LLM-Generated RDF Triples for Agricultural Species: A Comparative Evaluation Using AGROVOC Grounding

Leonardo Vianna do Nascimento[1]

[1]Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Sul (IFRS)
Campus Alvorada - Alvorada – RS – Brazil

leonardo.nascimento@alvorada.ifrs.edu.br

Abstract. *The construction of RDF knowledge bases for specialized domains is costly and requires close collaboration between domain experts and knowledge engineers. This paper evaluates three commercial LLMs — Claude, ChatGPT, and Gemini — in generating RDF Turtle triples for 38 plant species relevant to Brazilian agriculture. All models received an identical prompt combining five prompt engineering techniques, including few-shot exemplification and external file grounding via a CSV file with correct AGROVOC URIs. Outputs were assessed for AGROVOC URI precision, common name correctness against Embrapa reference sources, and syntactic conformance. Claude achieved perfect URI precision (100%) and the highest recall for common names (82.9%). ChatGPT reached the highest common name precision (96.1%) but poor URI precision (16.2%). Gemini showed similar recall (59.8%) and worse URI precision (8.1%). All models produced syntactically valid Turtle. Results indicate that grounding effectiveness varies across models and that programmatic URI validation is essential in LLM-assisted knowledge base construction.*

1. Introduction

Agricultural knowledge management increasingly relies on structured, machine-readable representations of domain information[Gong and Li 2025]. In Brazil, the agricultural sector plays a central role in the national economy, and the organization of data about cultivated species — including their botanical classification, common names, and associations with international controlled vocabularies — is essential for interoperability across research systems, databases, and public policies. Knowledge Graphs (KGs) built upon Semantic Web standards such as RDFS[1] and SKOS[2] provide a principled framework for this purpose, enabling data integration and reasoning across heterogeneous sources.

However, the manual construction of RDF knowledge bases is a time-consuming and costly process, typically requiring the involvement of both domain specialists and knowledge engineers[Pan et al. 2024]. This bottleneck has motivated growing interest in the use of Large Language Models (LLMs) as tools for knowledge acquisition and structured data generation. Recent advances in models such as Claude (Anthropic), ChatGPT (OpenAI), and Gemini (Google) have demonstrated remarkable capabilities in natural language understanding and generation, raising the question of whether these models can reliably produce well-formed, semantically accurate RDF triples when guided by carefully designed prompts[Mavridis et al. 2025].

[1]https://www.w3.org/TR/rdf-schema/
[2]https://www.w3.org/2004/02/skos/

Prompt engineering — the practice of designing inputs to elicit desired outputs from LLMs — has emerged as a key discipline in this context. Techniques such as few-shot prompting, domain context embedding, cardinality constraints, and grounding through external reference files have shown promise in improving the quality and consistency of LLM outputs for structured tasks[Brown et al. 2020, Schulhoff et al. 2024]. Nevertheless, systematic evaluations of these techniques applied to RDF generation in specialized domains remain scarce, particularly for the Portuguese-speaking context and agricultural ontologies.

This paper presents a comparative evaluation of three LLMs — Claude, ChatGPT, and Gemini — in the task of generating RDF Turtle triples for a set of 38 plant species relevant to Brazilian agriculture. Each model received the same prompt, which combined multiple prompt engineering techniques, including a few-shot example, embedded ontological context, property-level cardinality instructions, and an external CSV file containing the correct AGROVOC URIs for grounding. The generated outputs were evaluated along three dimensions: coverage and correctness of Brazilian Portuguese common names, syntactic conformance of the Turtle serialization, and precision of the exactMatch links to the AGROVOC controlled vocabulary. The results highlight the critical role of prompt engineering choices — particularly the use of external reference files — and suggest that the effectiveness of grounding depends significantly on the model's ability to prioritize provided context over internal knowledge.

The remainder of this paper is organized as follows. Section 2 reviews related work on Semantic Web standards, AGROVOC, and LLM-based knowledge graph construction. Section 3 describes the methodology, including the prompt design and evaluation protocol. Section 4 presents the results, and Section 5 discusses their implications. Section 6 concludes the paper and outlines directions for future work.

