μmol/L) whose conversion factors depend on analyte molecular weight (18.016 for glucose vs. 2.8 for BUN and 88.4 for creatinine), and mix absolute weights with relative changes ('lost 6 kg over two months').
3. **Concurrent EHR chart races between staging and publication:** Because LLM tool planning takes hundreds of milliseconds to seconds, a nurse or lab interface can append an addendum or correct a vital sign mid-inference; unchecked read-then-write commits stale outputs, whereas whole-note revision locks needlessly discard valid calculations on benign addenda.

Existing agentic safety layers respond to extraction errors either by invoking a second LLM critic [9] or by defining dialogue-level programmable rails [10]. However, ungrounded verbal critique adds latency and inference cost without enforcing deterministic invariants over note provenance or database serializability, while general-purpose topical guardrails do not verify analyte stoichiometry or concurrent chart freshness. Conversely, pure deterministic regular-expression extractors fail whenever notes record multiple serial laboratory draws or require categorical clinical reasoning across discrete risk criteria, and rigid question templates break when clinicians ask for alternate output units such as Bazett QTc in seconds rather than milliseconds [11].

This study introduces MedCalc-Enforce (v0.6), a zero-dependency Python and SQLite WAL runtime enforcement architecture built around a strict separation of concerns: an untrusted LLM proposes candidate tool selections and span-grounded extractions, while a deterministic kernel verifies discourse scope, unit canonicalization, multi-tool snapshot equality, and field-projected commit serializability before any numerical result can be released. Specifically, we investigate four research questions:

- **RQ1 (Cross-Family Clinical Safety, Attribution, and Accuracy):** Across open-weight LLM backbones spanning four architecture families from 1.5B to 8B parameters, how effectively does discourse-stateful runtime verification reduce unsafe nonmatching numerical releases and selective-prediction risk relative to schema/unit gates, clause-local verification, and LLM-free regex extraction?
- **RQ2 (Adversarial and Out-of-Vocabulary Discourse Generalization):** Can a frozen paragraph-level discourse automaton eliminate false releases across multi-sentence kin/phase traps and transfer zero-shot to unseen clinical note surface forms without sacrificing recall on multi-section patient charts?
- **RQ3 (Bounded Diagnostic Repair vs. Ungrounded Critique):** Does returning typed deterministic rejection codes for a single bounded retry ($k = 1$) recover malformed or misbound proposals safely without introducing valid-call regressions?
- **RQ4 (Concurrent Commit Freshness and Throughput Scaling):** Can atomic field-projected revalidation inside SQLite WAL transactions simultaneously prevent stale-data commits and eliminate false withholdings from benign chart addenda at multi-worker hospital ledger scale?

Summary of Contributions. (1) We formalize discourse-stateful clinical argument binding with explicit lexical-precondition and serial-draw boundaries, proving cross-sentence non-patient/incompatible-phase soundness alongside field-projected snapshot serializability (Section 2.2). (2) We implement an auditable 16-calculator verification engine using 42-digit decimal arithmetic across 10 continuous equations, 6 discrete/hybrid clinical risk scores (CHA_2DS_2-VASc, CURB-65, Wells DVT, Glasgow Coma Scale, MELD-Na, and Cockcroft–Gault), chained two-stage oncology dosing checks, and tamper-evident SHA-256 SQLite WAL audit logging. (3) We execute an 860-run live open-weight evaluation across five pinned GGUF models, decompose provenance attribution across LLM proposal vs. repair vs. fallback tiers, audit all residual errors case-by-case, evaluate a frozen 64-case out-of-vocabulary clinical transfer benchmark, and verify concurrency and 100,000-note scalability at $0 API spend.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 26 November 2026 https://doi.org/10.20944/preprints202611.2737.v1

1.1. Related Work and Positioning

Clinical LLM Evaluation, Hallucination, and Tool-Augmented Calculation. Despite rapid gains on multiple-choice medical exams [1], controlled clinical evaluations demonstrate that LLMs frequently misinterpret raw laboratory panels, violate diagnostic guidelines when note order shifts [2], and do not automatically improve clinician diagnostic accuracy in randomized vignette trials without structured workflow integration [12]. Multi-dimensional evaluation frameworks such as QUEST [13] and physician-rubric benchmarks such as HealthBench [14] emphasize safety, factual grounding, and abstention under missing clinical context as primary deployment gates. For quantitative tasks specifically, MedCalc-Bench [3] showed that chain-of-thought prompting suffers from both formula recall errors and arithmetic drift. Tool-calling paradigms such as ReAct [4] and Toolformer [5] eliminate arithmetic errors once canonical inputs are supplied, yet small and mid-sized models still misbind parameters from dense clinical prose, confuse uncorrected QT with QTc, or substitute post-stabilization values when admission labs are requested. MedCalc-Enforce targets this exact provenance gap between raw clinical text and deterministic code execution.

Clinical Discourse Context, Experiencer Attribution, and EHR Temporal Extraction. Determining whether an extracted clinical concept applies to the patient or a family member, whether it is negated, and where it sits along an encounter timeline is a foundational challenge in biomedical natural language processing. Rule-based algorithms such as ConText [15] propagate lexical triggers across local token windows to tag negation, historical status, and non-patient experiencers, while recent comparative studies adapt encoder and decoder LLMs to clinical temporal relation extraction across unstructured EHR timelines [16]. However, neural temporal extractors remain probabilistic and struggle on sparse cross-sentence relations, whereas intra-sentence trigger windows lose scope as soon as bare vital signs follow a framing sentence ('Measurements below belong to his brother. Weight 94 kg.'). MedCalc-Enforce bridges rule-based interpretability and LLM tool calling by enforcing a deterministic paragraph-stateful discourse automaton directly over proposed calculator spans and fallback candidates, resetting cleanly at explicit section boundaries.

Programmable Guardrails, Constrained Decoding, and Selective/Conformal Risk Control. Runtime safety toolkits such as NeMo Guardrails [10] interpose programmable dialogue rails around LLM calls, and finite-state grammar engines [8] constrain token logits so emitted payloads satisfy JSON types. During our five-model evaluation, we uncovered an undocumented interaction in server-side grammar compilation: chat templates whose generation-prompt suffix injects partial role tags into the grammar stack can trigger sampler exceptions unless template rendering is decoupled from GBNF enforcement. More fundamentally, syntactic schemas cannot verify analyte stoichiometry or cross-sentence subject identity, and verbal self-correction loops [9] frequently overwrite valid initial extractions. Selective classification [17] and Conformal Language Modeling [18] formalize risk–coverage trade-offs and calibrated rejection rules for neural predictors, but conformal calibration typically assumes exchangeable holdout splits and multi-sample generation. MedCalc-Enforce achieves selective risk control deterministically in a single greedy pass by mapping span provenance into discrete verification tiers and withholding execution whenever discourse or stoichiometric invariants fail.

Optimistic Concurrency Control and Recent Medical Informatics Advances. Classical snapshot isolation and optimistic concurrency protocols [19] detect read-write conflicts at commit time using row or document revision counters. In active inpatient charts, whole-document revision locking is overly conservative: a nursing disposition addendum increments the note revision even when admission height, weight, and serum creatinine remain untouched. MedCalc-Enforce introduces field-projected semantic revalidation inside SQLite WAL immediate transactions, re-evaluating deterministic span projection on the live chart only when the revision advances and committing safely whenever the bound clinical projection and discourse state remain invariant. Complementary 2026 studies in Computers have examined rationale-prediction alignment in clinical decision support [31], auditable healthcare resource pipelines [32], resource-efficient on-device transformer models [33], and evidence-grounded longitudinal EHR temporal QA benchmarks [34];

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 26 November 2026 https://doi.org/10.20944/preprints202611.2737.v1

MedCalc-Enforce extends this line of inquiry to stateful clinical calculator provenance, exact molar canonicalization, and field-projected transactional commit.

2. Materials and Methods

2.1. System Architecture and Threat Model

Figure 1 illustrates the three-stage MedCalc-Enforce runtime pipeline. An encounter record is stored as an immutable-history row (p, e, r, N_r, d_r) indexed by patient identifier p, encounter identifier e, monotonically increasing integer revision $r \geq 1$, UTF-8 note text N_r, and cryptographic digest d_r = SHA256(N_r). Given a clinical question q and snapshot (r_0, N_{r0}, d_{r0}), an untrusted LLM acts first as a planner selecting a candidate calculator identifier $c \in C \cup \{NONE\}$ and second as an extractor proposing a structured argument map $\hat{A}$ containing numeric/categorical values, observed unit strings, and verbatim evidence quotes.

