Separating Clinical Reasoning from Termination Authority in Agentic LLMs

Dr. Lina Kankevičienė, Julija Daranda

KAUNO KOLEGIJA
Higher Education Institution

CORE RESULT **Prompted stopping violated the coverage rule in up to 48% of degraded cases. Hard gating enforced zero CVT by construction, with measurable escalation and accuracy costs.** Specification compliance ≠ clinical safety

1 Why this matters

Clinical dialogue has two distinct tasks:

decide what to ask and decide when enough evidence has been collected to stop.

Research question: should the model's own sufficiency judgment control termination, or should stopping be authorized by a verifiable state rule?

2 What we compared

Same clinical task; different evidence-state signal and stopping authority.

PROMPTED STOPPING

Safety agent (full checklist) → LLM estimate (missing items) → Coordinator (ASK / ASSESS) → Terminate? (model judgment)

DETERMINISTIC AUTHORIZATION

Case state (observed facts) → Coverage check (programmatic) → Hard gate (all items covered?) → ASSESS (authorized)

ESTIMATION ≠ AUTHORIZATION

3 Study at a glance

300 acute-care cases (synthetic)	**6** complaint domains (acute care)	**4** information levels (100 · 75 · 50 · 25%)
6,000 Qwen evaluations (main evaluation matrix)	**qwen3.5:9b** primary model (local Ollama)	

PRIMARY ENDPOINT

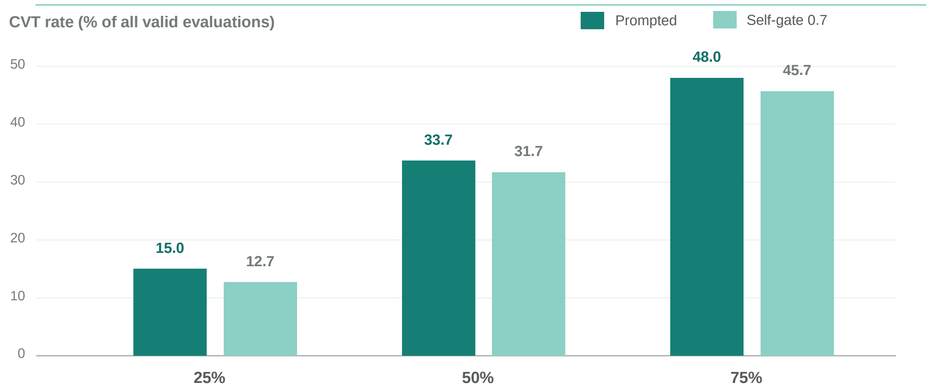
CVT = concluding while ≥1 required evidence item remains unaddressed.

Denominator: all valid evaluations within each method × information-level condition, including escalations.

Urgency accuracy: agreement with rule-generated labels (1–3; 3 = highest urgency); escalation was scored as 3.

4 Prompted stopping violated coverage

CVT increased from 15.0% to 48.0% across degraded-information conditions.



Final sufficiency score vs complete critical-evidence coverage

AUROC: 0.82 (25%) → 0.70 (50%) → 0.62 (75%)

Reading the chart: because escalations remain in the CVT denominator, a policy that escalates more often has fewer opportunities to terminate with missing evidence.

5 Self-gating exposed a sharp trade-off

Post hoc terminal filter only: low-score ASSESS → ESCALATE; no additional question.

0.70 threshold — CVT 12.7–45.7%	**0.95** threshold — CVT 1.7–3.0%	**6.3–20.7%** conclusion rate — at threshold 0.95

Lower CVT was achieved mainly by refusing more terminal ASSESS decisions – not by continuing information acquisition.

6 Deterministic gate: compliance guaranteed, utility measured

0 CVTs — 1,200 corrected-rerun evaluations	**72.7–98.3%** concluded — degraded information	**27.3%** escalated — at 25% availability

A terminal-transition bypass in the original implementation was corrected before the full rerun.

Zero CVT is a property of the enforced rule—not evidence that the model learned safer clinical reasoning.

7 Operational consequences

QWEN

Urgency accuracy remained above the single-pass baseline at ≤50% information availability.

GEMMA

−10 percentage points

urgency accuracy vs prompted CLARITY at 25% availability

paired 95% CI: −19 to −1 points

Comparators are reported exactly as analyzed: Qwen vs single-pass; Gemma vs prompted CLARITY.

8 What the comparison actually tests

- The Safety agent saw the full checklist; the prompted coordinator received an LLM-estimated missing-item list.
- Deterministic authorization used programmatically verified coverage.
- The contrast therefore changes both evidence-state estimation and termination authority.
- The threshold sweep is a terminal filter, not a live policy that can resume questioning.

9 Limits to interpretation

- Critical-evidence checklists and urgency labels were author-designed and not independently clinically validated.
- Escalation was scored as urgency 3, which favors conservative handling of urgent cases.
- External text transfer was limited when required schema fields were absent.

TAKE-HOME MESSAGE

Separate sufficiency judgment from termination authority.

A verifiable invariant can govern when an agent may conclude; validating that invariant remains a clinical requirement.

DAMSS 2026 · DATA ANALYSIS METHODS FOR SOFTWARE SYSTEMS · 26–28 NOVEMBER 2026 · DRUSKININKAI