2. Related Works

The use of KGs for organizing agricultural information has a well-established history. A central resource in this context is AGROVOC [Caracciolo et al. 2013], maintained by the Food and Agriculture Organization of the United Nations (FAO). Born in the early 1980s as a multilingual agricultural thesaurus, AGROVOC has steadily evolved into a SKOS-XL concept scheme published as Linked Open Data, containing links and references to many other datasets. Its role as a shared controlled vocabulary makes it a natural grounding target for agricultural knowledge bases seeking interoperability.

More recent efforts have extended KG construction to specific agricultural subdomains [Drury et al. 2019]. KGs and ontologies have emerged as a powerful paradigm for integrating and managing diverse agricultural datasets at scale — from soil health and climate conditions to genetic information — and several efforts have produced domain-specific vocabularies and ontologies, including AgroPortal[3] and Planteome[4] projects, which serve as hubs for agricultural vocabularies and ontologies.

The manual construction of ontologies and knowledge graphs remains costly and labor-intensive. The involved limitations have motivated a growing body of research on

[3]https://agroportal.eu/
[4]https://planteome.org/

the use of LLMs as tools for automated or semi-automated knowledge acquisition. In the domain of RDF generation specifically, [Frey et al. 2023] conducted a benchmark study evaluating the proficiency of several LLMs in tasks involving RDF Turtle serialization, finding evidence that even leading commercial models show difficulties in adhering strictly to output formatting constraints when generating knowledge graphs.

More recently, [Mavridis et al. 2025] conducted a comparative evaluation of six systems — including GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro — for ontology mapping and RDF KG construction in the medical domain, using precision, recall, and F1-score as evaluation metrics. The present work extends this line of research to the agricultural domain, evaluating the same three commercial models on a structured RDF generation task grounded in an external controlled vocabulary.

The quality of LLM outputs for structured tasks is heavily influenced by the design of the input prompt. [Brown et al. 2020] established that few-shot prompting — providing input-output examples within the prompt — substantially improves model performance on a wide range of tasks. [Caufield et al. 2024] propose the SPIRES method (Structured Prompt Interrogation and Recursive Extraction of Semantics), which employs zero-shot learning to extract information from text and populate knowledge bases according to predefined schemas. Similarly, [Norouzi et al. 2025] investigate the effectiveness of LLMs in populating the Enslaved.org knowledge graph, reporting that models such as GPT-4 can recover approximately 90% of the expected triples when guided by modular ontologies embedded in the prompt. These studies emphasize that performance improves significantly when the ontology schema is explicitly provided in the prompt—an approach also adopted in the present work.

3. Methodology

This section describes the experimental setup used to evaluate the three LLMs, the design of the prompt submitted to each model, and the criteria used to assess the quality of the generated outputs.

3.1. Experimental Setup

Three large language models were evaluated: Claude Sonnet (Anthropic), ChatGPT (OpenAI), and Gemini (Google). All models were accessed through their respective official web interfaces using fresh sessions with no prior conversation history, ensuring that no contextual bias from previous interactions could influence the outputs. The same prompt and the same AGROVOC reference CSV file[5] were submitted to all three models under identical conditions, enabling direct comparison of the generated outputs.

The task consisted of generating a set of RDF triples in Turtle serialization for 38 plant species relevant to Brazilian agriculture, modelling each species as an individual of the class PlantSpecies of the COnto ontology[6] and populating it with properties defined in the SKOS and RDFS vocabularies. The target species were selected to represent the main horticultural crops cultivated and consumed in Brazil, covering families such as *Cucurbitaceae*, *Solanaceae*, *Fabaceae*, *Brassicaceae*, and *Apiaceae*, among others.

