Semantic fidelity limits of language-model ontology compilation

Jimin Kang
AiNOS Research

August 2026

Abstract

Large language models are increasingly used to transform unstructured text into typed knowledge representations. We tested the stronger claim that a sufficiently orchestrated language-model workflow can produce a stable, lossless, canonical ontology that replaces the source as an authoritative reasoning substrate. The research programme combined a reported repetition study of 5,000 compilation outputs with a repository-verified legal-domain compiler comprising nine source identities, 1,364 immutable source units, 61 semantic windows, 3,090 objects, 5,803 assertions and 7,567 relations. We then performed a targeted source-to-ontology audit on five failure-prone windows. The resulting 593 relations all carried explicit evidentiary assertions, yet source comparison still identified material losses in actors, conditions, list membership and amendment scope. Two high-density source units illustrate the mechanism: a 13,455-character amendment unit was reduced to 33 broad claims, and a 4,219-character product-classification provision was reduced to ten claims that did not preserve its membership matrix. These observations separate graph integrity from semantic recoverability. They do not establish a universal impossibility theorem, but they reject open-ended probabilistic compilation as a sufficient basis for guaranteed canonical truth. We propose a source-first architecture in which immutable source units retain authority and generated ontologies operate as revisable semantic control planes for retrieval, navigation and candidate generation.

Keywords: large language models; ontology learning; semantic fidelity; knowledge graphs; source grounding; legal information systems

1 Introduction

Ontologies are designed to make a domain's conceptual commitments explicit and computationally usable[1]. Formal ontology languages distinguish entities, expressions and axioms under defined semantics[2], while knowledge graphs provide practical structures for integrating and querying heterogeneous information[3]. These properties make ontology construction attractive in regulated domains, where retrieval, reasoning and auditability must operate over large bodies of unstructured text.

Language models appear to remove a long-standing bottleneck in ontology engineering. Instead of manually defining every entity, claim and relation, a model can read a source document, propose semantic units, map them into a schema and connect the resulting records. Multi-stage orchestration can further decompose the task into extraction, normalization, identity resolution, relation construction and evaluation. Sampling several reasoning paths can improve answer accuracy on benchmark tasks[4], and retrieval-augmented generation can ground model outputs in external documents[5]. These results motivate a plausible engineering hypothesis: sufficiently careful orchestration might convert source text into a canonical graph while preserving the source's decision-relevant meaning.

That hypothesis is stronger than ordinary extraction quality. A useful graph may omit details and still improve discovery. An authoritative graph, by contrast, must be complete enough that a downstream system can rely on it without re-reading the source. It must also remain stable under repeated compilation, preserve the roles and conditions that determine legal effects, and expose enough provenance to reconstruct its claims. Contemporary work on hallucination and semantic uncertainty shows that fluent model outputs can remain ungrounded or semantically variable[6, 7], but those findings do not directly answer whether a large, carefully engineered semantic compiler can overcome the problem at the system level.

We therefore tested the architecture rather than a single prompt. The system was repeatedly strengthened with structured outputs, frozen schemas, evidence lineage, identity merging, relation completion, semantic postflight checks and targeted repair. We then compared compiled outputs against immutable source units. The central finding is that structural integrity and semantic recoverability are distinct. A graph can be internally clean, typed and evidence-linked while still failing to preserve material distinctions that disappeared before relation construction.

Our claim is deliberately bounded. We do not show that language models cannot generate useful ontologies, nor that constrained ontology construction with external adjudication is impossible. We show that open-ended probabilistic semantic compilation, by itself, did not provide the guarantees required to replace the source as canonical truth. This distinction leads to a different architecture: the source remains authoritative, while the ontology becomes a fallible but valuable semantic control plane.

2 Research hypothesis and falsification criteria

Let S denote an immutable source corpus, M a model, P the prompt and schema contract, and H_i the execution history of run i. A semantic compiler produces

$$O_i = C(S; M, P, H_i), \tag{1}$$

where O_i is an ontology containing typed objects, assertions and relations. The authoritative-compilation hypothesis requires more than a high-quality sample. It requires all of the following conditions.

1. **Semantic completeness.** Every material source condition, actor, exception, temporal qualification and enumerated membership required for downstream decisions is recoverable from O_i.
2. **Canonical stability.** Repeated compilations of the same source are semantically equivalent, even when their surface wording differs.
3. **Relational correctness.** Predicates, directions, endpoints and qualifiers preserve the source's meaning.
4. **Authority substitution.** A downstream reasoner can rely on O_i without consulting S for material claims.

