Literature-Grounded Language Models for Auditing Biomedical Ontologies

John Bang
Handshake AI and UC Berkeley
John_bang@berkeley.edu

Abstract

Biomedical ontologies encode clinical distinctions that downstream systems inherit, but auditing them is difficult: many apparent defects are legitimate modeling choices, while many real defects require both graph-level and domain-level evidence to recognize. We study whether language models can act as triage systems for biomedical ontology audit when their prompts expose progressively richer context: concept names alone, release-specific ontology neighborhoods, and ontology neighborhoods grounded in recent clinical literature. Our setting is deliberately narrower than automated ontology repair. The model's task is to propose candidate defects; validity, actionability, and novelty must be established by independent review. We build an evaluation framework around SNOMED CT clinical branches and UMLS source-preservation checks, with frozen request packets, repeated model runs, exact prompt and response logging, failure accounting, blinded proposal export, and reviewer-gated analysis. A retrospective source comparison across twelve clinical branches identifies 21 concepts whose measured SNOMED CT US records changed between March and September 2026, including 15 with changed transitive ancestry, providing naturally occurring release-change material for audit evaluation. In sleep-apnea development experiments, open-ended model audits surfaced candidate terminology issues, including a baclofen mechanism discrepancy later supported by external sleep-medicine expert correspondence. In September 2026, correspondence from Prof. Winfried Randerath's office stated that "Baclofen is indeed a GABA-B agonist, not GABA-A" and that the examined SNOMED CT classification is likely a terminology error [3]. These findings motivate, but do not replace, blinded adjudication. By separating candidate generation from validation, and by treating ontology source records, clinical literature, expert feedback, and curator disposition as distinct evidence layers, the framework tests a practical question for high-stakes biomedical AI: can language models improve the prioritization of terminology defects without collapsing plausible explanation into proof?

1 Introduction

Biomedical ontologies and clinical terminologies are infrastructure. Systems such as SNOMED CT and UMLS support cohort retrieval, coding, phenotype definitions, decision-support rules, knowledge graphs, and increasingly the context supplied to language-model systems [6, 8]. Because these artifacts are manually curated, historically layered, and distributed through release cycles, they can contain missing relationships, outdated synonyms, underspecified mechanisms, or ambiguous modeling choices. Yet an apparent error is often not an error. It may be a deliberate editorial decision, an inherited modeling convention, a post-coordination issue, or a mismatch between a researcher's intended clinical use case and the terminology's scope.

This paper asks whether language models can help audit such systems without being mistaken for authorities over them. The distinction matters. A model can produce persuasive text about a candidate defect, but persuasion is not validation. A literature-supported statement can identify tension with a terminology record, but it does not determine the correct authoring pattern. Two models agreeing with each other may indicate robustness, or only shared pretraining and prompt biases. In clinical terminology, the endpoint is not fluent explanation; it is an actionable, release-specific, independently reviewed editorial claim.

We therefore frame language models as triage systems for ontology audit. Given a focal concept and controlled context, a model proposes possible terminology defects. Those proposals are then deduplicated,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

blinded, and adjudicated by qualified human reviewers using the same canonical evidence packet. The scientific question is not whether a model can write a plausible curator request, but whether additional context changes the rate at which valid and actionable proposals are surfaced per scheduled audit request.

Our motivating domain is sleep-apnea terminology in SNOMED CT and its source representation in UMLS. Development work in this domain produced candidate issues such as the placement of "complex sleep apnea" and a baclofen receptor mechanism discrepancy. The latter illustrates the evaluation problem sharply: baclofen is widely described as a GABA-B agonist in sleep-medicine and pharmacology sources [5, 1, 2], while the examined terminology context represented it with a GABA-A receptor agonist disposition. A sleep-medicine expert group later confirmed by correspondence that the GABA-A classification is likely a terminology error. Specifically, correspondence from Prof. Winfried Randerath's office stated that baclofen is a GABA-B agonist rather than GABA-A, and described the SNOMED CT classification as most likely a terminology error [3]. This is valuable evidence, but it is still not the same as formal terminology-authority acceptance.

The contributions of this draft are fourfold:

1. We specify a release-specific evaluation framework for language-model ontology audit with three context conditions, repeated runs, frozen request hashes, failure accounting, blinded review export, and reviewer-gated analysis.

2. We reconstruct prior development experiments rather than retrofitting them into a single ablation, exposing output failures, self-review limitations, ceiling effects, and missing independent labels.

3. We construct a twelve-branch SNOMED CT sampling frame and a March–September 2026 source-change inventory over 318 concepts, finding 21 concepts with measured record changes and 15 with changed transitive ancestry.

4. We give an evidence-ladder formulation for candidate ontology defects: model proposal, source-record verification, literature support, expert feedback, curator submission, and curator disposition.