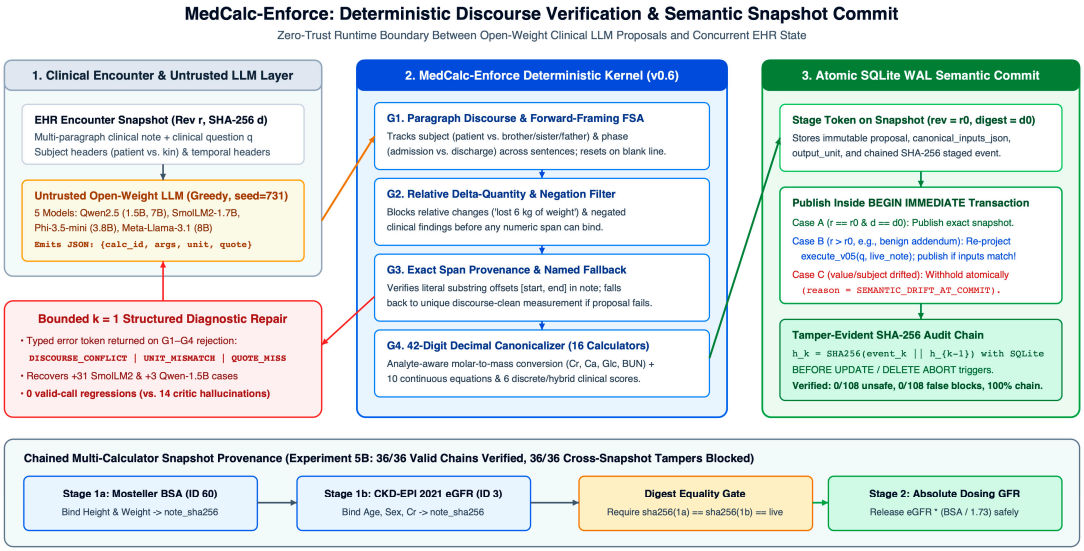


Figure 1. End-to-end MedCalc-Enforce (v0.6) runtime verification and transactional commit architecture across gates G1–G4, bounded k = 1 diagnostic repair, and atomic SQLite WAL BEGIN IMMEDIATE publication.

We treat the LLM as an untrusted proposer subject to four concrete fault classes: (i) hallucinated spans or mismatched numbers absent from the verbatim quote; (ii) discourse misattribution (binding vitals belonging to a relative or to a conflicting encounter phase such as discharge when q requests admission); (iii) unit/scale and relative-delta errors (molar vs. mass confusion, seconds vs. milliseconds, or binding weight loss Δw as total body weight w); and (iv) stale-snapshot writes when concurrent updates produce N_{r1} ($r_1 > r_0$) prior to publication.

2.2. Formal Discourse Provenance, Scope Preconditions, and Commit Guarantees

MedCalc-Enforce partitions N_r at blank-line and standalone section-header boundaries into paragraphs P_1, …, P_m and segments each paragraph into ordered sentences $S_{i,1}$, …, $S_{i,k}$. Let L_kin denote the frozen trigger vocabulary of non-patient experiencers (father, mother, son, daughter, brother, sister, spouse, husband, wife, sibling, partner, caregiver, grandfather, grandmother, twin), L_phase the frozen encounter-phase triggers (admission, on presentation, prior to treatment vs. discharge, follow-up, post-treatment, hospital day ≥ 2), and L_fwd forward-scoping headers and framing clauses ('Family history:', 'Measurements below belong/pertain/apply to...').

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 26 November 2026 https://doi.org/10.20944/preprints202611.2737.v1

Definition 1 (Paragraph-Stateful Discourse Automaton). *Let Σ_subj = {PATIENT, NON_PATIENT} and Σ_phase = {UNSPECIFIED, ADMISSION, LATER}. Within paragraph P_i, the automaton initializes at $\sigma_{i,0}$ = (PATIENT, UNSPECIFIED) and updates $\sigma_{i,j} = \delta(\sigma_{i,j-1}, S_{i,j})$: (1) if $S_{i,j}$ matches L_fwd with latest experiencer cue in L_kin and no patient-return pivot, subject state transitions to NON_PATIENT across subsequent sentences in P_i until an explicit patient-reset cue or paragraph boundary P_{i+1}; and (2) if $S_{i,j}$ matches L_fwd with phase cue $\tau \in$ L_phase, temporal state transitions to τ across subsequent bare-value sentences in P_i until overridden or reset at P_{i+1}.*

Definition 2 (Discourse-Clean Canonical Span Binding). *Given requested phase τ_q, a character span [a, b) in sentence $S_{i,j}$ with effective state (η_eff, τ_eff) (combining clause-local cues within a 64-character post-qualifier window and inherited paragraph state $\sigma_{i,j}$) is discourse-admissible for field f iff η_eff = PATIENT, τ_q = UNSPECIFIED or $\tau_eff \in$ {UNSPECIFIED, τ_q}, the local clause contains no negation, hypothetical/goal status, or relative-delta cue ('lost/gained x kg'), and the extracted raw quantity maps via exact analyte stoichiometry $\Phi_f(v, u)$ into physiological domain D_f.*

Theorem 1 (Soundness Under Recognized Cross-Sentence Kin and Phase Framing). *Fix frozen lexicons (L_kin, L_phase, L_fwd). If every domain-compatible mention of required field f* in note N_r resides inside spans [a, b) whose local clause or governing paragraph frame induces η_eff = NON_PATIENT or incompatible phase $\tau_eff \notin$ {UNSPECIFIED, τ_q} ($\tau_q \neq$ UNSPECIFIED), MedCalc-Enforce releases no numerical value (released = False) for any proposal $\hat{A}$.*

Proof of Theorem 1. Let f* ∈ F_c denote the required clinical field satisfying the stated precondition in note N_r. Both release paths—verified proposal repair (repair_call_v05) and unique named-measurement fallback (note_only_v05)—filter candidate observations through the identical discourse_conflict predicate; under the stated precondition, the clean candidate pool for f* is empty in both paths. Because Execute releases a numerical output only when every required f ∈ F_c resolves to a non-empty discourse-clean canonical observation, Execute returns released = False with typed rejection code DISCOURSE_CONFLICT or MISSING_GROUNDED_FIELD for every proposal $\hat{A}$. □

Remark 1 (Audited Scope Boundary). *Two regimes lie outside Theorem 1 and are measured empirically in Sections 3.3 and 4: (a) out-of-lexicon clinical entities or implicit timeline transitions lacking triggers in L_kin ∪ L_phase (e.g., 'living donor workup' or 'after stabilization'); and (b) unmarked intra-patient longitudinal trajectories containing multiple admissible draws for f, where note_only_v05 abstains (NONUNIQUE_NAMED_MEASUREMENT) while repair_call_v05 binds the draw proposed by $\hat{A}$.*

Definition 3 (Field-Projected Equivalence). *Given staged decision D_0 = Execute(q, N_{r0}, $\hat{A}$) with calculator c_0, canonical input vector x_0, output unit u_0, and result y_0, a newer revision N_{r1} ($r_1 > r_0$) is projection-equivalent ($N_{r1} \equiv N_{r0}$) iff D_1 = Execute(q, N_{r1}, $\hat{A}$) releases the same calculator c_0 and output unit u_0 with canonical inputs $x_1 = x_0$ and $|y_1 - y_0| \leq 10^{-9} \max(1, |y_0|)$.*

Theorem 2 (Atomic Commit Freshness and Zero Benign False Withholding). *Under SQLite WAL BEGIN IMMEDIATE serialization, publish_semantic acquires the exclusive writer lock before inspecting notes, verifying SHA256(N_cur) = d_cur and committing iff (r_cur, d_cur) = (r_0, d_0) or N_cur ≡ N_{r0}; any concurrent addendum separated by a blank-line paragraph reset outside F_c_0 preserves $x_cur = x_0$ and commits as published_revalidated without false withholding.*

Proof of Theorem 2. Because SQLite WAL serializes BEGIN IMMEDIATE writers, no concurrent transaction can mutate clinical_notes between revision inspection (r_cur, d_cur), SHA-256 verification, and audit insertion. If (r_cur, d_cur) ≠ (r_0, d_0), publish_semantic recomputes D_cur = Execute(q, N_cur, $\hat{A}$) inside the same write lock: any mutation altering a bound span or canonical input in F_c_0 violates N_cur ≡ N_{r0} and aborts with stale_note_conflict, whereas any blank-line-

