

Ontology-Governed LLM Search for Safe Drug Concept Mapping in Air-Gapped Healthcare Environments

Seung Ju Yun, Eun Kyung Ko, Myung Ji Yoo
Chonnam National University Hospital

Background

Large language models (LLMs) are increasingly used to standardize clinical vocabularies, including mapping local drug codes to OMOP concepts [1,2]. Three problems recur: they hallucinate [3], their output can vary between runs even under fixed settings [4], and they are black boxes, so when a mapping is wrong one cannot see where it went wrong [5]. Air-gapped hospitals, limited to small local models, feel this most.

In our design, a rule-based program built on the ontology (vocabulary relationships) controls the search and decides acceptance; the LLM runs inside it as a component, called for steps rules cannot do: reading the raw string, proposing candidate names, and measuring closeness in meaning. It never chooses the mapping. Because acceptance passes explicit rules, every outcome can be traced to the rule that allowed or stopped it.

The principle is the one we submitted; the implementation, including the feedback behind the submitted examples, is being rebuilt. This report asks how far its first pass gets, where control holds and where the design loses ground.

Methods

Setting. Each system received only the drug string as stored in the hospital drug master, with no cleaning, translation or prescription context, plus the OMOP vocabularies (RxNorm, RxNorm Extension) [6], a SapBERT encoder [7] and one quantized local LLM (Qwen3.8-27B or EXAONE-32B). The setting was kept minimal by design: no fine-tuning, no knowledge beyond the vocabulary, no human in the loop.

Pipeline (Figure 1, Algorithm 1). One pass runs three steps: the LLM forms hypotheses, the vocabulary bounds the search, and explicit rules decide.

Hypotheses. The string is read as six components (ingredient, brand, dose form, route, strength, supplier). For each, the LLM copies the fragment that states it and proposes up to four standard names. A name is kept only if it exists as a vocabulary concept, weighted by how directly the string supports it, and competes with an explicit UNKNOWN, so a component the string does not settle stays open. Prompts were brief and example-free.

Search. Vocabulary relationships expand these concepts into real drug concepts. Multiplying the component beliefs scores every combination, but only combinations a real concept carries survive; the belief left on the rest, m_{struct} , shows how well the LLM's reading fits what exists (near 0: real products; near 1: products that do not exist). The search then descends concept classes from ingredient to marketed product, measuring a few candidates per class by trajectory distance (TD; called semantic distance in [8]). The LLM writes continuations of the raw string and of the candidate name and scores each under both texts; texts that make the same continuations equally likely mean the same to the model, including what the string leaves unsaid.

Acceptance. A rule accepts the most specific single-ingredient candidate whose stated components match the hypotheses and whose added components the vocabulary uniquely entails, without contradiction; otherwise the program abstains and records which rule failed. We ran this single pass; returning the record to the LLM in further passes is future work.

Comparator. The same SapBERT retrieved the 20 nearest of 1,527,164 standard drug concepts, and the same LLM selected one or none.

Evaluation. We drew 200 codes at random from single-ingredient local codes whose Korean insurance (EDI) code maps to one standard drug concept ($N = 4,747$). The first author labeled each acceptance allowable or not under study-specific rules on ingredient, strength, brand, release and dose form. Because EDI-linked concepts do not always separate a salt from its base, strength followed the local product's basis of strength (BoSS; enalapril maleate 10 mg $\neq$ enalapril 10 mg). Proportions carry Wilson 95% intervals [9]; abstention analyses are post hoc lower bounds.