[5] `https://github.com/lvnascimento/encompif2026/blob/main/agrovoc.csv`
[6] `https://github.com/lvnascimento/sbbd2026/blob/main/conto.ttl`

3.2. Prompt Design

The prompt submitted to each model[7] combined five prompt engineering techniques, each targeting a specific dimension of output quality:

- *Explicit enumeration of input data*. All 38 species were listed by their scientific names within the prompt, accompanied by a reinforcement instruction specifying that all species must be included in the output. This was intended to prevent the model from truncating the response.
- *Embedded ontological context*. The formal definition of the target class *PlantSpecies*, including its RDF type declaration, multilingual labels, and description, was included in the prompt. This grounded the output in the target ontology and reduced the risk of the model inventing alternative class structures.
- *Property-level cardinality constraints*. Each property to be populated — *skos:prefLabel*, *skos:altLabel*, and *rdfs:comment* — was accompanied by explicit instructions regarding its expected cardinality and scope. In particular, *skos:altLabel* was constrained to Brazilian Portuguese common names of plants only, and *skos:prefLabel* was restricted to a single value per individual (the scientific name of each species).
- *External file grounding*. A CSV file containing the correct AGROVOC URIs for each species was provided alongside the prompt. The models were instructed to use this file as the sole source for populating the *skos:exactMatch* property, linking each individual to its corresponding concept in the AGROVOC controlled vocabulary.
- *Few-shot example*. A complete, well-formed Turtle individual — representing *Manihot esculenta* — was included as a concrete output example. This example covered all required properties and served as an implicit template for the structure, formatting, and language tags expected in the output.

3.3. Evaluation Dimensions

The outputs generated by each model were evaluated across three dimensions:

- *AGROVOC URI precision*. Each generated *skos:exactMatch* URI was compared programmatically against the reference values provided in the CSV file. A URI was considered correct only if it matched the reference exactly. This dimension was evaluated for 37 of the 38 species, as *Barbarea verna* had no corresponding entry in the reference file.
- *Common name coverage*. The skos:altLabel values generated by each model were evaluated against two reference sources maintained by Embrapa: the *Brazilian Vegetable Catalog*[8], which covers the 50 most commercially relevant species in Brazil and associates each with its common names and regional variants, and the Embrapa species portal, which provides complementary nomenclature data for individual crops. The comparison was conducted for all 38 species and assessed three aspects: correctness of the generated names (whether each name is genuinely used in Brazil for the indicated species), coverage of relevant names that

[7] Available at `https://github.com/lvnascimento/encompif2026/blob/main/prompt.txt`

[8] `https://www.infoteca.cnptia.embrapa.br/infoteca/bitstream/doc/887213/1/Catalogohortalicas.pdf`

were not generated by the model, and incorrect regional attribution (names associated with the wrong species or with regions outside Brazil). The total number of *skos:altLabel* values per model is also reported as a proxy for output richness.

- *Turtle syntactic conformance*. Each output file was inspected for syntactic validity of the Turtle serialization, including correct use of prefixes, proper termination of triples, well-formed language tags, and absence of structural errors.

4. Results

This section presents the results of the evaluation across the three dimensions defined in Section 3: AGROVOC URI precision, common name coverage and correctness, and Turtle syntactic conformance.

4.1. AGROVOC URI Precision

Table 1 summarizes the results of the programmatic comparison between the skos:exactMatch URIs generated by each model and the reference values provided in the AGROVOC CSV file, for the 37 species with a known correct URI.

Tabela 1. AGROVOC URI precision by model.

Model	Correct URIs	Incorrect URIs	Precision
Claude	37	0	100.0%
ChatGPT	6	31	16.2%
Gemini	3	34	8.1%

Claude achieved perfect precision, generating the correct URI for all 37 species. ChatGPT and Gemini performed substantially worse, with precision of 16.2% and 8.1% respectively. Notably, the incorrect URIs produced by both models were numerically plausible — close to the correct values but not matching them — suggesting that these models relied on parametric knowledge acquired during training rather than consulting the reference file provided in the prompt. This pattern was consistent across species and model versions, and constitutes the most significant finding of this study.

4.2. Common Name Coverage and Correctness

Table 2 presents the results of the evaluation of *skos:altLabel* values against the two Embrapa reference sources, reporting total names generated, correct names, incorrect names, absent names, precision, and recall for each model.

