Inferring Semantic Relations Between Terms with Large Language Models

Giulia Speranza
University of Naples "L'Orientale" - Via Chiatamone, 60/61 - 80121 Naples (Italy)
gsperanza@unior.it

Abstract

The purpose of this paper is to investigate the ability of Large Language Models (LLMs) to identify relations among terms, with the goal of facilitating and accelerating the construction of thesauri and terminological resources. We investigate whether the use of LLMs in this context can provide a valuable initial set of relations, serving as a basis upon which professional terminologists can build, validate, and enrich domain-specific knowledge representations.

1 Introduction

The identification and formalization of semantic relations among terms constitute a fundamental task in the development and refinement of thesauri and terminological resources.

In any specialized knowledge domain, terms do not exist in isolation but form a complex net of semantic relations. These relationships enable the structuring of domain knowledge, facilitate precise communication, and support tasks such as indexing, searching, and reasoning.

Early foundational work in terminology science, such as by Wüster (1975), emphasized the importance of defining clear hierarchical relations (broader-narrower) and associative relations to support knowledge organization and retrieval. Recent theoretical developments, such as Frame-Based Terminology (Faber, 2022), further highlight the central role of semantic relations in the organization of specialized knowledge.

Systematic terminological standards, including ISO 25964-1:2011 and ISO 704:2022 for thesauri, codify relations among terms into three main types:

- Hierarchical Relations: represent relations of superordination and subordination, where the superordinate concept represents a class or whole, and subordinate concepts refer to its members or parts. These relations form taxonomies or classification hierarchies that structure knowledge from general to specific. The most common forms are Broader Term (BT) and Narrower Term (NT);
- Equivalence Relations: establish links between terms that are near-synonyms within the domain, enabling consolidation of lexical variants and preventing ambiguity.
- Associative Relations: denote semantic connections that are neither hierarchical nor equivalence-based but are meaningful in context. Examples include Related Terms (RT) according to relationships of cause-effect, part-whole, or functionally related terms.

Formal models such as SKOS (Simple Knowledge Organization System) have standardized the representation of semantic relations in thesauri, providing an interoperable framework for semantic web applications (Miles and Bechhofer, 2009).

This work aims to analyze and compare the out-of-the-box performance of two LLMs (ChatGPT and Gemini) in identifying terminological semantic relations, with the aim of supporting the development of thesauri and other terminological resources. The case study and the gold standard dataset is in the domain of building materials and constructions.

2 Related Works

Historically, thesauri have been constructed manually by domain experts, who defined semantic relations based on domain knowledge. Hierarchical structures were typically conceptual while associative relations were more complex, often reflecting functional, causal, or contextual connections.

One of the first attempts to automatically identify hierarchical relations was proposed by Hearst (1992), who introduced lexico-syntactic patterns (e.g., "X such as Y", "Y is a type of X") as indicators of hyponymy.

One influential approach in this area is the use of knowledge patterns, as introduced by Meyer (2001), which have been instrumental in the extraction and identification of semantic relations. Automatic systems based on this approach have been developed to support the extraction of semantic relations from corpora. Notable examples include the EcoLexicon Semantic Sketch Grammar (ESSG) (León-Araúz and Martín, 2018), which applies knowledge patterns within a frame-based approach to identify and organize conceptual relations in domain-specific texts, and Corpógrafo (Maia and Matos, 2008), a tool designed to facilitate corpus-based terminological analysis by detecting and visualizing term relationships and co-occurrence patterns.

Studies by Cimiano et al. (2005); Velardi et al. (2013); Spiteri (2002) extended this by combining pattern-based extraction with distributional semantics, enabling more nuanced identification of various semantic relations beyond simple hierarchies.

In parallel, graph-based methods have gained traction for modeling and identifying term relations (Velardi et al., 2013). Tools like WordNet (Miller, 1992) exemplify a rich network of lexical-semantic relations structured as a graph, which has influenced the design of domain-specific thesauri.