The current evidence supports feasibility and motivates a benchmark. It does not yet establish that literature grounding improves precision, clinical reasoning, or patient outcomes. We make that limitation explicit because high-stakes biomedical AI needs evaluation designs that preserve the difference between candidate generation and validation.

2 Problem Setting

Let $G_r = (V_r, E_r)$ denote a release-specific terminology graph for release r, where nodes are active concepts and edges include hierarchical and selected attribute relationships. Each audit request is centered on a focal concept $u \in V_r$ within a clinical branch. A model receives one of three context packets:

A: name-only context. The focal identifier, fully specified name, and synonyms.

B: ontology context. A plus release-specific branch records, inferred relationships, ancestor context, attributes, language acceptability, and stated limitations.

C: ontology plus literature context. B plus a frozen literature packet selected for the branch.

The model emits up to three proposed defects. A proposal is a structured claim about a release-specific terminology record: for example, an incorrect attribute value, missing parent, synonym-scope problem, or candidate concept gap. The model is not allowed to modify the ontology. Its output is an object for review.

We distinguish three quantities that are often blurred in informal audits. Proposal rate is the number of delivered proposals per scheduled request. Precision is the fraction of delivered proposals independently labeled valid. Actionable yield is the number of proposals labeled both valid and directly actionable per scheduled request. Fewer proposals are not automatically fewer false positives; they can also be missed opportunities.

Table 1: Current public sampling frame and planned request scale. Counts are release- and pipeline-specific, not population estimates.

Quantity	Value
Selected clinical branches	12
Active concepts in sampling frame	318
Root-stage scheduled requests	216
Full concept-level scheduled requests	5,724

For branch b, model m, condition a, and repetition r, let P_{bmar} be the number of schema-valid proposals and let J_{bmar} be the number independently labeled valid and directly actionable. With N_{bma} scheduled requests, the primary paired effect for model m is

$$\Delta_{C-B,m} = \frac{1}{B} \sum_{b} \left(\frac{\sum_r J_{bmCr}}{N_{bmC}} - \frac{\sum_r J_{bmBr}}{N_{bmB}} \right). \tag{1}$$

Failed requests remain in the scheduled-request denominator for yield, while their unobserved scientific output is treated as unknown rather than as a true negative. Precision is reported as V/P among delivered proposals, with unresolved-label bounds.

3 Data Sources and Processing

3.1 Terminology records

The target terminology is SNOMED CT [6]. The work uses release-specific RF2 Snapshot records. Hierarchy edges are oriented child-to-parent. For a concept u, $A(u)$ denotes strict ancestor closure; a missing direct edge (u, v) is not a missing subsumption if $v \in A(u)$. The importer retains active concepts, active inferred relationships, selected stated OWL records, US English language membership, and selected MRCM files. It does not claim complete description-logic classification or complete historical-association modeling.

UMLS is treated as an integration and provenance system rather than as an independent ontology against which SNOMED CT can be naively judged [8]. For source-preservation checks, MRCONSO atom identifiers connect UMLS source atoms to SNOMED CT identifiers, and MRREL is restricted to the same SNOMED CT source and unsuppressed records. Relationship direction follows the RRF convention: for isa, the second atom bears the recorded relationship to the first; inverse_isa is normalized accordingly. This verifies extraction and source preservation for selected records; it is not a semantic completeness test over all UMLS sources.

3.2 Branch sampling frame

The planned multi-branch benchmark uses twelve disjoint clinical branches from the September 2026 US release. Candidate roots were selected from ordered clinical topics, constrained to branches with 5–80 concepts, and overlap was rejected. Sleep apnea and drug-mechanism leads were retained as development material and excluded from the prospective sampling frame. The selected frame contains 318 active concepts. With two models, three context conditions, and three repetitions per concept, the complete concept-level design schedules 5,724 requests. A smaller root-stage calibration schedules 216 requests.

3.3 Literature packets

A development sleep-apnea corpus contained 19 substantive publications and three commentary or correction records. For the multi-branch design, a later exploratory retrieval pass selected 24 full-text sources, two per branch, from 3,000 focused Europe PMC branch-paper matches. These selected sources were assistant-screened and hashed, but they are not a systematic review and not an independent human reference standard. The protocol requires two independent screeners before the literature condition can be treated as confirmatory.

4 Evaluation Design

4.1 Generation

The planned generator uses two frontier language-model APIs with frozen request packets. Each request stores the requested model identifier, returned model-version field when available, UTC timestamp, full visible prompt, generation configuration, response text, finish reason, usage metadata, elapsed time, and request hash. Existing attempt markers prevent silent retries. A version change or rescue run is a separately labeled experiment.

Outputs must follow a constrained schema: target concept, proposed defect type, proposed action, evidence anchors, uncertainty, and rationale. Maximum proposal count is capped to discourage long speculative lists. Malformed outputs are retained as failures.

