

Identifying Semantic Relationships Between Research Topics Using Large Language Models in a Zero-Shot Learning Setting

Tanay Aggarwal^{1,*}, Angelo Salatino¹, Francesco Osborne^{1,2} and Enrico Motta¹

¹Knowledge Media Institute, The Open University, Milton Keynes, UK

²Department of Business and Law, University of Milano Bicocca, Milan, IT

Abstract

Knowledge Organization Systems (KOS), such as ontologies, taxonomies, and thesauri, play a crucial role in organising scientific knowledge. They help scientists navigate the vast landscape of research literature and are essential for building intelligent systems such as smart search engines, recommendation systems, conversational agents, and advanced analytics tools. However, the manual creation of these KOSs is costly, time-consuming, and often leads to outdated and overly broad representations. As a result, researchers have been exploring automated or semi-automated methods for generating ontologies of research topics. This paper analyses the use of large language models (LLMs) to identify semantic relationships between research topics. We specifically focus on six open and lightweight LLMs (up to 10.7 billion parameters) and use two zero-shot reasoning strategies to identify four types of relationships: *broader*, *narrower*, *same-as*, and *other*. Our preliminary analysis indicates that Dolphin2.1-OpenOrca-7B performs strongly in this task, achieving a 0.853 F1-score against a gold standard of 1,000 relationships derived from the IEEE Thesaurus. These promising results bring us one step closer to the next generation of tools for automatically curating KOSs, ultimately making the scientific literature easier to explore.

Keywords

Zero-Shot Learning, Large Language Models, Ontology Generation, Research Topics, Scholarly Knowledge, Scientific Knowledge Graphs,

1. Introduction

Knowledge Organization Systems (KOS), like ontologies, taxonomies, and thesauri, designed for research topics, are instrumental for structuring, managing, and retrieving information from digital libraries, enabling efficient knowledge discovery [1]. Major publishers like ACM, IEEE, PubMed, and SpringerNature employ KOS like the ACM Computing Classification System¹, the IEEE Thesaurus, Medical Subject Headings, and the SpringerNature Taxonomy to organise their published content [2, 3, 4]. KOSs play also a crucial role in enabling intelligent systems to navigate and interpret academic literature effectively [5, 6], serving as the foundation for

Sci-K 2024: 4th International Workshop on Scientific Knowledge: Representation, Discovery, and Assessment 11/12 November 2024 - Baltimore, MD, USA


*Corresponding author.

✉ tanay.aggarwal@open.ac.uk (T. Aggarwal); angelo.salatino@open.ac.uk (A. Salatino);

francesco.osborne@open.ac.uk (F. Osborne); enrico.motta@open.ac.uk (E. Motta)

🆔 0009-0009-9477-7112 (T. Aggarwal); 0000-0002-4763-3943 (A. Salatino); 0000-0001-6557-3131 (F. Osborne);

0000-0003-0015-1952 (E. Motta)

 © 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

¹ACM Computing Classification System - <https://dl.acm.org/ccs>

tools like search engines [6], conversational agents [7], analytics dashboards [8], and academic recommender systems [5]. Additionally, KOS provide a robust representation of research topics, crucial for various AI-driven analyses of scientific literature [9, 10].

However, maintaining these KOSs is a growing challenge. The rapid expansion of scientific literature, with an estimated 2.5 million new papers published annually [11], necessitate continuous curation to reflect the latest advancements. Besides, manually curating them has become increasingly costly and time-consuming, highlighting the need for innovative solutions to keep pace with the evolving scientific landscape.

In contrast, the emergence of Large Language Models (LLMs) has revolutionised the field of Artificial Intelligence (AI), enabling deeper language comprehension, including grasping the semantic of word and sentences, inferring relationships between concepts, resolving ambiguities, and understanding overall text meaning [12].