Identifying these relations often requires domain expertise combined with semi-automated methods, such as supervised learning approaches trained on annotated corpora.

Studies such as Petroni et al. (2019) have shown that LLMs can retrieve encyclopedic facts through cloze tasks (e.g., "Paris is the capital of ___"), but the ability to infer more abstract conceptual relations (e.g., hypernymy, meronymy) remains underexplored.

Recently, some works have analyzed the implicit presence of structured semantic knowledge in LLMs. For example, Davison et al. (2019) investigated the extent to which LLMs can correctly answer questions about relations between entities. However, these studies almost always rely on rich contextual input (sentences, definitions, knowledge triples), whereas our work focuses on a minimal setting in which the model receives only a list of terms and must infer possible relations.

Trained on massive text corpora, LLMs have demonstrated a strong capacity to understand and generate human-like text. They are also capable of handling a wide range of linguistic tasks with little need for explicit guidance. As a result, recent studies have increasingly explored their effectiveness across a variety of NLP tasks. These capabilities raise a compelling question: can LLMs effectively support or even partially automate the process of finding relations among terms for terminological resource construction purposes?

In this regard, despite a growing interest in the use of LLMs in Automatic Terminology Extraction tasks for the creation of resources, glossaries, and thesaurus (Banerjee et al., 2024; Tran et al., 2024), there is still no systematic evaluation of their ability to infer semantic relations from simple term lists, without context, examples, or explicitly provided background knowledge.

3 Case study

Our study explores the zero-shot capabilities of LLMs in inferring relations between terms, without access to external texts or the use of fine-tuning techniques. We analyze the LLMs ability in this task by comparing their output against a thesaurus about construction materials of monuments, buildings, and structures taken as Gold Standard reference. No examples are provided in the prompt; the model relies only on prior knowledge.

3.1 LLM Selection

We selected two state-of-the-art large language models: ChatGPT and Gemini in order to conduct a comparative evaluation of different models' designs in the context of terminology work. These models were chosen based on their widespread adoption, usage, and testing in various Natural Language Processing (NLP) tasks, their architectures, and training paradigms such as Reinforcement Learning from Human Feedback.

ChatGPT is developed by OpenAI and is based on the GPT architecture. For this study, we used the GPT-4 version, which represents a significant advance over previous iterations in terms of coherence and domain generalization capabilities. GPT-4 was trained on a massive corpus of publicly available text and licensed data using a transformer-based decoder architecture. It is capable of few-shot and zero-shot learning, meaning it can perform tasks with little to no task-specific fine-tuning, relying instead on prompt engineering.

Gemini is developed by Google DeepMind and represents the evolution of the earlier PaLM (Pathways Language Model) family. It is a multimodal LLM, trained not only on text but also on images

and code, although our experiment utilizes the text-only capabilities of Gemini. The model is designed for factual grounding, task adaptability, native multimodality and long context window.

This approach allows us to compare the different LLMs strengths and weaknesses in handling the task and provides a broader perspective on how general-purpose language models can be employed in terminological tasks.

To simulate and closely reproduce a real-world use case scenario, we chose to interact with the selected LLMs in the same way a professional terminographer would in a practical work environment. This means that we did not apply any fine-tuning, domain-specific retraining, or technical modifications to the models. Instead, we relied on and evaluated exclusively their out-of-the-box capabilities, using the standard user interfaces provided by the respective platforms. This approach reflects the typical conditions under which language professionals, such as terminologists, translators, or linguists, without strong technical and engineering skills, would engage with LLM to perform terminology tasks. Our objective was to evaluate the actual usability and effectiveness of the models in supporting terminology work without requiring advanced technical expertise or custom integration efforts.