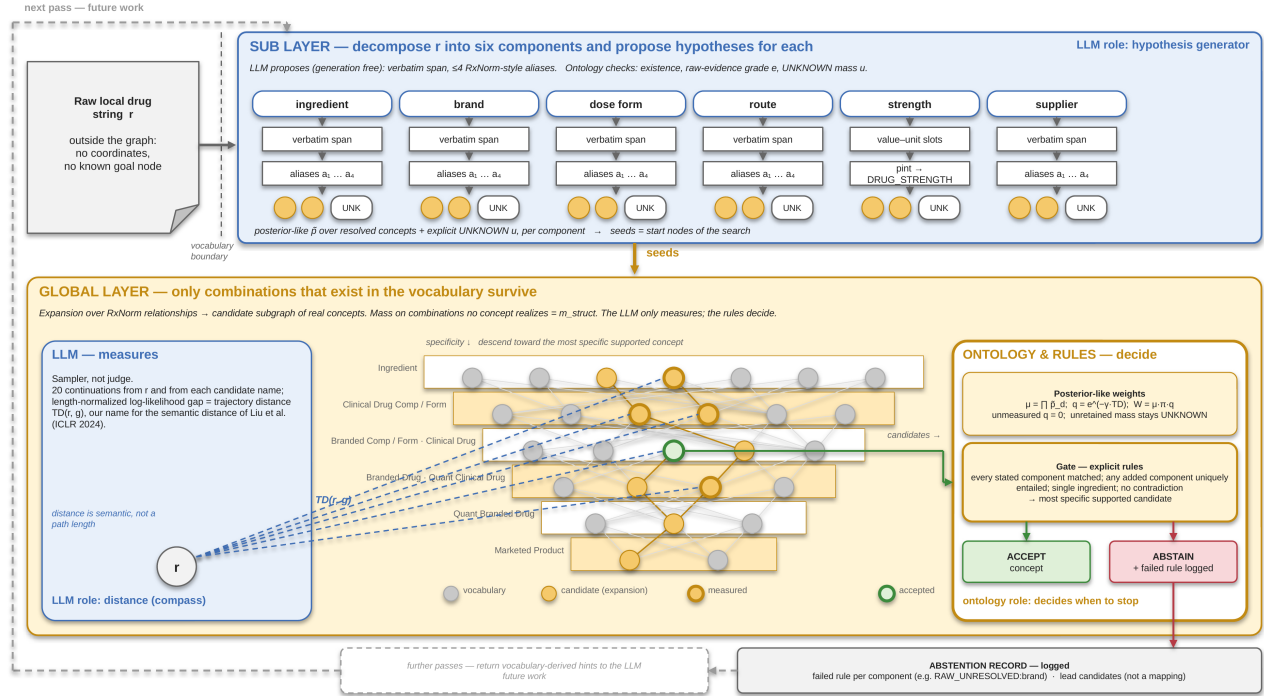


Figure 1. A rule-based program with the LLM as a component. Upper panel: the LLM proposes, for each of six components, the fragment it read and up to four standard names; the vocabulary keeps only existing concepts and sets an explicit UNKNOWN beside them. Lower panel: only combinations carried by real concepts survive (unrealized belief = m_{struct}); trajectory distance (TD; called semantic distance in Liu et al. [8]) guides the descent through concept classes; explicit rules accept or abstain and log the failed rule. Grey dashed: returning vocabulary-derived hints to the LLM in further passes, future work.

Algorithm 1 Ontology-governed drug concept mapping (one pass)

Input: raw string r ; vocabulary V ; local LLM M

Output: concept c^* , or ABSTAIN with the failed rule

Hypotheses: the LLM proposes, the vocabulary verifies

- 1: **for** each component $d \in \{\text{ingredient, brand, form, route, strength, supplier}\}$ **do**
- 2: $s_d \leftarrow M.\text{copy}(r, d)$ $\triangleright$ fragment of r stating d , or $\emptyset$
- 3: $A_d \leftarrow M.\text{propose}(s_d, k=4) \cap V$ $\triangleright$ names that exist as concepts
- 4: $p_d \leftarrow \text{belief over } A_d \cup \{\text{UNK}\}$, weighted by support in r

Search: the vocabulary bounds, the LLM measures

- 5: $C \leftarrow \text{concepts reached from } \bigcup_d A_d \text{ via relations of } V$ $\triangleright |C| \leq 1000$
- 6: $\mu(a) \leftarrow \prod_d p_d(a_d)$; $m_{\text{struct}} \leftarrow \sum_{a \notin C} \mu(a)$ $\triangleright$ belief on non-existent products
- 7: **for** class $\ell = \text{Ingredient} \rightarrow \dots \rightarrow \text{Marketed Product}$ **do**
- 8: $w(c) \leftarrow \mu(c) e^{-\gamma \text{TD}(r, c)}$ **for** ≤ 3 candidates $c \in C_\ell$ $\triangleright \leq 24$ per string