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 26 November 2026 https://doi.org/10.20944/preprints202611.2737.v1

separated addendum outside F_c_0 leaves all bound spans and paragraph frames in F_c_0 invariant, preserving x_cur = x_0 and committing as published_revalidated. □

2.3. 16-Calculator Registry, High-Precision Canonicalization, and Audit Chain

In expanded_registry.py, MedCalc-Enforce sets 42-digit decimal precision (decimal.getcontext().prec = 42) for unit canonicalization, covering analyte-specific molar conversions (creatinine μmol/L ÷ 88.4 → mg/dL; calcium mmol/L × 4.008 → mg/dL; glucose mmol/L × 18.016 → mg/dL; BUN mmol/L × 2.8 → mg/dL), platelet volume aliases (/mm^3, /uL, ×10^3/μL, ×10^9/L), and output-unit scaling (Bazett QTc in seconds vs. milliseconds). Alongside 10 continuous equations—CKD-EPI 2021 eGFR [6], MAP, BMI, Payne albumin-corrected calcium, Bazett QTc [11], FIB-4 [20], Hillier corrected sodium [21], serum osmolality, anion gap, and Mosteller BSA [7]—the registry enforces 6 discrete/hybrid risk scores: CHA_2DS_2-VASc [22], CURB-65 [23], Wells DVT [24], Glasgow Coma Scale [25], OPTN MELD-Na [26], and Cockcroft–Gault creatinine clearance [27]. Multi-calculator oncology workflows (GFR_abs = eGFR × BSA / 1.73) enforce cryptographic snapshot equality (d_BSA = d_eGFR = d_live), and every state transition appends a SHA-256 hash-chained row protected by SQLite BEFORE UPDATE/DELETE abort triggers.

2.4. Experimental Design, Co-Primary Safety Endpoints, and Clustered Statistical Inference

Table 1 summarizes the five open-weight Q4_K_M GGUF models—spanning Qwen2.5 [28], SmolLM2, Phi-3.5 [29], and Llama 3.1 [30]—evaluated across all 124 held-out official MedCalc-Bench cases (5 × 124 = 620 official runs) and 48 cross-sentence distractor traps (5 × 48 = 240 adversarial runs; 860 total live workflows). Consistent with institutional privacy requirements for local medical NLP inference and recent studies in Computers [31–34], all 860 workflows executed locally via llama.cpp with deterministic greedy decoding (temperature = 0, seed = 731; --no-jinja on Phi-3.5-mini): 292 initial official runs executed on local x86-64 CPU (llama-server build b12211: 124 Qwen2.5-1.5B, 124 SmolLM2-1.7B, and the initial 44 Qwen2.5-7B cases) and 568 runs executed on an NVIDIA Tesla T4 GPU (15 GB VRAM, llama-server build b12428, CUDA 12.8: the 80-case Qwen2.5-7B extension, 124 Phi-3.5-mini, 124 Llama-3.1-8B, and all 5 × 48 = 240 live distractor runs).

Table 1. Pinned open-weight Q4_K_M GGUF model panel evaluated across N = 124 official clinical cases and N = 48 live cross-sentence distractor traps (860 total live executions at $0 API spend).

Model ID	Architecture Family	Params	License	GGUF SHA-256 Prefix
qwen2.5_1.5b_q4km	Qwen2.5-1.5B-Instruct [28]	1.5B	Apache-2.0	6a1a2eb6d15622bf...
smollm2_1.7b_q4km	SmolLM2-1.7B-Instruct	1.7B	Apache-2.0	decd2598bc2c8ed0...
phi3.5_mini_q4km	Phi-3.5-mini-instruct [29]	3.8B	MIT	e4165e3a71af97f1...
qwen2.5_7b_q4km	Qwen2.5-7B-Instruct (2 shards) [28]	7.6B	Apache-2.0	dfce12e3... / 539cf93f...
llama3.1_8b_q4km	Meta-Llama-3.1-8B-Instruct [30]	8.0B	Llama 3.1	7b064f5842bf9532...

We designate Unsafe Numerical Release Rate and Selective-Prediction Risk (AURC) [17] as co-primary safety endpoints alongside Benchmark Reference Agreement. Because 7B–8B models already reach 85.5%–87.1% baseline agreement on N = 124 cases (leaving 12–14 discordant accuracy pairs per model), pooling across the five-model panel provides high statistical power when within-encounter correlation is modeled explicitly. We treat the N = 124 unique patient encounters as primary clusters, computing 10,000-resample patient-case-clustered calculator-stratified paired bootstrap 95% confidence intervals (CIs), exact two-sided cluster-level paired sign tests across the 124

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 26 November 2026 https://doi.org/10.20944/preprints202611.2737.v1

independent clusters, and Holm–Bonferroni / Benjamini–Hochberg (q-value) adjustments across per-model McNemar tests.

3. Results

3.1. Experiment 1: Five-Model Held-Out Safety, Agreement, and Case-Clustered Inference

Auditing canonical ground-truth annotations across all 124 held-out official cases confirmed that expanded_registry.py reconstructs 124/124 (100.0%) reference answers within benchmark tolerance bounds, resolving the official-283 platelet /mm^3 alias gap (vs. 123/124 in v0.4). Across 620 live runs (Table 2, Figure 2), the baseline schema/unit/domain gate (L_2) releases unsafe outputs on 107/620 (17.3%) encounters, whereas pure LLM-free regex (note_only_v05) achieves low unsafe risk (1/124, 0.8%) but stalls at 94/124 (75.8%) agreement by abstaining on 29 multi-draw notes.

Table 2. Official held-out benchmark evaluation across pure LLM-free regex (note_only_v05) and five open-weight models (N = 124 independent patient cases per model; 620 live runs).

Model / Extractor	Enforcement Policy	Matched (%)	Unsafe (%)	Withheld / AURC
No-LLM Regex	note_only_v05 (pure regex)	94 (75.8%)	1 (0.8%)	29 / —
Qwen2.5-1.5B	baseline_unit_domain (L_2)	98 (79.0%)	16 (12.9%)	10 / 0.0627
Qwen2.5-1.5B	reviewed_v04 (clause-local)	109 (87.9%)	2 (1.6%)	13 / —
Qwen2.5-1.5B	**discourse_v05 (MedCalc-Enforce)**	110 (88.7%)	2 (1.6%)	12 / 0.0117
SmolLM2-1.7B	baseline_unit_domain (L_2)	63 (50.8%)	54 (43.5%)	7 / 0.2586
SmolLM2-1.7B	reviewed_v04 (clause-local)	106 (85.5%)	2 (1.6%)	16 / —
SmolLM2-1.7B	**discourse_v05 (MedCalc-Enforce)**	107 (86.3%)	2 (1.6%)	15 / 0.0166
Phi-3.5-mini (3.8B)	baseline_unit_domain (L_2)	102 (82.3%)	16 (12.9%)	6 / 0.0797
Phi-3.5-mini (3.8B)	reviewed_v04 (clause-local)	110 (88.7%)	3 (2.4%)	11 / —
Phi-3.5-mini (3.8B)	**discourse_v05 (MedCalc-Enforce)**	110 (88.7%)	4 (3.2%)	10 / 0.0304
Qwen2.5-7B	baseline_unit_domain (L_2)	108 (87.1%)	11 (8.9%)	5 / 0.0610
Qwen2.5-7B	reviewed_v04 (clause-local)	112 (90.3%)	1 (0.8%)	11 / —
Qwen2.5-7B	**discourse_v05 (MedCalc-Enforce)**	114 (91.9%)	1 (0.8%)	9 / 0.0006
Llama-3.1-8B	baseline_unit_domain (L_2)	106 (85.5%)	10 (8.1%)	8 / 0.0232
Llama-3.1-8B	reviewed_v04 (clause-local)	111 (89.5%)	2 (1.6%)	11 / —
Llama-3.1-8B	**discourse_v05 (MedCalc-Enforce)**	113 (91.1%)	2 (1.6%)	9 / 0.0106
Pooled (N = 620)	**baseline_unit_domain (L_2)**	**477 (76.9%)**	**107 (17.3%)**	**36 / —**
Pooled (N = 620)	**discourse_v05 (MedCalc-Enforce)**	**554 (89.4%)**	**11 (1.8%)**	**55 / —**