The hypothesis is falsified operationally if repeated runs select materially different semantic units, if source comparison reveals unrecoverable information after structural validation, or if a downstream repair stage cannot restore omissions without returning to the source. This is an architectural falsification criterion, not a proof that no future model can construct any valid ontology.

3 Methods

3.1 Study design

The research programme combined three complementary studies (Fig. 1). Study A examined output stability under repeated compilation. Study B examined whether a large end-to-end compiler could achieve

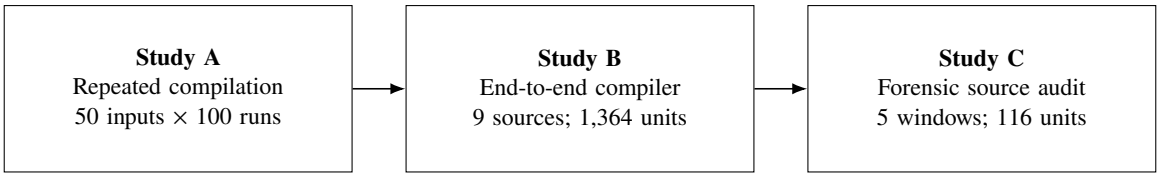

Figure 1. Three-part evaluation of authoritative semantic compilation. Study A tests repeated-output stability, Study B tests end-to-end structural compilation, and Study C compares selected source units with their compiled semantic records. The studies evaluate different necessary conditions and are not treated as interchangeable estimates of one metric.

Table 1. Scale of the repository-verified compiler and targeted audit. Counts describe stored records and source boundaries; they do not by themselves measure semantic completeness.

Artifact	Full corpus	Target audit	Interpretation
Source identities	9	5 selected windows	Frozen legal sources and audit scope
Immutable source units	1,364	116	Raw authority retained for comparison
Semantic windows	61	5	Context boundaries used by the compiler
Objects	3,090	211	Canonical entities and concepts
Assertions	5,803	428	Typed semantic claims
Relations	7,567	593	Typed connections among records

structurally coherent ontology construction at production-relevant scale. Study C held the source boundaries fixed and conducted targeted forensic comparisons between raw source units and compiled ontology records.

3.2 Study A: repeated semantic compilation

In an earlier internal experiment, 50 inputs were independently processed 100 times by an agent designed to recover source meaning and compile it into an ontology. The resulting 5,000 outputs were compared using semantic and structural vector similarity. The observed variation was not limited to paraphrase: runs differed in claim atomicity, entity boundaries, type assignment, relation selection and qualifier ownership.[E7]

The original per-run ledger and aggregate similarity distributions were not retained. We therefore classify the protocol and qualitative direction as REPORTED evidence. We do not report an unverified mean, variance, threshold, confidence interval or *P* value, and we do not use Study A to estimate an effect size.

3.3 Study B: end-to-end legal ontology compiler

The repository-verified baseline contained nine insurance-law source identities, 1,364 immutable source units and 61 semantic windows. The compiled artifact contained 3,090 canonical objects, 5,803 assertions and 7,567 relations.[E1] The pipeline included source-unit segmentation, semantic-card extraction, schema mapping, identity merging, relation completion, semantic postflight evaluation, bounded repair, replay and artifact freezing.

The experiment was intentionally adversarial to the proposed conclusion: the compiler was improved rather than held at an obviously weak baseline. Across iterations, it received structured-output contracts, evidence lineage, role-group lineage, qualifier materialization and narrower relation-patching logic. These changes reduced several known endpoint and evidence defects, allowing a later audit to distinguish upstream semantic loss from downstream relation corruption.

3.4 Study C: targeted source-to-ontology audit

Five historically failure-prone windows were recompiled against the same frozen source boundaries. Together they contained 116 source units and produced 211 objects, 428 assertions and 593 relations.[E2][E3] We materialized each selected source unit and compared its raw text with every compiled object, assertion and relation linked to that unit. The audit recorded what was preserved, compressed, reassigned or omitted, and assessed whether the difference could change retrieval or downstream decision-making.

Relation evidence was audited separately from source fidelity. Of 593 relations, 578 had exactly one supporting assertion and 15 had exactly two; none had zero or more than two. This allowed us to ask whether a structurally well-evidenced graph was also semantically recoverable.