Acceptance: explicit rules decide

- 9: $E \leftarrow \{c : \text{one ingredient, stated } d \text{ match } p_d, \text{ added } d \text{ uniquely entailed, no contradiction, } w(c) > 0\}$
- 10: **if** $E \neq \emptyset$ **then return** most specific $c^* \in E$
- 11: **else return** ABSTAIN, failed rule per component, lead candidates

$\text{TD}(r, c) = \frac{1}{2n} \sum_{t \in T_r \cup T_c} |\log M_r(t) - \log M_c(t)|$, a sampled estimate of the semantic distance of Liu et al. [8], with $n = 20$ continuations sampled from each text and $\log M_s(t)$ the mean per-token log-probability of t after s ; $\gamma = 2$.

Results

Across 200 items (Table 1), the pipeline accepted 44 with Qwen and 11 with EXAONE, of which 16 (36.4%; 95% CI 23.8–51.1) and 3 (27.3%; 9.7–56.6) were non-allowable. The same LLMs selecting from retrieved candidates accepted 134 and 189, with 37.3% (29.6–45.7) and 51.1% (44.0–58.1) non-allowable; EXAONE as selector named another ingredient 28 times, the pipeline never.

Table 1. The same LLM selecting from retrieved candidates (baseline) and working as a component of the pipeline, on 200 items.

	Qwen		EXAONE	
	selects	pipeline	selects	pipeline
All 200 items				
Allowable acceptance	84	28	92	8
Non-allowable acceptance	50	16	96	3
Acceptance not assessable ^a	0	0	1	0
Abstained	—	156	—	189
No acceptance on record ^b	66	—	11	—
Non-allowable, % of assessed acceptances	37.3	36.4	51.1 ^a	27.3
95% CI	29.6–45.7	23.8–51.1	44.0–58.1	9.7–56.6
... naming another ingredient ^c	1	0	28	0
Items the selector accepted wrongly				
Pipeline: allowable / non-allowable / abstained	50	0/10/40	96	2/1/93
Abstentions: nearest candidate's ingredient matched ^d		37/39		75/86
... its dose form: matched / omitted / contradicted / none / added ^e		10/16/2/7/2		18/35/5/15/2

Notes. Labels: the first author judged each acceptance under this study's allowability criteria. Selects = SapBERT top 20 + the same LLM picks one or none. a. One EXAONE selector acceptance could not be assessed (strength basis unverified) and is excluded from the denominator (188). b. Selector records list acceptances only, so abstention and no answer cannot be told apart. c. A different active ingredient, not a salt or derivative of the right one; includes one item where every selector chose a bovine instead of a porcine brain preparation. d. Nearest = the measured candidate with the lowest trajectory distance. Ingredient match follows the existing component scoring, among abstentions with a scored candidate; it does not mean the whole candidate was allowable (confirmed in 2 and 5 cases; most not reviewed). A candidate whose biological source is unverified is not counted as a match. e. Among nearest candidates with a matching ingredient; omitted = stated in the string but not carried, none = neither stated nor carried, added = carried but not stated. Qwen with thinking selected 146, 56 non-allowable (38.4%; 30.9–46.4), 9 naming another ingredient; on those 56 items the Qwen pipeline accepted 1 allowable and 10 non-allowable and abstained on 45.

All 19 pipeline errors concerned product-level detail: salt, derivative or strength basis (6; the pipeline resolves standard ingredients only), brand or product variant (5), an omitted release characteristic (4), dose form or device (3), and one overly general concept. The same model's selector also erred on all six salt-type items but was allowable on seven others, including Invega ER: Qwen omitted the release detail and EXAONE accepted a capsule.

