Beyond Ontology Generation: Evaluating Multi-Level Modelling Semantics in LLM-Generated Ontologies

Introduction

- **Explore the potential of LLMs** for automated ontology generation and their ability to capture complex ontological structures.
- **Investigate semantic correctness** in LLM-generated ontologies, focusing on multi-level modelling principles.
- **Evaluate multi-level representations** with respect to abstraction-level correctness, meta-level representation, instantiation chains, and cross-level relationships.
- **Identify recurring semantic limitations** by comparing LLM-generated ontologies against gold-standard ontologies from the literature.
- **Inform improvements in LLMs** to enhance their understanding and generation of semantically correct multi-level ontological representations.

Main Contributions

- **MLT-based evaluation framework:** Defines semantic criteria for assessing abstraction levels, meta-level representations, relationships, and structural completeness.
- **Comparative empirical evaluation:** Evaluates three LLMs across four prompting strategies and three domains using IR metrics and MLT-based criteria.
- **Semantic error analysis:** Identifies recurring errors, including class–instance confusion, abstraction-level mixing, and invalid cross-level relationships.

Overview Of The Experiment

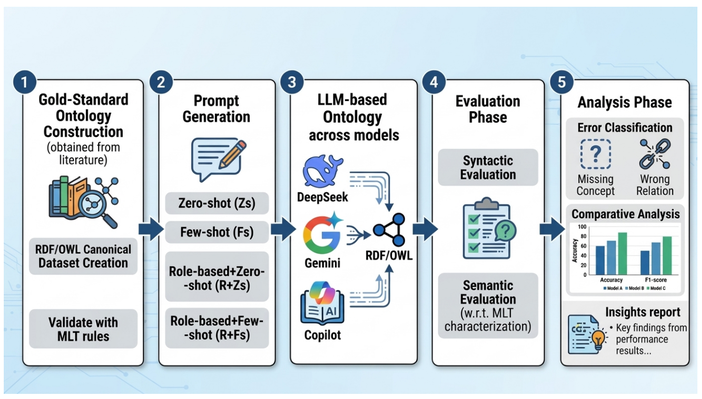


Figure 1. Overview of the experimental methodology for generating and evaluating multi-level ontologies using LLMs.

Semantic Evaluation Criteria

1. **Abstraction-Level Correctness:** Correct assignment of token individuals, first-order types, and higher-order types.
2. **Meta-Class Correctness:** Correct representation of higher-order types and class classification.
3. **Intra-Level Semantic Correctness:** Correct use of rdfs:subClassOf and mlt:isSubordinationOf
4. **Cross-Level Semantic Correctness:** Correct use of MLT relationships, including `mlt:isPowertypeOf` and `mlt:isCharacterizationOf`, across abstraction levels.
5. **Structural Completeness :** Coverage of the expected domain concepts and relationships.

Prompting Strategies Performance

We compare four prompting strategies: Zero-Shot (ZS), Role-Based Zero-Shot (R+ZS), Few-Shot (FS), and Role-Based Few-Shot (R+FS).

Mean MLT semantic compliance and violations by prompting strategy.

Strategy	C1	C2	C3	C4	C5	Mean	Violations
ZS	77.1	87.6	58.9	68.1	64.2	71.18	25.3
R+ZS	73.4	**92.9**	73.9	**72.5**	76.5	77.84	19.8
FS	**100.0**	87.8	78.8	64.0	81.3	82.38	**13.8**
R+FS	**100.0**	88.2	**81.2**	64.4	**82.3**	**83.22**	13.9

Key observations:

- Few-shot prompting achieves higher overall semantic compliance than zero-shot prompting.
- Role-based prompting improves meta-class correctness in zero-shot settings.
- R+FS achieves the highest overall mean (83.22%), while FS records the fewest violations (13.8).

Cross-Domain Evaluation

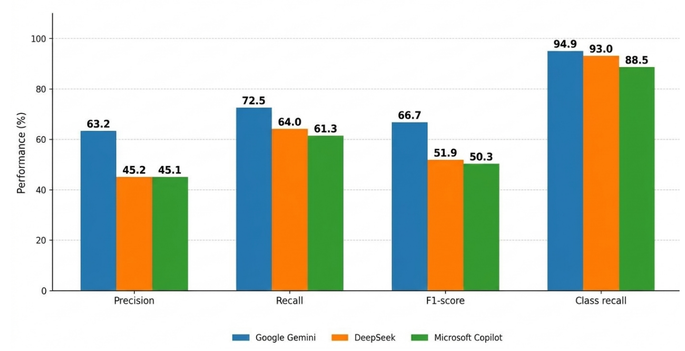
Semantic performance varies across the three domains, reflecting differences in the difficulty of representing multi-level structures.

Domain	C1	C2	C3	C4	C5	Violations
Taxonomy	**96.1**	90.8	73.1	**77.8**	72.0	15.6
Business	86.4	76.6	63.3	58.4	76.6	23.7
Person	80.5	**100.0**	**83.3**	65.5	**79.5**	**15.3**

Key observations:

- **Taxonomy** highest abstraction-level and cross-level scores.
- **Person** perfect meta-class correctness and the highest intra-level and structural scores.
- **Business** lowest semantic scores and the highest violation count.

LLM Performance Against Reference Ontologies



Comparative performance of Gemini, DeepSeek, and Copilot.

Recurring Multi-Level Modelling Errors

1. **Abstraction-Level Mixing:** Assigning entities incompatible roles across abstraction levels. Example: declaring `:Person` and `:Woman` as both OWL classes and named individuals.
2. **Category Inversion:** Confusing higher-order classification with subclassing. Example: modelling `:Species` and `:Genus` as subclasses of `:TaxonomicRank`, instead of using instantiation.
3. **Incorrect Intra-Level Relationships:** Misusing `mlt:isSubordinateOf` for ordinary class hierarchies, e.g., connecting `:FemaleManager`, `:MaleManager`, and `:Manager`.
4. **Incorrect Cross-Level Relationships:** Misusing `mlt:isPowertypeOf` and `mlt:isCharacterizationOf` through incorrect directions or endpoints.
5. **Incomplete and Excessive Modelling:** Omitting foundational MLT structures or generating unsupported entities and relationships. One output contained 134 triples against 60 in the reference, with 106 false positives and 20.9% precision.

Conclusion

- **LLM Limitations:** LLMs generate multi-level ontologies but exhibit severe semantic errors in cross-level relationships and intra-level modeling.
- **Prompting Impact:** Few-shot prompting improves semantic compliance, with **Role-Based Few-Shot** achieving the highest mean score.
- **Domain Variance:** Taxonomy and Person domains achieve comparable compliance, whereas Business presents greater modeling challenges.
- **Model Performance:** Google Gemini leads in overall Information Retrieval metrics (precision, recall, F1-score, and class recall).

COMENIUS UNIVERSITY BRATISLAVA

MATFYZ CONNECTIONS November 25, 2026 Bratislava, Slovakia

Zekeri Adams
Supervisors: Martin Homola & Ján Kluka
Comenius University Bratislava, Slovakia