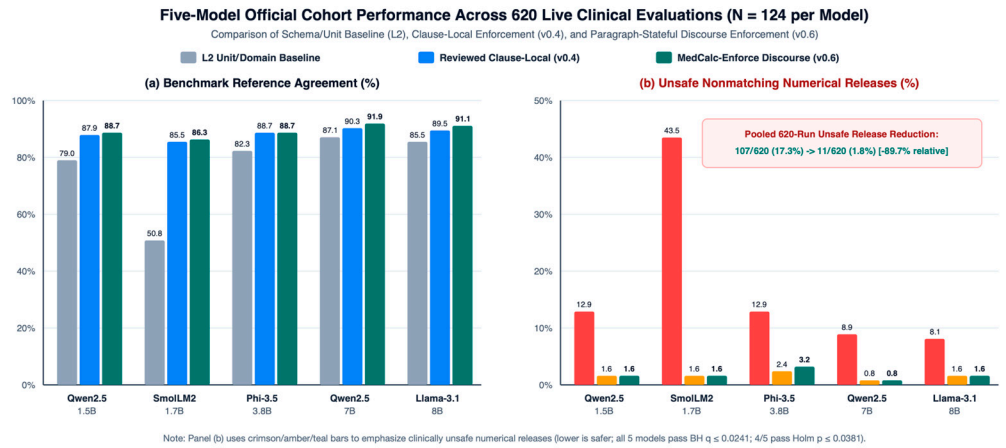


Figure 2. Five-model official cohort performance (N = 124 per model, 620 total runs): (a) benchmark reference agreement (%) and (b) co-primary unsafe nonmatching numerical release rate (%).

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 26 November 2026 https://doi.org/10.20944/preprints202611.2737.v1

MedCalc-Enforce lifts pooled agreement to 554/620 (89.4%; case-clustered 95% CI: 84.5%–93.9%)—peaking at 91.9% on Qwen2.5-7B and 91.1% on Llama-3.1-8B—while cutting unsafe outputs by 89.7% to 11/620 (1.8%; case-clustered 95% CI: 0.3%–3.7%). Safety reductions pass Benjamini–Hochberg FDR control across all 5/5 models ($q \leq 0.0241$) and Holm–Bonferroni family-wise control on 4/5 models ($p_{Holm} \leq 0.0381$; Llama-3.1-8B: $q = 0.0241$, $p_{Holm} = 0.0859$; Table 3). Across the N = 124 independent patient clusters, case-clustered Δ Agreement is +12.4% (+7.1% to +17.4%; 62 improved vs. 7 worsened clusters, $p = 4.10 \times 10^{-12}$) and Δ Unsafe is −15.5% (−18.5% to −12.3%; 62 improved vs. 2 worsened clusters, $p = 2.26 \times 10^{-16}$). Selective-prediction AURC drops up to 99.0% (0.0610 → 0.0006 on Qwen2.5-7B).

Table 3. Per-model bootstrap comparisons (N = 124) and pooled patient-case-clustered inference (620 runs across 124 clusters) comparing L_2 vs. MedCalc-Enforce.

Cohort / Model	Δ Agreement (95% CI)	Acc. q / Cluster p	Unsafe Elim./Add	Safety q (Holm / Cluster p)
Qwen2.5-1.5B	+9.7% (+3.2% to +16.1%)	q = 0.0850	15 / 1	0.0011 (0.0054)
SmolLM2-1.7B	+35.5% (+27.4% to +43.5%)	$q = 1.5 \times 10^{-8}$	52 / 0	$<10^{-10}$ ($<10^{-10}$)
Phi-3.5-mini	+6.5% (+1.6% to +11.3%)	q = 0.2063	13 / 1	0.0034 (0.0147)
Qwen2.5-7B	+4.8% (−0.8% to +10.5%)	q = 0.3589	11 / 1	0.0087 (0.0381)
Llama-3.1-8B	+5.6% (+0.0% to +11.3%)	q = 0.2839	9 / 1	0.0241 (0.0859)
Pooled Clustered (620)	**+12.4% (+7.1% to +17.4%)**	**$p = 4.10 \times 10^{-12}$**	**100 / 4 (−15.5%)**	**$<10^{-15}$ (2.26×10^{-16})**

3.2. Provenance-Tier Attribution and Bounded Diagnostic Repair

To verify that high agreement stems from LLM extraction rather than silent regex takeover, Table 4 decomposes all 565 released official runs across provenance tiers. Pooled across all five models, 529/565 (93.6%) of releases originate directly from verified LLM proposals—449 at Tier 1.0 (exact canonical match; 98.7% precision) and 80 at Tier 0.85 (unit/span repair; 95.0% precision)—while only 36/565 (6.4%) invoke Tier 0.70 named fallback (31 on 1.7B SmolLM2). On 3.8B–8B models, verified LLM proposals drive 99.1% of releases on Phi-3.5-mini and Qwen2.5-7B and 100.0% (115/115; zero fallback calls) on Llama-3.1-8B, rescuing +13 to +20 multi-draw notes per model where pure regex abstains (0 losses; McNemar $p \leq 2.44 \times 10^{-4}$; case-clustered Δ = +13.5%). On the 120 valid discrete/hybrid clinical risk trials, pure regex matches 0/120 (0.0%) vs. 120/120 (100.0%) with structured proposals.

Table 4. Provenance-tier attribution of all 565 released official runs under MedCalc-Enforce and paired rescue over pure regex (note_only_v05: 94/124 matched).

Model (N = 124)	Tier 1.0 / Tier 0.85 LLM	Tier 0.70 Fallback	LLM Proposal Share	Net Rescue vs. Regex (p)
Qwen2.5-1.5B	92 (91 ok) / 17 (16 ok)	3 (3 ok)	109/112 (97.3%)	+16 / 0 loss (3.05×10^{-5})
SmolLM2-1.7B	57 (56 ok) / 21 (21 ok)	31 (30 ok)	78/109 (71.6%)	+13 / 0 loss (2.44×10^{-4})
Phi-3.5-mini	100 (97 ok) / 13 (12 ok)	1 (1 ok)	113/114 (99.1%)	+16 / 0 loss (3.05×10^{-5})
Qwen2.5-7B	101 (101 ok) / 13 (12 ok)	1 (1 ok)	114/115 (99.1%)	+20 / 0 loss (1.91×10^{-6})
Llama-3.1-8B	99 (98 ok) / 16 (15 ok)	0 (0 ok)	115/115 (100.0%)	+19 / 0 loss (3.81×10^{-6})
Pooled (620 Runs)	**449 (98.7%) / 80 (95.0%)**	**36 (97.2%)**	**529/565 (93.6%)**	**+84 (+13.5%, 1.91×10^{-6})**

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 26 November 2026 https://doi.org/10.20944/preprints202611.2737.v1

In Experiment 2 (fallback disabled, Table 5), we compare the deterministic ungrounded-critic baseline (simulate_ungrounded_critic) against bounded structured diagnostic repair (execute_bounded_repair, k = 1), where gating a single retry (k = 1) strictly on typed rejection codes recovers 31 failed SmolLM2-1.7B proposals (77 → 108 matches without fallback), 3 Qwen2.5-1.5B calls, and 1 call each on Qwen2.5-7B and Phi-3.5-mini with 0 valid-call regressions across all 620 trajectories, whereas an ungrounded critic degrades Qwen2.5-7B to 85/124 (68.5%).

Table 5. Experiment 2: Bounded structured diagnostic repair (k = 1, zero fallback) vs. ungrounded critic and strict k = 0 across N = 124 official cases.