3.5 Forensic canaries

Two dense source units were selected as canaries because they expose different compression risks. The first was a 13,455-character amendment unit from the Financial Consumer Protection Act. A mechanical scan found 23 occurrences of the Korean equivalent of "delete", 16 of "establish", 54 of "replace with", and 27 of "replace and". The v3 decomposition linked 33 claims to the unit.[E4] The second was a 4,219-character, 98-line enforcement-decree provision defining product-specific professional-consumer classes. It contained 43 numbered-item markers and 48 Korean sub-item markers; the v3 decomposition linked ten claims to the unit.[E5]

These counts are not treated as a gold ontology or as one-to-one estimates of the correct claim count. They are diagnostic indicators of source density. The substantive audit examined whether the compiled records preserved the actor-condition-effect pairings necessary to reconstruct the provision.

3.6 Evidence grading and statistical analysis

Evidence was graded as repository-verified (VERIFIED), principal-investigator reported (REPORTED), or bounded inference from those sources (INFERRED). Exact counts are reported only for locally verified artifacts. No inferential statistical test was applied because the available studies did not provide exchangeable observations under a common sampling design, and the raw Study A ledger was not retained. Accordingly, the paper reports descriptive counts and qualitative failure mechanisms rather than significance claims.

3.7 Generative-AI assistance

Language models were used in the ontology-compilation systems under study and to assist with manuscript drafting and LaTeX preparation. Quantitative claims labelled VERIFIED were checked against local artifacts. Responsibility for the study design, interpretation and final manuscript rests with the author.

4 Results

4.1 Repeated runs did not identify one canonical semantic structure

Study A reported variation in semantic decisions, not merely in wording. A proposition represented as one claim in one run could be divided across several claims in another. Entity boundaries and types changed, as did the choice of which relations to materialize and where to attach qualifiers. Vector similarity could identify broadly related outputs, but it did not establish equivalence of actor-condition-exception structures.[E7]

Table 2. Structural observations and their semantic interpretation in the five-window audit. Clean graph records are necessary for traceability but do not establish that all material source meaning survives compilation.

Observation	Structural result	Source-fidelity result
116 source units	Every unit remained in a frozen window	Inclusion did not ensure preservation of every material distinction
428 assertions	Typed records carried source references	Some actors and conditions survived only in broad prose descriptions
593 relations	578 had one evidence assertion; 15 had two	Evidence-linked edges could not restore distinctions absent from upstream claims
Endpoint and evidence repairs	Known whole-list and empty-evidence defects were removed	Residual failures localized to source-to-semantic compression

Deterministic decoding may increase string-level reproducibility within a fixed runtime, but it does not show that the repeated output is complete or uniquely determined by the source. Changes to the model, schema, prompt order or upstream extraction can expose another plausible representation. Stability of a generation procedure and semantic identifiability of its target are therefore different properties.

4.2 Graph integrity did not imply semantic recoverability

The targeted audit produced a structurally clean relation set. Every relation had one or two explicit supporting assertions, and the previously observed omnibus-evidence problem was absent. Nevertheless, source comparison found that material semantics had already been compressed before relation construction (Table 2).

This separation matters. A validator that checks schema conformance, endpoint existence and evidence cardinality can certify the graph that exists. It cannot detect a proposition that was never generated unless it independently returns to the source.

4.3 Amendment scope was compressed into broad claims

The 13,455-character amendment unit contained many explicit amendment markers spanning multiple statutes and clauses. The 33 compiled claims retained the general fact that consumer-protection provisions were being added, deleted or realigned, but several claims grouped multiple statutory operations into a single description.[E4] From the ontology alone, an investigator could not reliably reconstruct which clause, paragraph, citation or exception was changed by each operation.

Adding relations downstream would not repair this loss. The atomic amendment operations were not represented as separate upstream claims, so relation completion had no evidence-bearing endpoints from which to reconstruct them. The failure boundary was the first semantic projection, not the final graph materialization.

4.4 A product-membership matrix was flattened

The enforcement-decree provision defined different categories of professional financial consumers for deposit, loan, investment and insurance products. The ten compiled claims preserved the broad product classes and the fact that eligibility differed among them. They did not preserve the complete pairing matrix that specifies which person, institution, corporation or intermediary is included or excluded for each product class.[E5]

This is not a total extraction failure. It is a resolution mismatch. The generated claims were adequate for topical retrieval but insufficient for a query such as “Is institution x a professional consumer for insurance

product y?" When a coarse representation is promoted to authoritative truth, useful compression becomes an unrecoverable decision error.