Tabela 2. Common name evaluation against Embrapa reference sources.

Model	Generated	Correct	Incorrect	Absent	Precision	Recall
Claude	116	92	24	19	79.3%	82.9%
ChatGPT	51	49	2	36	96.1%	57.6%
Gemini	64	52	12	35	81.3%	59.8%

The results reveal a clear trade-off between precision and recall across models. ChatGPT achieved the highest precision (96.1%), generating very few incorrect names, but at the cost of low recall (57.6%) — it produced a conservative set of names, missing 36

names present in the reference sources. Claude adopted the opposite strategy, generating the largest total number of names (116) and achieving the highest recall (82.9%), though with lower precision (79.3%) due to 24 incorrect names. Gemini occupied an intermediate position in terms of volume (64 names), but its recall (59.8%) remained close to that of ChatGPT, suggesting that generating more names did not translate proportionally into better coverage of the reference set.

4.3. Turtle Syntactic Conformance

All three models produced syntactically valid Turtle files. No prefix declaration errors, malformed triple terminations, invalid language tags, or structural inconsistencies were detected in any of the outputs. This result suggests that current commercial LLMs are broadly capable of adhering to the syntactic constraints of the Turtle serialization format when provided with clear prompt instructions and a concrete output example.

5. Discussion

The results presented in Section 4 allow for a nuanced analysis of the strengths and limitations of each model across the four evaluation dimensions, and raise broader questions about the conditions under which LLMs can be reliably used for structured knowledge base construction.

AGROVOC URI generation and the grounding problem. The most striking finding of this study is the near-complete failure of ChatGPT and Gemini to utilize the external reference file for AGROVOC URI generation, despite receiving the same CSV file and the same grounding instruction as Claude. The incorrect URIs produced by both models were not random — they were numerically close to the correct values, which strongly suggests that these models retrieved URI patterns from parametric knowledge acquired during pre-training rather than consulting the provided file. This behavior is consistent with findings in the hallucination literature, which has documented that LLMs frequently prioritize internal knowledge over externally provided context, particularly when the model has prior exposure to the domain. Claude's perfect precision on this dimension suggests that it was more effective at following the grounding instruction, though the reasons for this difference — whether architectural, related to instruction-following training, or to differences in AGROVOC exposure during pre-training — cannot be determined from this experiment alone and constitute an open question for future work.

The precision-recall trade-off in common name generation. The results for *skos:altLabel* reveal a consistent trade-off between precision and recall that reflects distinct generation strategies across models. ChatGPT adopted a conservative approach, generating few names per species but with high accuracy, which may reflect a tendency to prioritize confidence over coverage. Claude generated the largest volume of names and achieved the best recall, but at the cost of lower precision — a pattern that may be partly explained by the model including names used in other Portuguese-speaking regions, names associated with cultivated varieties rather than the species as a whole, or names that are plausible but not attested in the reference sources. Gemini's intermediate volume did not translate into proportionally better recall than ChatGPT, suggesting that simply generating more names is not sufficient if the additional names do not align with the reference. For a knowledge base intended for Brazilian agriculture, recall is arguably the

more critical metric, as missing names reduce the findability of species records; however, incorrect names introduce semantic noise that may propagate errors downstream.

Syntactic conformance. The universal syntactic validity of the three outputs is a positive finding that suggests current commercial LLMs have largely internalized the formal constraints of Turtle serialization.

This study has several limitations that should be acknowledged. The evaluation was conducted using the web interfaces of the three models at a specific point in time; results may differ with other versions or access methods. The common name evaluation relied on two Embrapa reference sources, which, while authoritative, may not cover the full range of regional names used across Brazil. Finally, only one prompt design was evaluated; a systematic comparison of prompt variants — for instance, with and without the few-shot example, or with and without the external CSV file — would provide stronger evidence about the contribution of each technique to output quality, and is left for future work.