3.2 Prompting strategy

To guide the language models in generating domain-relevant and semantically coherent output, we adopted a persona prompting strategy. This approach involves framing the prompt in such a way that the model is instructed to assume the role of a specific expert or domain specialist—referred to as a persona. By embedding this expert persona into the prompt, we aimed to steer the models' responses toward more precise, terminologically consistent, and semantically appropriate outputs. The persona was specified at the beginning of the prompt and further contextualized with task-specific instructions, such as identifying hierarchical, associative, and equivalence relations among domain-specific terms. This technique leverages the models' ability to adapt to match the expectations associated with a particular professional role. To force the model to act like a specific person, adopting a certain perspective on the task to be performed, one effective strategy is to provide the persona's job title, which should elicit a set of associated attributes and competencies. The 'persona-pattern' or 'role-play' prompting techniques have been widely used in several studies for different tasks (Kong et al., 2023; Olea et al., 2024; Mzwri and Turcsányi-Szabo, 2025) as it is much more accessible, compared to fine-tuning the model from an engineering point of view. This prompting approach belongs to the so-called 'Output Customization category' according to White et al. (2023) or to the 'LLM Role-Playing' category according to Tseng et al. (2024).
For our experiment, we queried the language models using the prompt detailed in Example 1. It's important to highlight that all experimental sessions were carried out in May 2025.

LLM Prompt

```
You are an expert terminologist specialized
in the domain of building materials. You
are given a simple list of domain-specific
terms ordered alphabetically. Your task
is to identify all relevant semantic
relations (Broader Term, Narrower Term,
Related Term) among the terms in the list.

Instructions:

  1. Only consider relations valid within
     the context of building materials and
     construction.

  2. Use only the terms provided in the
     list: don't add or omit any term.

  3. Each relation should be directional
     where applicable (e.g., Term A is a
     broader term for Term B).
```

Example 1. Prompt with instructions

3.3 Dataset

As a case study, we used the FISH Building Materials Thesaurus[1], a controlled vocabulary employed for documenting primary construction materials of monuments, buildings, and structures associated with the built and buried heritage of the British Isles. This thesaurus is maintained by Historic England on behalf of the FISH (Forum on Information Standards in Heritage) Terminology Working Group and is openly available for reuse. It can be downloaded in multiple formats, including PDF (both alphabetical and hierarchical listings), CSV, and semantic web-compatible linked-data formats. A sample of the dataset can be found in Example 2.

[1] https://heritagedata.org/live/schemes/eh_tbm.html

GRANITE

SN: A hard, coarse grained, durable igneous rock. Can be used decoratively or as a main material. Found through-out the country.

BT: STONE

NT: ABERDEEN GRANITE, MOORSTONE, MOUNTSORREL GRANITE, PETERHEAD GRANITE, SHAP GRANITE

RT: ELVAN, GNEISS, GREENSTONE, PORPHYRY.

Example 2. Example of the FISH Building Materials Thesaurus

We fed the LLMs only with the list of terms (prefLabel) ordered alphabetically, without explicitly expressing the relations. This list of 232 terms (including TopConcepts) was obtained by querying the SPARQL endpoint. The complete dataset, including all term relations (broader, narrower, related), was retained as a Gold Standard reference for the subsequent evaluation phase.

4 Results and Evaluation

The following examples showcase representative outputs produced by each model, offering qualitative insight into their ability to identify semantic relations among the listed terms.

GRANITE

BT: Stone

NT: Aberdeen Granite, Granite, Peterhead Granite, Shap Granite

RT: Gneiss, Diorite, Feldspar

Example 3. ChatGPT output

GRANITE

BT: Stone

NT: Aberdeen Granite, Mountsorrel Granite, Peterhead Granite, Shap Granite

RT: Elvan, Gneiss, Greenstone

Example 4. Gemini output

To assess the ability of LLMs to correctly identify the semantic relation between terms, we adopted standard classification metrics such as Precision, Recall and F-Measure (the harmonic mean of Precision and Recall). The results of our evaluation are presented in Table 1, where we compare the performance of ChatGPT and Gemini.