5 Mechanisms of failure

5.1 Semantic non-identifiability

Formal ontology engineering begins with an explicit conceptual commitment[1]. Open-ended text does not generally contain a unique graph decomposition waiting to be recovered. The same sentence may validly be represented as an obligation, a procedure, an exception, an evidentiary requirement or a temporal rule, depending on the intended queries. Without an external task contract, "the correct ontology" is underdetermined.

The OWL 2 specification itself distinguishes structural equivalence from semantic equivalence[2]. That distinction is relevant here: two generated graphs may differ structurally while supporting similar inferences, or appear structurally consistent while omitting a source distinction required by a particular decision. A vector-similarity score does not resolve this mismatch.

5.2 Lossy projection into a finite schema

Lists, cross-references, exceptions, amendment instructions and role combinations must be projected into a finite inventory of objects, assertion types, relation types and slots. Every projection makes choices about granularity and ownership. Expanding the schema can represent additional cases, but it also increases mapping burden and does not provide a general guarantee of losslessness for unseen source forms.

5.3 Error lock-in across stages

Once a source distinction is absent from the semantic-card stage, later mapping, merging and relation completion operate on an already compressed state. Structural validators can improve self-consistency and remove malformed edges, but they cannot infer which unrepresented source proposition should have existed. If a validator rereads the raw source to find omissions, then the ontology has not replaced the source; the source remains the authority needed to judge the ontology.

5.4 Why model scaling is not a sufficient guarantee

Stronger models can reduce average extraction error, improve long-context handling and generate more useful candidate structures. Semantic-entropy methods can also help detect uncertainty across generations[7]. None of these improvements converts empirical average performance into a logical guarantee of completeness, uniqueness, stability and verifiability. In particular, representation choice and task-conditioned granularity are specification problems, not only model-capacity problems.

6 Discussion

6.1 What the evidence supports

The combined studies support three conclusions. First, repeated open-ended compilation can vary in meaning selection and graph construction. Second, a large compiler can achieve substantial structural integrity while retaining material source-fidelity gaps. Third, evidence linkage proves where a generated record came from, but not that every decision-relevant source proposition was generated.

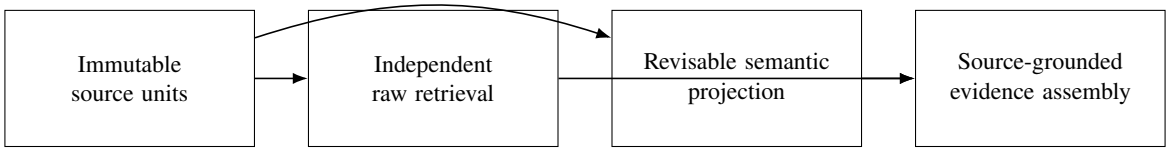

Figure 2. Source-first semantic control plane. Immutable source units retain factual authority. Raw retrieval and generated semantic projections provide independent candidate paths whose identifiers are expanded back to source text before final reasoning. Projection errors remain performance failures rather than silent corruption of the truth layer.

The strongest defensible conclusion is therefore operational: open-ended probabilistic semantic compilation alone is not a sufficient basis for guaranteeing a lossless canonical replacement of unstructured source text. This conclusion concerns architectural authority, not the usefulness of semantic projections.

6.2 What the evidence does not support

The evidence does not prove that every future model will fail to generate a useful ontology. It does not rule out constrained construction for a fixed query family, externally specified schema, deterministic identity policy and human adjudication. It also does not show that generated objects, aliases, relations or summaries are useless for retrieval. Finally, the unavailable Study A ledger prevents quantitative claims about the distribution or magnitude of repeated-run variance.

6.3 Threats to validity

The forensic audit focused on five failure-prone windows rather than a random sample of all 61 windows. It is therefore designed to localize mechanisms, not estimate corpus-wide prevalence. The corpus consists of Korean insurance law and may overrepresent dense amendments, enumerations and cross-references. Some loss judgements depend on the downstream decision task; a broad claim can be sufficient for discovery while insufficient for adjudication. The repeated-run study is reported from earlier internal experiments and cannot be independently reproduced from the retained artifacts. These limitations reduce the scope of prevalence claims but do not invalidate the observed examples of unrecoverable compression.