Model (N = 124)	Ungrounded Critic	Strict k = 0	Repair k = 1	Repaired / Regressed
Qwen2.5-1.5B	95 match / 3 unsafe	107 / 2	110 / 2	3 / 0
SmolLM2-1.7B	63 match / 14 unsafe	77 / 1	108 / 2	31 / 0
Phi-3.5-mini	102 match / 7 unsafe	109 / 4	110 / 4	1 / 0
Qwen2.5-7B	85 match / 2 unsafe	113 / 1	114 / 1	1 / 0
Llama-3.1-8B	106 match / 2 unsafe	113 / 2	113 / 2	0 / 0

3.3. Paraphrases, Live Multi-Sentence Traps, and Zero-Tuning Out-of-Vocabulary Discourse Transfer

In Experiment 3, evaluating 27 fixed clinical question templates (P = 3 distinct phrasings across 9 calculators, yielding 372 evaluations across the 124 official notes, including Bazett QTc in seconds, replaying archived Qwen2.5-1.5B proposals), clause-local v0.4 matches 187/372 (50.3%) with 41 nonmatching outputs, whereas MedCalc-Enforce matches 316/372 (84.9%) with 7 nonmatching outputs ($p < 10^{-12}$). Prepending a family-history paragraph followed by a blank-line break before all 124 official notes preserves 110/124 (88.7%) agreement. Across Experiment 8's 240 live adversarial workflows on 48 cross-sentence family and wrong-timepoint traps (Table 6, Figure 3b), L_2 falsely released 217/240 (90.4%) and v0.4 released 240/240 (100.0%), whereas MedCalc-Enforce intercepted 240/240 (0.0% false releases; Holm-adjusted $p \leq 5.96 \times 10^{-8}$, BH $q \leq 5.96 \times 10^{-8}$).

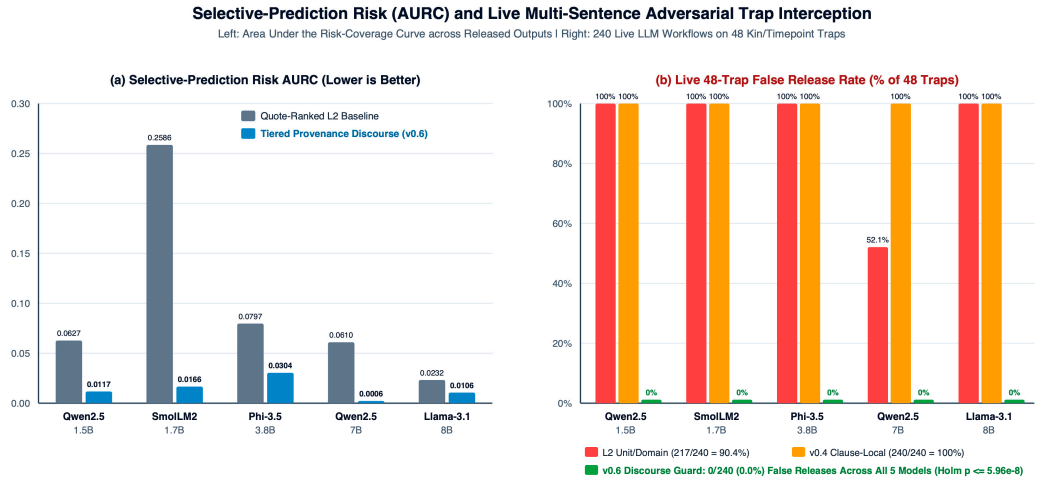


Figure 3. (a) Selective-prediction risk (AURC) on the 124-case official cohort. (b) Live 5-model adversarial evaluation across 48 cross-sentence kin and wrong-timepoint traps (240 live workflows).

Table 6. Experiment 8: Live 5-model inference across 48 cross-sentence family-attribution and wrong-timepoint clinical traps (240 live LLM workflows).

Model (48 Traps)	L_2 False Releases	v0.4 False Releases	MedCalc-Enforce	Holm p (BH q)
Qwen2.5-1.5B	48 (100.0%)	48 (100.0%)	0 (0.0%)	3.55×10^{-14} (8.88×10^{-15})
SmolLM2-1.7B	48 (100.0%)	48 (100.0%)	0 (0.0%)	3.55×10^{-14} (8.88×10^{-15})

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 26 November 2026 https://doi.org/10.20944/preprints202611.2737.v1

Model (48 Traps)	L_2 False Releases	v0.4 False Releases	MedCalc-Enforce	Holm p (BH q)
Phi-3.5-mini	48 (100.0%)	48 (100.0%)	0 (0.0%)	3.55×10^{-14} (8.88×10^{-15})
Qwen2.5-7B	25 (52.1%)	48 (100.0%)	0 (0.0%)	5.96×10^{-8} (5.96×10^{-8})
Llama-3.1-8B	48 (100.0%)	48 (100.0%)	0 (0.0%)	3.55×10^{-14} (8.88×10^{-15})
Total (240 Runs)	**217 (90.4%)**	**240 (100.0%)**	**0 (0.0%)**	—

Zero-Tuning Out-of-Vocabulary (OOV) Clinical Discourse Transfer (N = 64). To eliminate rule-to-benchmark circularity, we froze discourse_guard.py (SHA-256: dc74af9b...) and release_policy.py (SHA-256: ba74626e...) with zero rule edits and evaluated an independent 64-case OOV clinical transfer suite (data/oov_discourse_transfer_64.jsonl; Table 7). Across Stratum A (24 unseen clinical surface forms covering collateral H&P notes, ICU post-treatment sign-outs, hospital-day-4/6 progress notes, dry-weight goals, and dysphagia weight-loss deltas), L_2 falsely released 23/24 (95.8%) and v0.4 released 13/24 (54.2%), whereas frozen MedCalc-Enforce released 0/24 (100.0% zero-shot interception). Across Stratum B (8 intentionally out-of-lexicon probes: living kidney donor, roommate, field EMS, maternal uncle, newborn infant, 2-year-old wellness panel, adjacent-bed specimen mix-up, and manikin simulation), all 8 bypassed lexical triggers as predicted by Remark 1, yielding an honest 24/32 (75.0%; 95% Wilson CI: 57.9%–86.7%) overall OOV interception rate vs. 3.1% for L_2 ($p = 2.38 \times 10^{-7}$) and 34.4% for v0.4 ($p = 2.44 \times 10^{-4}$). On Stratum C (32 unseen multi-section valid patient charts with SI molar panels), frozen MedCalc-Enforce matched 31/32 (96.9%; 0 unsafe) vs. 19/32 (13 unsafe) under L_2 and 14/32 (12 unsafe) under v0.4 ($p = 1.53 \times 10^{-5}$).

Table 7. Zero-tuning out-of-vocabulary (OOV) clinical discourse transfer audit (N = 64 disjoint notes evaluated on byte-frozen discourse_guard.py).

OOV Stratum (Cases)	Clinical Scenario Scope	L_2 Baseline	Clause-Local v0.4	Frozen MedCalc-Enforce
Stratum A (24 traps)	Unseen collateral, ICU sign-out, HD 4–6, goals	23 (95.8%) false rel.	13 (54.2%) false rel.	0 (0.0%) false rel.
Stratum B (8 traps)	Out-of-lexicon (donor, EMS, uncle, neonate)	8 (100.0%) false rel.	8 (100.0%) false rel.	8 (100.0%) false rel.
Pooled Traps (32)	**All OOV non-patient / phase traps**	**31 (96.9%) false rel.**	**21 (65.6%) false rel.**	**8 (25.0%; 75.0% blocked)**
Stratum C (32 valid)	Multi-section valid charts + SI molar units	19 match / 13 unsafe	14 match / 12 unsafe	31 (96.9%) match / 0 unsafe

3.4. Experiments 4–7: Falsification Controls, Discrete/Chained Scores, Concurrency, and Scalability

All three falsification controls confirmed active invariant enforcement: Control A (117 cross-patient note–proposal permutations) fooled L_2 on 114/117 (97.4%) vs. 0/117 (0.0%) under MedCalc-Enforce; Control B (45 perturbed molar conversions) collapsed agreement from 45/45 (100.0%) to 3/45 (6.7%); and Control C (24 polarity-inverted headers) flipped releases from 24/24 (100.0%) to 0/24 (0.0%). Across the 6 discrete/hybrid risk calculators (240 trials), MedCalc-Enforce released 120/120 valid calls and blocked 120/120 fault-injected calls, while verifying 36/36 valid and blocking 36/36 weight-tampered carboplatin BSA + eGFR chains.

Across 324 two-connection SQLite WAL concurrency trials (108 per policy; Table 8, Figure 4a), unchecked read_then_write published 24/108 stale/unsafe results, and strict_revision_lock suffered 24/108 benign false withholdings. Field-projected semantic_field_commit_v05 achieved 0/108 unsafe commits and 0/108 benign false withholdings (rescuing all 24 benign updates), with 108/108 intact SHA-256 audit chains. Across the 17-cell scalability matrix (Table 9, Figure 4b), single-worker throughput reached 2,881.8 checks/s ($p_{50} = 0.357$ ms) at 1,000 notes and sustained 1,852.0 checks/s ($p_{50} = 11.118$ ms, $p_{95} = 28.969$ ms) at 100,000 notes under 32 workers (309.94 MB per 10^6 audit events).