Model	Prec. (%)	Rec. (%)	F1 (%)
ChatGPT	55.0	65.0	59.6
Gemini	75.4	84.2	79.6

Table 1: Evaluation results of LLM outputs

The comparative evaluation of ChatGPT and Gemini highlighted a significant difference in their ability to capture hierarchical depth in conceptual structures.

The recall is slightly lower than precision across models, indicating a tendency to be conservative in identifying relations, often abstaining when unsure.

In addition to the quantitative metrics, a qualitative evaluation was carried out to understand the nature of the errors and the strengths of each model.

While both LLMs performed almost adequately in detecting first-level (direct) conceptual relations, such as hypernymy (e.g., 'granite' or 'limestone' is a type of 'stone'), none of the models showed great performance in capturing deeper taxonomic structures and inferring multi-step hierarchical relations, which can in some cases be very complex and reach deep levels, as in Figure1.

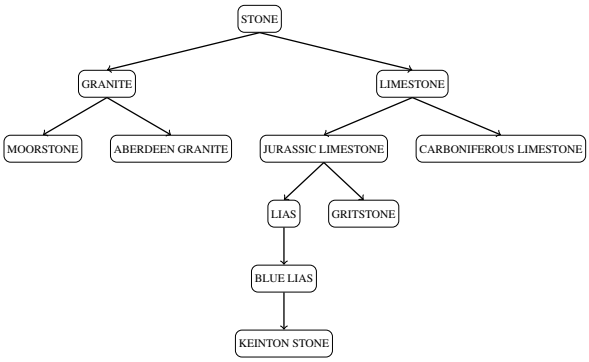

Figure 1: Hierarchical relations of stone types

Generally speaking, Gemini retains and operationalizes more granular domain knowledge, possibly due to a more extensive training dataset or improved alignment mechanisms. Conversely, ChatGPT tended to limit more often its inferences to surface-level relations and classifications (e.g., stone → limestone) and often failed to recognize nested or domain-specific sub-classifications, particularly when terms were more technical or less frequent. These findings suggest that Gemini may have stronger capabilities in ontology-oriented processing, making it better suited for tasks involving the structuring or expansion of specialized terminologies.

Both LLMs, however, may still provide valuable support to the terminographer, especially in the identification of first-level semantic relations—such as broader term (BT), narrower term (NT), and related term (RT) links—across well-represented conceptual domains. While Gemini may offer more depth in modeling nested hierarchies and domain-specific nuances, ChatGPT can nonetheless assist in generating baseline classifications or verifying established semantic patterns, particularly when supplemented with domain-specific prompts or curated input examples.

5 Conclusion and Future Works

This study investigated the ability of LLMs to infer semantic relations between terms in a minimal-input, zero-shot setting, without access to external context, training data, or fine-tuning.

Our results suggest that, while LLMs demonstrate promising capabilities in identifying first-level types of conceptual relations, they still struggle with a deeper level (second and third level) of analysis.

One of the key contributions of this work is to highlight that LLMs encode a certain degree of structured semantic knowledge that can be activated even in the absence of linguistic context or definitional cues. At the same time, our findings underline the limits of this implicit competence, especially when models are required to infer second-level hierarchies.

From a terminological practice perspective, these insights suggest that LLMs could serve as lightweight tools for semi-automatic relation extraction, particularly in the early stages of resource construction or when dealing with low-resource domains. However, their use should be complemented with expert validation or human-in-the-loop, given the high degree of variability and wrong hierarchy structuring.

As future works, several directions emerge from this initial study: further analysis could take into account different prompting strategies (e.g., few-shot, chain-of-thought) to improve the quality and explainability of relational inferences. Moreover, extending the range of relation types (e.g., made-of, causes, used-for) would allow for a broader assessment of the models' semantic competence. Future experiments may also focus on reproducing this experiment on different domain-specific term lists (e.g., in medicine, law, or engineering) as well as on other languages in order to generalize the output results and evaluation.