6. Conclusion

This paper presented a comparative evaluation of three large language models — Claude, ChatGPT, and Gemini — in the task of generating RDF Turtle triples for a set of 38 plant species relevant to Brazilian agriculture. The evaluation assessed three dimensions of output quality: precision of links to the AGROVOC controlled vocabulary, coverage and correctness of Brazilian Portuguese common names, and syntactic conformance of the Turtle serialization.

The results revealed substantial differences among the three models. Claude achieved perfect precision in AGROVOC URI generation (100%), the highest recall for common names (82.9%), and the closest adherence to the prompt's structural conventions. ChatGPT achieved the highest precision for common names (96.1%) but low recall (57.6%) and poor AGROVOC URI precision (16.2%). Gemini performed similarly to ChatGPT in recall (59.8%) and worse in AGROVOC URI precision (8.1%). All three models produced syntactically valid Turtle files, suggesting that current commercial LLMs have broadly internalized the formal constraints of this serialization format.

The most significant finding of this study concerns the grounding behavior of the models. Despite all three receiving the same external CSV file with correct AGROVOC URIs, only Claude consistently utilized this reference, while ChatGPT and Gemini appear to have overridden the provided context with parametric knowledge acquired during pre-training. This result has direct practical implications: external file grounding for controlled vocabulary URIs cannot be assumed to work uniformly across models, and programmatic validation of generated URIs is a necessary step in any LLM-assisted knowledge base construction pipeline.

This work contributes to the growing body of research on LLM-based knowledge graph construction by providing an evaluation grounded in a real agricultural domain use case, using authoritative Brazilian reference sources for common name validation and a formal controlled vocabulary for URI verification. The prompt design evaluated here — combining few-shot examples, embedded ontological context, cardinality constraints, and external file grounding — represents a replicable methodology that can be extended to other domains and vocabularies.

Future work includes the systematic evaluation of prompt variants to isolate the contribution of individual techniques, the extension of the approach to the broader cultivar knowledge base of which this species vocabulary is a component, the inclusion of additional models such as open-source LLMs, and the application of the methodology to other agricultural subdomains and languages.

Referências

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Caracciolo, C., Stellato, A., Morshed, A., Johannsen, G., Rajbhandari, S., Jaques, Y., and Keizer, J. (2013). The agrovoc linked dataset. *Semantic Web*, 4(3):341–348.

Caufield, J. H., Hegde, H., Emonet, V., Harris, N. L., Joachimiak, M. P., Matentzoglu, N., Kim, H., Moxon, S., Reese, J. T., Haendel, M. A., et al. (2024). Structured prompt interrogation and recursive extraction of semantics (spires): A method for populating knowledge bases using zero-shot learning. *Bioinformatics*, 40(3):btae104.

Drury, B., Fernandes, R., Moura, M.-F., and de Andrade Lopes, A. (2019). A survey of semantic web technology for agriculture. *Information Processing in Agriculture*, 6(4):487–501.

Frey, J., Meyer, L.-P., Arndt, N., Brei, F., and Bulert, K. (2023). Benchmarking the abilities of large language models for rdf knowledge graph creation and comprehension: how well do llms speak turtle? In *ISWC: Workshop Deep Learning for Knowledge Graphs*.

Gong, R. and Li, X. (2025). The application progress and research trends of knowledge graphs and large language models in agriculture. *Computers and electronics in agriculture*, 235:110396.

Mavridis, A., Tegos, S., Anastasiou, C., Papoutsoglou, M., and Meditskos, G. (2025). Large language models for intelligent rdf knowledge graph construction: results from medical ontology mapping. *Frontiers in Artificial Intelligence*, 8:1546179.

Norouzi, S. S., Barua, A., Christou, A., Gautam, N., Eells, A., Hitzler, P., and Shimizu, C. (2025). Ontology population using llms. In *Handbook on Neurosymbolic AI and Knowledge Graphs*, pages 421–438. IOS Press.

Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., and Wu, X. (2024). Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3580–3599.

Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., Li, Y., Gupta, A., Han, H., Schulhoff, S., et al. (2024). The prompt report: A systematic survey of prompt engineering techniques. *arXiv preprint arXiv:2406.06608*.