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 26 November 2026 https://doi.org/10.20944/preprints202611.2737.v1

Table 8. Experiment 6: Three-policy SQLite WAL concurrency evaluation across 9 edit schedules × 12 repeats (108 trials per policy; 72 safe publishable states).

Commit Policy (108 Trials)	Published	Stale / Unsafe	Benign False Blocks	Revalidated Saves
read_then_write (unchecked)	96	24	0	0
strict_revision_lock (v0.4)	48	0	24	0
semantic_field_commit_v05	72	0	0	24

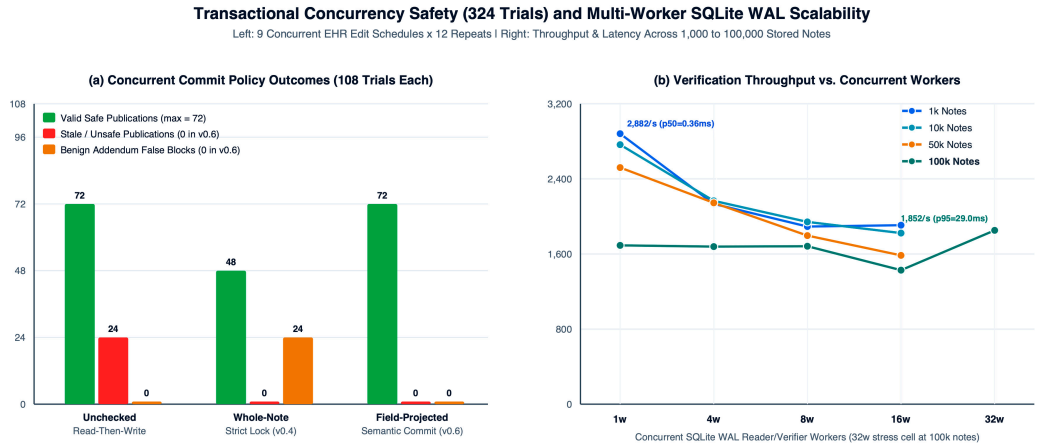


Figure 4. (a) Concurrent SQLite WAL commit outcomes across 108 trials per policy. (b) Verification throughput across {1k, 10k, 50k, 100k} stored notes and {1, 4, 8, 16, 32} concurrent workers.

Table 9. Selected cells from the 17-cell SQLite WAL scalability matrix (800 verified operations per cell).

Workload (Notes × Workers)	Throughput (checks/s)	p_{50} (ms)	p_{95} (ms)	p_{99} (ms)
1,000 notes × 1 worker	2,881.8	0.357	0.419	0.465
10,000 notes × 1 worker	2,764.9	0.363	0.467	0.564
50,000 notes × 4 workers	2,145.0	0.932	5.854	11.084
100,000 notes × 1 worker	1,691.8	0.595	0.733	0.921
100,000 notes × 16 workers	1,427.7	0.795	45.438	65.963
100,000 notes × 32 workers	1,852.0	11.118	28.969	39.944

4. Discussion

4.1. Why Parameter Scaling and Pure Regex Each Fall Short Alone

Comparing SmolLM2-1.7B through Llama-3.1-8B reveals an instructive divide: parameter scaling cuts baseline L_2 errors on single-patient notes from 43.5% to 8.1%, yet 8B Llama-3.1 and 3.8B Phi-3.5-mini still exhibit a 100.0% (48/48) false-release rate on cross-sentence family and wrong-timepoint traps when grammar-constrained to emit JSON calculator calls. Conversely, pure regex (note_only_v05) stalls at 75.8% on continuous equations and 0.0% on discrete clinical scores, whereas combining span-grounded LLM proposals (99.1%–100.0% of releases on 3.8B–8B models) with stateful discourse verification lifts agreement up to 91.9% [2, 14].

4.2. Case-Level Clinical EHR Provenance Walkthrough and Residual Root-Cause Taxonomy

Figure 5 illustrates how MedCalc-Enforce operates at the span and transaction level across three representative benchmark records: (a) a multi-paragraph family-history record (oov-ctrl-30, CKD-EPI 2021 eGFR) where a sibling framing cue in Paragraph 1 propagates η_eff = NON_PATIENT to block a non-patient lab span before resetting cleanly at the blank-line break for the patient's own labs in Paragraph 2; (b) an SI molar metabolic panel (oov-ctrl-14, Hillier glycemic sodium correction)

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 26 November 2026 https://doi.org/10.20944/preprints202611.2737.v1

where molar concentration in mmol/L is canonicalized to mg/dL at 42-digit precision and committed as published_revalidated without false withholding when a concurrent addendum arrives under SQLite WAL; and (c) a longitudinal multi-draw record (official-397, Remark 1) illustrating the exact scope boundary when an unmarked post-stabilization repeat draw coexists with an abbreviated admission measurement.

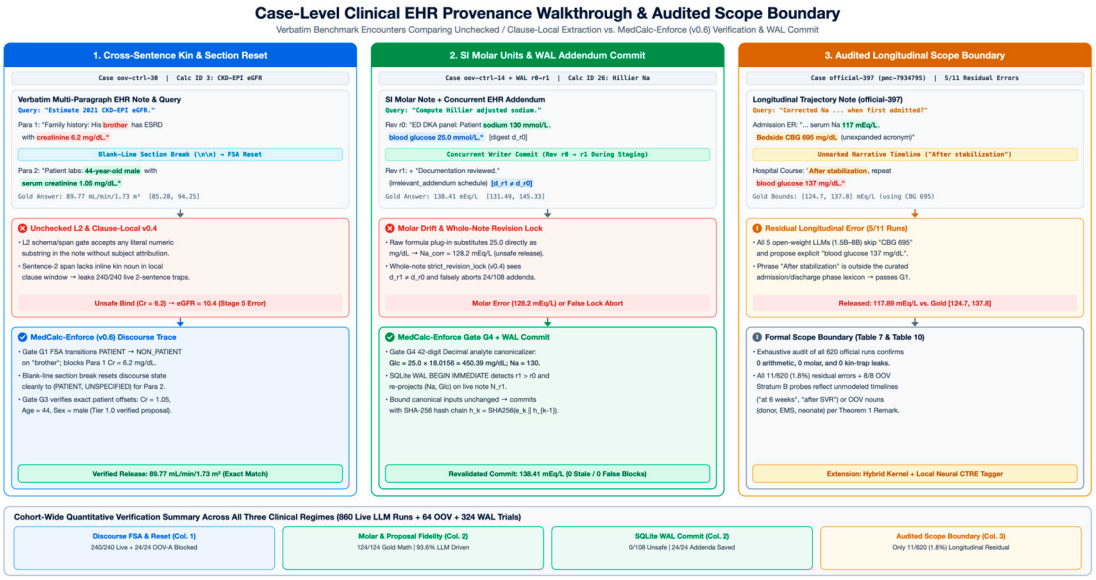


Figure 5. Case-level clinical EHR provenance and commit walkthrough across three audited records: (a) cross-sentence relative framing interception and blank-line section reset (oov-ctrl-30, CKD-EPI 2021 eGFR); (b) exact SI molar-to-mass canonicalization and field-projected concurrent SQLite WAL revalidation (oov-ctrl-14, Hillier sodium correction); and (c) audited longitudinal timeline boundary on unmarked post-stabilization draws (official-397, Remark 1).

Auditing all 11/620 (1.8%) residual nonmatching outputs across the five models (Table 10) confirms that 0/11 stem from arithmetic drift, molar unit misconversion, fabricated numbers, or kin misattribution. All 11/11 arise on 6 longitudinal notes containing both an admission draw and an unmarked post-therapy trajectory, including official-397 (5/11 across all models, where admission glucose was abbreviated as 'Bedside CBG 695 mg/dL' while recovery wrote 'blood glucose 137 mg/dL' under 'After stabilization') and official-299 (explaining the exact 3 → 4 unsafe shift on Phi-3.5-mini between v0.4 unanswered and v0.5 when expanded k/μL support validated Phi-3.5-mini's ICU day-3 nadir platelet binding). Recording transport transcripts in invocations.jsonl also uncovered how llama-server's Jinja processor split Phi-3.5-mini's <|assistant|> role tag into assistant|>, breaking GBNF sampler initialization until --no-jinja isolated template rendering.