Abstentions were not empty: an allowable candidate had been measured in 44 of 156 (Qwen) and 63 of 189 (EXAONE) abstentions and was the nearest in 20 and 24 (lower bounds). Where the selector had erred, the nearest candidate usually matched in ingredient (37/39, 75/86) but left the stated dose form out more often than it matched it (16 vs 10; 35 vs 18). Within one pass, search was bounded: the pool was capped in 57 and 59 abstentions, and class layers were not fully sampled in 132 and 162. In two exploratory items, adding 20 package-description chunks lowered TD for every candidate but moved the ingredient-level concept to the top.

Conclusions

In this minimal setting, explicit rules kept the LLM's errors traceable but did not by themselves make acceptances more correct: with Qwen, the error rate matched that of the same LLM selecting alone, at a third of the coverage. Still, no pipeline acceptance crossed the coarse boundary between ingredients. We assumed real products sit far apart, so brief hypotheses would suffice and TD would recover unstated detail. The results show otherwise: close siblings differing only in salt basis, release, device or brand all exist, which an existence check cannot separate; brief hypotheses often dropped what the string stated; and TD tended toward the ingredient level. The distinguishing detail must therefore enter as a hypothesis and be checked by its component's rule. Next, the LLM should build a richer hypothesis space from fuller prompts, external knowledge and vocabulary hints returned across passes, narrowing the search. Limits: 200 items, one adjudicator, no clinician reference standard, bounded search, no combination products.

References

1. Adams MC, Perkins ML, Hudson C, Madhira V, Akbilgic O, Ma D, et al. Breaking digital health barriers through a large language model-based tool for automated Observational Medical Outcomes Partnership mapping: development and validation study. *J Med Internet Res*. 2025;27:e69004. doi:10.2196/69004
2. Zhang G, Fang Y, Chen F, Ta C, Hripcsak G, Ryan P, et al. A neural-symbolic AI agent system for biomedical concept mapping. *npj Digit Med*. 2026;9:425. doi:10.1038/s41746-026-02594-6
3. Ji Z, Lee N, Frieske R, Yu T, Su D, Xu Y, et al. Survey of hallucination in natural language generation. *ACM Comput Surv*. 2023;55(12):248. doi:10.1145/3571730
4. Atıl B, Aykent S, Chittams A, Fu L, Passonneau RJ, Radcliffe E, et al. Non-determinism of “deterministic” LLM system settings in hosted environments. In: *Proceedings of the 5th Workshop on Evaluation and Comparison of NLP Systems (Eval4NLP)*. Mumbai: ACL; 2025. p. 135-48. doi:10.18653/v1/2025.eval4nlp-1.12
5. Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell*. 2019;1(5):206-15. doi:10.1038/s42256-019-0048-x
6. Reich C, Ostropolets A, Ryan P, Rijnbeek P, Schuemie M, Davydov A, et al. OHDSI Standardized Vocabularies—a large-scale centralized reference ontology for international data harmonization. *J Am Med Inform Assoc*. 2024;31(3):583-90. doi:10.1093/jamia/ocad247
7. Liu F, Shareghi E, Meng Z, Basaldella M, Collier N. Self-alignment pretraining for biomedical entity representations. In: *Proceedings of NAACL-HLT 2021*. ACL; 2021. p. 4228-38. doi:10.18653/v1/2021.naacl-main.334
8. Liu TY, Trager M, Achille A, Perera P, Zancato L, Soatto S. Meaning representations from trajectories in autoregressive models. In: *International Conference on Learning Representations (ICLR)*; 2024. arXiv:2310.18348
9. Wilson EB. Probable inference, the law of succession, and statistical inference. *J Am Stat Assoc*. 1927;22(158):209-12. doi:10.1080/01621459.1927.10502953

Acknowledgments

This research was supported by a grant of the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number : RS-2025-02373039).

Use of generative AI

ChatGPT Astral (OpenAI) pre-annotated, in October 2026, which components each drug string states, to help build the reference labels; it saw drug strings only, no patient data, and the first author adjudicated every label. Claude (Anthropic) assisted with drafting text, analysis scripts and figure layout. The authors reviewed and take responsibility for all content. Neither evaluated system uses a remote model.