4.2 Blinded adjudication

Two qualified human reviewers independently label each deduplicated proposal. One reviewer must have terminology or SNOMED CT expertise, and one must have relevant clinical-domain expertise for the branch. Reviewers see the same canonical source and evidence packet regardless of the originating model condition. They do not see model identity, condition, repetition, self-review confidence, or allocation metadata.

Each proposal receives labels for validity, actionability, novelty, and error type. Validity distinguishes supported defects from already represented, unsupported, out-of-scope, or ambiguous claims. Actionability distinguishes no supported action, editorial investigation, and specific justified correction. Agreement is computed on initial labels; adjudicated labels determine final estimates.

4.3 Analysis

The primary estimand is branch-mean actionable yield difference between C and B. Secondary measures include proposal rate, precision, unresolved-label bounds, per-model effects, run-to-run variability, failure rate, latency, token usage, reviewer time, and error-type distribution. We plan paired branch bootstrap intervals and branch sign-flip tests for the primary comparison, with multiplicity adjustment across model-specific tests. With twelve purposive branches, intervals describe uncertainty over selected clusters rather than population-wide generalization.

5 Development Results

5.1 Release-change inventory

We compared selected measured fields for the 318-concept sampling frame between March and September 2026 SNOMED CT US Snapshot records. The comparison included active concept records, active US English synonyms, inferred relationships, and strict ancestor closures. Nonzero role groups were normalized by member sets to avoid counting simple group renumbering as a structural change.

Twenty-one concepts had at least one observed change. Of these, 15 had changed transitive ancestry, 12 had changed inferred-relationship structures, 10 had changed definition status, and 3 had changed active US synonym inventories. Three concepts changed only in transitive ancestry among measured fields, showing why direct-edge comparisons alone are insufficient.

This inventory is an external editorial-source signal, not a gold standard. Some changes may be editorial reorganizations rather than corrections; unchanged concepts are not validated negatives. Nevertheless, it supplies naturally occurring release-change material that can be used to test whether audit systems recover consequential graph changes.

5.2 Historical model-call reconstruction

We reconstructed 31 main experiment calls from archived project outputs: six branch-audit feasibility requests, four frontier discovery/self-review requests, three drug discovery/challenge requests, twelve reading-

Table 2: Observed March–September 2026 source changes in the 318-concept selected frame. Categories overlap.

Measured change category	Concepts
Any observed measured change	21
Transitive ancestry changed	15
Inferred relationship structure changed	12
Definition status changed	10
Active US synonym inventory changed	3
Changed active OWL expression text among changed concepts	12

Table 3: Open-ended audit outputs reconstructed from archived experiments. Proposal counts are not precision.

Experiment	Condition/model	Calls	Usable	Proposals
Primary feasibility	evidence / 2.5 Flash	2	1	8
Compact follow-up	ontology / 2.5 Flash	1	1	8
Compact follow-up	evidence / 2.5 Flash	1	1	11
Frontier discovery	3.1 Pro Preview	1	1	4
Frontier discovery	3.8 Flash	1	1	4
Drug discovery	ontology / 3.8 Flash	1	1	6
Drug discovery	literature / 3.8 Flash	1	1	4

ablation requests, and six mechanism-answering requests. Early engineering attempts and selected-dossier calls were excluded from this inventory.

Table 3 reports open-ended proposal-producing experiments that returned usable structured candidates. The primary feasibility run also revealed a practical engineering limitation: several scheduled requests ended at the output limit or failed validation. Later compact and frontier runs produced candidate proposals, but these outputs remain development observations until independently adjudicated. Same-model review retained some model-proposed findings, but self-review is not independent validation.

Repeated reading and mechanism-answering experiments showed ceiling behavior rather than a context benefit. In the reading ablation, both models matched the investigator key on all clinical and graph questions in every run under both baseline and reading conditions; the only disagreement concerned an uncertainty label with multiple defensible interpretations. In the mechanism-answering experiment, all six requests matched all 24 investigator-keyed questions across three context variants. These results show that the development questions were too easy or too keyed to existing model knowledge to establish that context improves audit performance.

5.3 Case study: baclofen mechanism discrepancy

The strongest development case is baclofen. In the examined SNOMED CT records, baclofen substance concept 387342009 had a stated disposition to GABA-A receptor agonist and was inferred under an A-agonist substance class. UMLS preserved that SNOMED CT source representation, while other source definitions for the same baclofen CUI described GABA-B activity [7, 4]. Literature sources in sleep medicine and pharmacology also described baclofen as a GABA-B agonist [5, 1, 2]. After outreach, correspondence from Carla Miltz on behalf of Prof. Winfried Randerath responded that “Baclofen is indeed a GABA-B agonist, not GABA-A” and that the SNOMED CT classification is most likely a terminology error [3].