Given the challenges in maintaining up-to-date ontologies of research areas, and the recent advancements in AI, this paper presents an analysis showcasing the capabilities of LLMs in identifying semantic relationships between pairs of research topics. Specifically, we focus on six open and lightweight models (up to 10.7 billion parameters), and we employ two zero-shot reasoning strategies to identify four relationships types (e.g., *broader*, *narrower*, *same-as*, and *other*), on a gold standard of 1,000 relationship (250 per relationship).

Our main objective is to develop an innovative pipeline to automatically generate and maintain ontologies of research topics. This pipeline will allow us to curate existing ontologies, create more granular representations of research concepts, and expand the development of ontologies into other scientific fields. In this context, our research focuses on determining whether LLMs can effectively aid in this process.

In brief, this paper contributes to the literature with a preliminary analysis of whether zero-shot reasoning can effectively identify semantic relationships between research topics using open, smaller models, with a focus on sustainability. The gold standard and the code we used to run our experiments are available on a GitHub repository².

The remainder of this paper is structured as follows. Section 2 provides an overview of the literature. Section 3 describes the dataset and the approach we have devised to conduct our experiments. Section 4 presents and discusses our results. Finally, in Section 5 we conclude the paper and provide future directions.

2. Literature Review

In this section, we review the literature focusing on two key aspects relevant to our work: ontologies of research areas and automatic ontology generation.

2.1. Ontologies of research areas

In the literature there are several ontologies, or more in general KOSs, within the scientific ecosystem, which can support the exploration process in digital libraries, the production of scholarly analytics, and modelling research dynamics [2, 13]. These include Medical Subject

²Gold Standard and Code for experiments - <https://github.com/ImTanay/LLM-Semantic-Relationship-Analysis>

Headings (MeSH) [3], ACM Computing Classification System (CCS), the Computer Science Ontology (CSO) [13], AGROVOC, Mathematical Subject Classification (MSC) [2], and Physics Subject Headings (PhySH) [14]. MeSH is a comprehensive controlled vocabulary, with more than 30K concepts, developed and maintained by the National Library of Medicine [3]. It is widely used in the medical and health sciences, and it receives yearly updates. The ACM Computing Classification System³ is a taxonomy of research topics in the field of Computer Science, covering about 2K research topics. It is curated by ACM, the world's largest educational and scientific computing society, and the last update dates back to 2012. CSO is the largest ontology of research topic in Computer Science, covering 14K research areas [13]. It has been automatically generated using the Klink-2 algorithm [15] on a dataset of 16 million scientific articles, and receives yearly updates. The IEEE Thesaurus mainly covers the field of Engineering but also contains different concepts relevant to Computer Science. It is curated by the Institute of Electrical and Electronics Engineers⁴. It contains around 5.6K topics and 24K relationships, and receives yearly updates. MSC is a comprehensive taxonomy with over 6.5K concepts, covering a wide range of mathematical disciplines, from pure mathematics to applied mathematics and statistics. It is curated by the American Mathematical Society and zbmATH, and receives updates every 10 years [2].

Most of the existing KOSs are curated manually, usually by a committee of domain experts who periodically meet to discuss the updates for the next version. This process, however, makes ontologies evolve slowly and hence prone to becoming outdated. Additionally, such a manual curation is costly and increasingly unsustainable given the rapid rate at which new research is published [11, 13]. Automating the creation of research area ontologies can overcome existing limitations by ensuring these ontologies remain current. This, in turn, can enhance cataloguing, retrievability, and the various downstream applications mentioned earlier.

2.2. Automatic Ontology Generation

The automatic generation of ontologies is a field of research whose objective is to overcome the challenges of traditional ontology creation, hence making it more scalable and efficient. Techniques that are usually employed include natural language processing, clustering techniques, or statistical methods [16, 17, 15]. For instance, Text2Onto [16] is a tool that can create ontologies from a corpus of documents. This method identifies synonyms, sub-/superclass hierarchies, and through the application of NLP techniques it can learn hierarchical structures between terms, leveraging phrases such as “such as...” and “and other...”.