7 A source-first semantic control plane

The results suggest relocating authority rather than abandoning ontology. Immutable source units should remain the truth layer. Generated semantic structures should be treated as revisable projections used to discover, connect and rank source material. The final reasoning stage should expand selected identifiers back to raw source text before asserting a material claim.

This architecture retains the most valuable compiler assets: immutable source boundaries, versions, hierarchy, content hashes, stable high-confidence identities, source-local claims, evidence lineage and query expansion. It reduces or removes responsibilities that require unsupported completeness guarantees, including corpus-wide relation completion, exhaustive extension generation, uncertain global object merging and release states that declare semantic completeness.

The evaluation target must change accordingly. Compiler success should be measured by authoritative source-unit recall, query-conditioned context precision, final-answer correctness, citation entailment, ontology-omission recovery, latency, cost and auditability. Record counts and graph self-consistency remain useful diagnostics but are not primary measures of truth preservation.

8 Conclusion

We evaluated the hypothesis that careful language-model orchestration can recover unstructured source meaning as a stable, canonical and authoritative ontology. Repeated compilation reported semantic and structural variance. A large legal-domain compiler achieved clean evidence linkage and substantial structural integrity, yet source-level audits found material compression of amendment operations and product-specific membership rules.

The root failure was not simply insufficient model intelligence. The architecture assumed that an open-ended source had one recoverable graph decomposition and then assigned authority to a probabilistic projection. Better models may produce better maps, but a better map does not become the territory merely by being more accurate on average.

Immutable sources should therefore remain authoritative. Language-model ontologies are most defensible as semantic control planes: fallible, revisable structures that improve navigation and evidence discovery while preserving a direct path back to the source. This is not a retreat from ontology engineering. It is a narrower and more testable definition of what ontology can safely do in high-stakes systems.

Data availability

The repository-verified artifacts used for Studies B and C are retained in an internal AiNOS research repository and are enumerated in the evidence registry (Appendix A). They are not currently deposited in a public archive; a reproducible subset is available from the author on reasonable request. The raw per-run ledger for Study A was not retained.

Code availability

Compiler code and evaluation scripts are retained in the internal project repository. They are not currently released under a public archival identifier and are available from the author on reasonable request.

Author contributions

J.K. conceived the study, designed and built the compilation systems, performed the experiments and forensic audits, and wrote the manuscript.

Competing interests

The author is affiliated with AiNOS, which develops ontology-based artificial-intelligence systems and may benefit from the architectural conclusions reported here. No other competing interests are declared.

References

[1] Gruber, T. R. Toward principles for the design of ontologies used for knowledge sharing. *Int. J. Hum.-Comput. Stud.* **43**, 907–928 (1995). doi:10.1006/ijhc.1995.1081.

[2] Motik, B., Patel-Schneider, P. F. & Parsia, B. *OWL 2 Web Ontology Language: Structural Specification and Functional-Style Syntax* (Second Edition). W3C Recommendation (2012). `https://www.w3.org/TR/owl2-syntax/`.

[3] Hogan, A. *et al.* Knowledge graphs. *ACM Comput. Surv.* **54**, 71:1–71:37 (2021). doi:10.1145/3447772.

[4] Wang, X. *et al.* Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations* (2023).

[5] Lewis, P. *et al.* Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems* **33**, 9459–9474 (2020).

[6] Ji, Z. *et al.* Survey of hallucination in natural language generation. *ACM Comput. Surv.* **55**, 248:1–248:38 (2023). doi:10.1145/3571730.

[7] Farquhar, S., Kossen, J., Kuhn, L. & Gal, Y. Detecting hallucinations in large language models using semantic entropy. *Nature* **630**, 625–630 (2024). doi:10.1038/s41586-024-07421-0.

A Evidence registry

Table 3. Evidence classes and uses in this manuscript.

ID	Class	Use in this manuscript
E1	VERIFIED	Full compiler scale and frozen manifest
E2	VERIFIED	Five-window output counts and relation evidence cardinality
E3	VERIFIED	Five immutable window plans containing 116 source units
E4	VERIFIED	Amendment-unit character count, lexical markers and 33 claims
E5	VERIFIED	Enforcement-decree unit dimensions, list markers and ten claims
E6	VERIFIED	Architecture transition record
E7	REPORTED	Repetition protocol and qualitative variance only