This case demonstrates the proposed evidence ladder. The model-assisted audit surfaced and documented a candidate. Source-record checks localized the issue to a stated and inferred mechanism, not a synonym display artifact. Literature and cross-source definitions supported the pharmacologic conflict. External expert feedback strengthened the case. The remaining steps are terminology-authority submission, formal disposition, and regression checks over affected substances and products. Until those steps occur, the correct

label is “expert-supported candidate defect,” not accepted ontology correction.

6 Discussion

The development work suggests that language models are useful for producing inspectable audit candidates, but it also shows why ordinary LLM evaluation habits are unsafe for biomedical ontology QA. A model can generate eight plausible proposals while three quarters of sibling requests fail schema validation. A second model can agree with a first model without supplying an independent reference standard. A reading intervention can appear scientifically motivated while the actual answer task is already at ceiling. A literature-backed claim can expose a terminology tension while leaving the correct editorial action unresolved.

The main design lesson is that context should be evaluated through adjudicated yield and precision, not through plausibility or proposal volume. Literature may reduce false positives by grounding the model in current clinical usage; it may also increase false positives by encouraging the model to treat every clinical phenotype as a missing precoordinated concept. Graph context may prevent proposals for already represented relationships; it may also anchor the model to source errors. Only blinded, branch-paired review can distinguish these effects.

For machine learning, biomedical ontology audit is an attractive evaluation setting because it combines retrieval, structured reasoning, uncertainty, and high-stakes domain grounding. Unlike many question-answering benchmarks, a correct answer may require saying that the evidence is insufficient, that a candidate is already represented transitively, or that the issue should be submitted for editorial review rather than patched locally. This makes the task a useful testbed for model calibration and tool-mediated reasoning.

7 Limitations

The most important limitation is that the expanded three-condition experiment has not yet produced model outputs or independent human labels. Therefore this draft cannot estimate human-validated precision, actionable-yield differences, or context effects. The release-change inventory is retrospective and based on a September-defined sampling frame, so it is not a population-wide change-rate estimate. The sleep-apnea and drug cases were development material and are subject to adaptive selection. Expert correspondence for baclofen is external feedback, not formal curator acceptance; raw correspondence should not be redistributed without permission from the sender. The terminology importer does not perform full description-logic classification or exhaustive historical modeling. Finally, model API identifiers may point to mutable systems, so exact replication requires archived prompts, outputs, model-version metadata, and dates.

8 Ethics Statement

This work concerns terminology and ontology records, not patient-level data. No clinical recommendation or diagnosis is made. The main ethical risks are overclaiming model authority, releasing restricted terminology content, and implying institutional or expert endorsement where none exists. The proposed release plan therefore separates public code, prompts, hashes, bibliographic metadata, derived identifiers, and aggregate statistics from licensed source records, raw requests containing licensed content, credentials, reviewer identities, and private expert correspondence. Short excerpts from correspondence may support a manuscript only with appropriate permission and context. Candidate defects should be submitted through appropriate terminology-maintenance channels rather than applied as informal local corrections.

9 Reproducibility Statement

The repository contains the parsing code, prompt builders, historical reconstruction scripts, release-change inventory code, reviewer rubric, planned statistical analysis, and sanitized derived outputs. Exact licensed terminology packets are rebuilt locally from edition checksums and are not redistributed. Each model call is designed to record request hashes, response metadata, failures, and costs. The current draft reports

development and infrastructure results; confirmatory claims require executing the frozen request set and completing blinded human adjudication.

References

[1] Simona Franchini et al. Baclofen: from laboratory to clinical use. ACS Chemical Neuroscience, 2020. Pharmacology source supporting GABA-B activity.

[2] Cyrille Goudet et al. Baclofen and analogues as gaba-b receptor ligands. Journal of Medicinal Chemistry, 2013. Primary pharmacology source supporting GABA-B receptor activity.

[3] Carla Miltz and Winfried Randerath. Email correspondence concerning baclofen receptor mechanism and snomed ct classification, 2026. Personal communication from Carla Miltz on behalf of Prof. Winfried Randerath, September 2026; cited as expert correspondence, not terminology-authority disposition.

[4] National Cancer Institute. NCI Dictionary of Cancer Terms: Baclofen, 2026.

[5] Winfried Randerath et al. Medication-induced central sleep apnea: beyond the unification of concepts, seizing the opportunity to precision medicine. Sleep, 2024. Used as sleep-medicine motivation for medication-associated central sleep apnea and baclofen mechanism review.

[6] SNOMED International. SNOMED CT Editorial Guide and Release File Specification, 2026.

[7] U.S. National Library of Medicine. MeSH entry for Baclofen, 2026.

[8] U.S. National Library of Medicine. UMLS Reference Manual: Rich Release Format, 2026.