Table 10. Exhaustive root-cause taxonomy of all 11/620 (1.8%) residual nonmatching outputs across the five open-weight models (6 unique longitudinal cases).

Case ID (Calc)	Models (Count)	Out / Gold Bounds	Bound vs. Gold Span	Clinical Root Cause
official-397 (Na)	All 5 models (5/11)	117.89 / [124.7, 137.8]	Glu 137 vs. CBG 695	CBG acronym + 'After stabilization'
official-119 (Ca)	Qwen-1.5B, Phi-3.5 (2/11)	8.10 / [8.99, 9.93]	Alb 3 vs. 1.3 g/dL	'day of discharge' >64 chars past span
official-299 (FIB-4)	Phi-3.5-mini (1/11)	20.53 / [3.88, 4.28]	Plt 33 vs. 166 k/μL	ICU 'third day' nadir (v0.4 abstained)
official-290 (FIB-4)	SmolLM2-1.7B (1/11)	2.76 / [6.33, 6.99]	Plt 256k vs. 106.3k	Post-therapy 'after 8 years of SVR'

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 26 November 2026 https://doi.org/10.20944/preprints202611.2737.v1

Case ID (Calc)	Models (Count)	Out / Gold Bounds	Bound vs. Gold Span	Clinical Root Cause
official-23 (eGFR)	Phi-3.5-mini (1/11)	69.88 / [46.0, 50.9]	Cr 0.98 vs. 1.33	Post-hydration/calcitonin draw
official-31 (eGFR)	Llama-3.1-8B (1/11)	52.99 / [26.7, 29.5]	Cr 1.77 vs. 3 mg/dL	'at 6 weeks' post-chemotherapy draw

4.3. Architectural Implications, Limitations, and Future Work

Viewed alongside recent Computers investigations into clinical rationale-prediction alignment [31], auditable healthcare data pipelines [32], local quantized transformer architectures [33], and longitudinal EHR temporal benchmarks [34], our findings indicate that deterministic provenance state machines and field-projected database serialization provide high-leverage safety guarantees for small on-premise models without incurring secondary LLM critic latency. Three limitations nonetheless guide future work. First, as shown by Stratum B of our OOV audit (Table 7) and the 11 longitudinal errors (Table 10), curated lexical triggers cannot intercept out-of-lexicon entities (donor, neonate, specimen mix-up) or implicit treatment timelines ('after stabilization', 'at 6 weeks'); pairing our deterministic stoichiometric and commit kernel with lightweight local temporal/experiencer taggers [15, 16] is a high-priority extension. Second, prospective validation of the 6 discrete risk scores and 72 chained oncology workflows on multi-institution FHIR streams with clinician adjudication remains essential. Third, distributed hospital deployments should map field-projected revalidation onto row-level serializable enterprise relational databases.

5. Conclusions

MedCalc-Enforce closes the provenance and concurrency safety gap left open by syntactic JSON tool calling. Across 860 live open-weight LLM workflows (1.5B–8B) at $0 API spend, MedCalc-Enforce lifted held-out agreement to 89.4% (with 93.6%–100.0% of releases driven by verified LLM proposals), eliminated 89.7% of unsafe numerical releases, intercepted 240/240 live multi-sentence traps and 24/24 unseen surface-form OOV traps, and prevented both stale writes and benign-addendum false blocks under concurrent SQLite WAL load at sub-millisecond single-worker latency.

Supplementary Materials: The complete replication archive MedCalc_Enforce_Experiments_v0_6.zip (295 files comprising 294 SHA-256-verified entries plus SHA256SUMS.txt, wheel medcalc_enforce_research-0.6.0-py3-none-any.whl, all 860 live LLM traces, the 64-case OOV transfer suite, and 75 unit tests) accompanies this manuscript, is mirrored at https://github.com/ai-developers33/medcalc-enforce/releases/tag/v0.6.0 (release v0.6.0, commit ae35863; repository main commit 403ed84 with reproduce_all.py and interactive Colab T4 notebook at https://colab.research.google.com/github/ai-developers33/medcalc-enforce/blob/main/MedCalc_Enforce_v0_6_Colab_Live_Reproduction.ipynb), and is permanently archived on Zenodo at https://doi.org/10.5281/zenodo.24293301.

Author Contributions: Conceptualization, R.S.T.; methodology, R.S.T.; software, R.S.T.; validation, R.S.T. and A.N.; formal analysis, R.S.T. and A.N.; investigation, R.S.T.; data curation, R.S.T. and A.N.; writing—original draft preparation, R.S.T.; writing—review and editing, R.S.T. and A.N.; visualization, R.S.T. and A.N.; supervision, R.S.T. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable; only publicly available de-identified benchmark records and synthetic concurrency/adversarial test vectors were analyzed.

Informed Consent Statement: Not applicable.

Data Availability Statement: All source code, locked cohort manifests, raw llama.cpp inference logs, OOV clinical transfer suites, statistical evaluation scripts, and the Google Colab T4 GPU reproduction notebook (https://colab.research.google.com/github/ai-developers33/medcalc-

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 26 November 2026 https://doi.org/10.20944/preprints202611.2737.v1

enforce/blob/main/MedCalc_Enforce_v0_6_Colab_Live_Reproduction.ipynb) are openly available in the public repository at https://github.com/ai-developers33/medcalc-enforce (main commit 403ed84; tagged release v0.6.0 at commit ae35863) and permanently archived on Zenodo under DOI 10.5281/zenodo.24293301 (https://doi.org/10.5281/zenodo.24293301, hosting the 295-file replication archive with SHA-256 c3121032ea6eaf6f25d31cb230d67c2eb37e5030e8026d28d9000a9e06e6e6cb plus 4 root metadata files). Original MedCalc-Enforce software and synthetic evaluation suites are released under the MIT License, while derived MedCalc-Bench benchmark adaptations comply with Creative Commons Attribution-ShareAlike 4.0 (https://creativecommons.org/licenses/by-sa/4.0/). Executing python3 reproduce_all.py verifies all 294 file checksums, 75 unit tests, locked transcripts, OOV transfer audits, and case-clustered statistical invariants offline using only the Python standard library.

Acknowledgments: The authors thank the open-source maintainers of MedCalc-Bench and llama.cpp for making public clinical benchmarks and local quantized inference runtimes openly accessible.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AURC	Area Under the Risk–Coverage Curve
BH	Benjamini–Hochberg
BMI	Body Mass Index
BSA	Body Surface Area
BUN	Blood Urea Nitrogen
CBG	Capillary Blood Glucose
CI	Confidence Interval
CrCl	Creatinine Clearance
CTRE	Clinical Temporal Relation Extraction
DVT	Deep Vein Thrombosis
eGFR	Estimated Glomerular Filtration Rate
EHR	Electronic Health Record
EMS	Emergency Medical Services
FDR	False Discovery Rate
GBNF	GGML Backus–Naur Form
GCS	Glasgow Coma Scale
LLM	Large Language Model
MAP	Mean Arterial Pressure
MELD-Na	Model for End-Stage Liver Disease–Sodium
OOV	Out-of-Vocabulary
SVR	Sustained Virologic Response
WAL	Write-Ahead Logging

References

1. Singhal, K.; Azizi, S.; Tu, T.; Mahdavi, S.S.; Wei, J.; Chung, H.W.; Scales, N.; Tanwani, A.; Cole-Lewis, H.; Pfohl, S.; et al. *Large Language Models Encode Clinical Knowledge. Nature* **2023**, *620*, 172–180. https://pubmed.ncbi.nlm.nih.gov/37438534/
2. Hager, P.; Jungmann, F.; Holland, R.; Bhagat, K.; Hubrecht, I.; Knauer, M.; Vielhauer, J.; Makowski, M.; Braren, R.; Kaissis, G.; Rueckert, D. *Evaluation and Mitigation of the Limitations of Large Language Models in Clinical Decision-Making. Nat. Med.* **2024**, *30*, 2613–2622. https://pubmed.ncbi.nlm.nih.gov/38965432/
3. Khandekar, N.; Jin, Q.; Xiong, G.; Lu, Z.; et al. MedCalc-Bench: Evaluating Large Language Models for Medical Calculations. In Advances in Neural Information Processing Systems (NeurIPS Datasets and Benchmarks); 2024; Volume 37. https://arxiv.org/abs/2406.12036
4. Yao, S.; Zhao, J.; Yu, D.; Du, N.; Narasimhan, K.; Cao, Y. ReAct: Synergizing Reasoning and Acting in Language Models. In Proceedings of the Eleventh International Conference on Learning Representations (ICLR); 2023.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 26 November 2026 https://doi.org/10.20944/preprints202611.2737.v1