References

Shubhanker Banerjee, Bharathi Raja Chakravarthi, and John Philip McCrae. 2024. Large language models for few-shot automatic term extraction. In *International Conference on Applications of Natural Language to Information Systems*, pages 137–150. Springer.

Philipp Cimiano, Andreas Hotho, and Steffen Staab. 2005. Learning concept hierarchies from text corpora using formal concept analysis. *Journal of artificial intelligence research*, 24:305–339.

Joe Davison, Joshua Feldman, and Alexander Rush. 2019. Commonsense knowledge mining from pretrained models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1173–1178, Hong Kong, China. Association for Computational Linguistics.

Pamela Faber. 2022. Frame-based terminology. In *Theoretical Perspectives on Terminology*, pages 353–376. John Benjamins Publishing Company.

Marti A. Hearst. 1992. Automatic acquisition of hyponyms from large text corpora. In *COLING 1992 Volume 2: The 14th International Conference on Computational Linguistics*.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2023. Better zero-shot reasoning with role-play prompting. *arXiv preprint arXiv:2308.07702*.

Pilar León-Araúz and A San Martín. 2018. The ecolexicon semantic sketch grammar: from knowledge patterns to word sketches. *arXiv preprint arXiv:1804.05294*.

Belinda Maia and Sérgio Matos. 2008. Corpografo v4-tools for researchers and teachers using comparable corpora. In *quot; In Pierre Zweigenbaum; Éric Gaussier; Pascale Fung (ed) Proceedings of the 6 th International Conference on Language Resources and Evaluation; LREC 2008 Workshop on Comparable Corpora (LREC 2008)(Marrakech 28-30 May 2008; 31 May 2008) European Language Resources Association (ELRA); 79-82*. European Language Resources Association (ELRA).

Ingrid Meyer. 2001. Extracting knowledge-rich contexts for terminography: A conceptual and methodological framework. In *Recent advances in computational terminology*, pages 279–302. John Benjamins Publishing Company.

Alistair Miles and Sean Bechhofer. 2009. *SKOS Simple Knowledge Organization System Reference*. W3C Recommendation. World Wide Web Consortium, United States.

George A. Miller. 1992. WordNet: A lexical database for English. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

Kovan Mzwri and Márta Turcsányi-Szabo. 2025. The impact of prompt engineering and a generative ai-driven tool on autonomous learning: A case study. *Education Sciences*, 15(2):199.

Carlos Olea, Holly Tucker, Jessica Phelan, Cameron Pattison, Shen Zhang, Maxwell Lieb, Doug Schmidt, and Jules White. 2024. Evaluating persona prompting for question answering tasks. In *Proceedings of th e 10th international conference on artificial intelligence and soft computing, Sydney, Australia*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473., Hong Kong, China. Association for Computational Linguistics.

L Spiteri. 2002. Word association testing and thesaurus construction: Defining inter-term relationships. In *Proceedings of the 30th Annual Conference of the Canadian Association for Information Science*, pages 24–33.

Hanh Thi Hong Tran, Carlos-Emiliano González-Gallardo, Julien Delaunay, Antoine Doucet, and Senja Pollak. 2024. Is prompting what term extraction needs? In *International Conference on Text, Speech, and Dialogue*, pages 17–29. Springer.

Yu-Min Tseng, Yu-Chao Huang, Teng-Yun Hsiao, Wei-Lin Chen, Chao-Wei Huang, Yu Meng, and Yun-Nung Chen. 2024. Two tales of persona in llms: A survey of role-playing and personalization. *arXiv preprint arXiv:2406.01171*.

Paola Velardi, Stefano Faralli, and Roberto Navigli. 2013. Ontolearn reloaded: A graph-based algorithm for taxonomy induction. *Computational Linguistics*, 39(3):665–707.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.

Eugen Wüster. 1975. Die ausbildung in terminologie und terminologischer lexikographie.