5. Schick, T.; Dwivedi-Yu, J.; Dessì, R.; Raileanu, R.; Lomeli, M.; Scialom, T.; et al. Toolformer: Language Models Can Teach Themselves to Use Tools. In Advances in Neural Information Processing Systems (NeurIPS); 2023; Volume 36, pp. 68539–68551.
6. Inker, L.A.; Eneanya, N.D.; Coresh, J.; Levey, A.S.; et al. *New Creatinine- and Cystatin C-Based Equations to Estimate GFR without Race. N. Engl. J. Med.* **2021**, *385*, 1737–1749.
7. Mosteller, R.D. *Simplified Calculation of Body-Surface Area. N. Engl. J. Med.* **1987**, *317*, 1098.
8. Willard, B.T.; Louf, R. Efficient Guided Generation for Large Language Models. *arXiv* **2023**, arXiv:2307.09702.
9. Madaan, A.; Tandon, N.; Gupta, P.; Welleck, S.; Clark, P.; et al. Self-Refine: Iterative Refinement with Self-Feedback. In Advances in Neural Information Processing Systems (NeurIPS); 2023; Volume 36, pp. 46534–46594.
10. Rebedea, T.; Dinu, R.; Sreedhar, M.N.; Parisien, C.; Cohen, J. NeMo Guardrails: A Toolkit for Controllable and Safe LLM Applications with Programmable Rails. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP): System Demonstrations; Association for Computational Linguistics: Singapore, 2023; pp. 431–445. https://aclanthology.org/2023.emnlp-demo.40/
11. Bazett, H.C. *An Analysis of the Time-Relations of Electrocardiograms. Heart* **1920**, *7*, 353–370.
12. Goh, E.; Gallo, R.; Hom, J.; Strong, E.; Weng, Y.; Kerman, H.; Cool, J.A.; Kanjee, Z.; Parsons, A.S.; Ahuja, N.; et al. Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial. *JAMA Netw. Open* **2024**, *7*, e2440969. https://pubmed.ncbi.nlm.nih.gov/39466245/
13. Tam, T.Y.C.; Sivarajkumar, S.; Kapoor, S.; Stolyar, A.V.; Polanska, K.; McCarthy, K.R.; Osterhoudt, H.; Wu, X.; Visweswaran, S.; Fu, S.; et al. *A Framework for Human Evaluation of Large Language Models in Healthcare Derived from Literature Review. npj Digit. Med.* **2024**, *7*, 258. https://pubmed.ncbi.nlm.nih.gov/39333376/
14. Arora, R.K.; Wei, J.; Hicks, R.S.; Bowman, P.; Quiñonero-Candela, J.; Tsimpourlas, F.; Sharman, M.; Shah, M.; Vallone, A.; Beutel, A.; Heidecke, J.; Singhal, K. HealthBench: Evaluating Large Language Models Towards Improved Human Health. *arXiv* **2025**, arXiv:2505.08775. https://arxiv.org/abs/2505.08775
15. Harkema, H.; Dowling, J.N.; Thornblade, T.; Chapman, W.W. ConText: An Algorithm for Determining Negation, Experiencer, and Temporal Status from Clinical Reports. *J. Biomed. Inform.* **2009**, *42*, 839–851. https://pubmed.ncbi.nlm.nih.gov/19435614/
16. He, J.; Sun, Z.; Rasmy, L.; Li, H.; Li, J.; Yu, E.; Zhi, D.; Tao, C. Optimizing Clinical Temporal Relation Extraction With Large Language Models: Comparative Analysis. *JMIR Med. Inform.* **2026**, *14*, e89957. https://medinform.jmir.org/2026/1/e89957
17. Geifman, Y.; El-Yaniv, R. Selective Classification for Deep Neural Networks. In Advances in Neural Information Processing Systems (NeurIPS); 2017; Volume 30. https://arxiv.org/abs/1705.08500
18. Quach, V.; Fisch, A.; Schuster, T.; Yala, A.; Sohn, J.H.; Jaakkola, T.S.; Barzilay, R. Conformal Language Modeling. In Proceedings of the Twelfth International Conference on Learning Representations (ICLR); 2024. https://arxiv.org/abs/2306.10193
19. Berenson, H.; Bernstein, P.; Gray, J.; Melton, J.; O'Neil, E.; O'Neil, P. A Critique of ANSI SQL Isolation Levels. In Proceedings of ACM SIGMOD; 1995; pp. 1–10.
20. Sterling, R.K.; Lissen, E.; Clumeck, N.; Torriani, F.J.; et al. Development of a Simple Noninvasive Index to Predict Significant Fibrosis in Patients with HIV/HCV Coinfection. *Hepatology* **2006**, *43*, 1317–1325.
21. Hillier, T.A.; Abbott, R.D.; Barrett, E.J. Hyponatremia: Evaluating the Correction Factor for Hyperglycemia. *Am. J. Med.* **1999**, *106*, 399–403.
22. Lip, G.Y.H.; Nieuwlaat, R.; Pisters, R.; Lane, D.A.; Crijns, H.J.G.M. *Refining Clinical Risk Stratification for Predicting Stroke and Thromboembolism in Atrial Fibrillation. Chest* **2010**, *137*, 263–272.
23. Lim, W.S.; van der Eerden, M.M.; Laing, R.; Macfarlane, J.T.; et al. *Defining Community Acquired Pneumonia Severity on Presentation to Hospital. Thorax* **2003**, *58*, 377–382.
24. Wells, P.S.; Anderson, D.R.; Rodger, M.; Kovacs, M.J.; et al. *Evaluation of D-Dimer in the Diagnosis of Suspected Deep-Vein Thrombosis. N. Engl. J. Med.* **2003**, *349*, 1227–1235.
25. Teasdale, G.; Jennett, B. Assessment of Coma and Impaired Consciousness: A Practical Scale. *Lancet* **1974**, *304*, 81–84.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org | NOT PEER-REVIEWED | Posted: 26 November 2026 https://doi.org/10.20944/preprints202611.2737.v1

26. Kim, W.R.; Biggins, S.W.; Kremers, W.K.; Therneau, T.M.; et al. *Hyponatremia and Mortality among Patients on the Liver-Transplant Waiting List. N. Engl. J. Med.* **2008**, *359*, 1018–1026.
27. Cockcroft, D.W.; Gault, M.H. *Prediction of Creatinine Clearance from Serum Creatinine. Nephron* **1976**, *16*, 31–41.
28. Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; et al. Qwen2.5 Technical Report. *arXiv* **2024**, arXiv:2412.15115. https://arxiv.org/pdf/2412.15115
29. Abdin, M.; Jacobs, S.A.; Awan, A.A.; Aneja, J.; Awadallah, A.; Awadalla, H.; Bach, N.; Bahree, A.; Bakhtiari, A.; Bao, J.; et al. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. *arXiv* **2024**, arXiv:2404.14219.
30. Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. The Llama 3 Herd of Models. *arXiv* **2024**, arXiv:2407.21783.
31. Hasan, H.M.Q.; Babiker, H.K.B.; Kim, M.-Y.; Goebel, R. Reason2Decide-C: Adaptive Cycle-Consistent Training for Clinical Rationales. *Computers* **2026**, *15*, 279. https://www.mdpi.com/2073-431x/15/5/279
32. Igweh, K.; Sultana, A.; Bruni, T.; Lin, W. *An Auditable AI Pipeline for Long-Tail Healthcare Resource Discovery and Directory Maintenance. Computers* **2026**, *15*, 824. https://www.mdpi.com/2073-431x/15/11/824
33. Reimer, L.M.; Pries, L.; Schweizer, F.; Nissen, L.; Jonas, S.M. *On-Device Transformer Architectures for Speech Evaluation in Neurodegenerative Disease Detection. Computers* **2026**, *15*, 287. https://www.mdpi.com/2073-431x/15/5/287
34. Tran, Q.A.Q.; To, H.Q.; Sheng, Y.; Liu, M. EHRForge: A Reproducible Workflow for Constructing an Evidence-Grounded Synthetic Temporal QA Benchmark from Longitudinal Clinical Notes. *Computers* **2026**, *15*, 751. https://www.mdpi.com/2073-431x/15/11/751

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.