Shan et al. [18] used a similar method to develop Fields of Study for Microsoft Academic, combining manually created concepts with topics derived from Wikidata. However, this approach relied heavily on Wikidata and did not utilize metadata associated with research papers. The OpenAlex team [19] also employed a similar strategy, expanding upon the “All Science Journal Classification” structure in Scopus and incorporating topics extracted from papers using citation analysis.

Some studies have explored a hybrid approach, combining ontology learning with crowdsourcing to integrate statistical measures and user feedback [20, 21]. Specifically, human support

³The ACM Computing Classification System – <http://www.acm.org/publications/class-2012>

⁴IEEE Taxonomy - <https://www.ieee.org/content/dam/ieee-org/ieee/web/org/pubs/ieee-taxonomy.pdf>

has been integrated to evaluate an automatically generated ontology [20].

More recently, researchers have begun to leverage LLMs for the creation of taxonomies, ontologies, and knowledge graphs [22]. For instance, Chen et al. [23] proposed a two-module approach for taxonomy generation. The first module predicts parent-child relationships, and the second module reconciles these predictions into tree structures. The model is trained on subtrees from Wordnet and evaluated on separate Wordnet subtrees.

Given the new opportunities unlocked by LLMs, we hypothesise that significant progress can be made to tackle the ongoing challenges of curating ontologies of research areas.

3. Material and Methods

This section describes the task, the gold standard, and the employed LLMs.

3.1. Task definition and experiments

The task is to classify the semantic relationship holding between pairs of research topics (t_A, t_B) according to four categories which are essential for ontology generation. More formally, this is single-label multi-class classification problem, and the categories are:

- *broader*: t_A is a parent topic of t_B . E.g., *ontological languages* is a broader area than *owl*
- *narrower*: t_A is a child topic of t_B . E.g., *nosql* is a specific area within *databases*
- *same-as*: t_A and t_B can be used interchangeably to refer to same concept. E.g., *haptic interface* and *haptic device*
- *other*: t_A and t_B do not relate according to the above categories. E.g., *blockchain* and *user interfaces*

Our experiments consist of two zero-shot reasoning strategies, as depicted in Fig. 1.

One-way Strategy: This experiment, highlighted by the red dashed box in Fig. 1, involves taking each pair of research topics, generating a prompt using a specially designed template (see Appendix A), submitting it to the LLM, and then analysing the response to determine the appropriate classification.

The prompt template is identical for both strategies and all models. It was carefully refined through an iterative process to ensure optimal performance and consistency across all models.

Two-way Strategy: This experiment, highlighted by the green dashed box in Fig. 1, involves running the one-way strategy twice. First, we identify the relationship between topics t_A and t_B . Then, in a fresh context, it identifies the relationship when the topics are swapped. This is possible because the relationships *broader* and *narrower* are inverses of each other, *same-as* is symmetric, and by definition also *other* is symmetric.

Finally, we set empirical rules (cyan box in Fig. 1) to mitigate the agreement/disagreement between the two branches of the two-way strategy. These rules are designed with the aim of prioritising the development of the hierarchical structure. Let $f(X)$ and $s(X)$ represent the relationship types returned by the first and second branches of our two-way strategy, respectively. Additionally, let $\text{len}(t_A)$ denote the length of the topic’s surface form. We defined the rules as follow:

1. broader :- f(broader) \wedge s(narrower)
2. narrower :- f(narrower) \wedge s(broader)
3. broader :- ((f(narrower) \wedge s(narrower)) \vee (f(broader) \wedge s(broader))) \wedge len(t_A) \leq len(t_B)
4. narrower :- ((f(narrower) \wedge s(narrower)) \vee (f(broader) \wedge s(broader))) \wedge len(t_A) $>$ len(t_B)
5. same-as :- f(same-as) \wedge s(same-as)
6. broader :- (f(broader) \wedge s(other)) \vee (f(other) \wedge s(narrower))
7. narrower :- (f(narrower) \wedge s(other)) \vee (f(other) \wedge s(broader))
8. :- f(X)

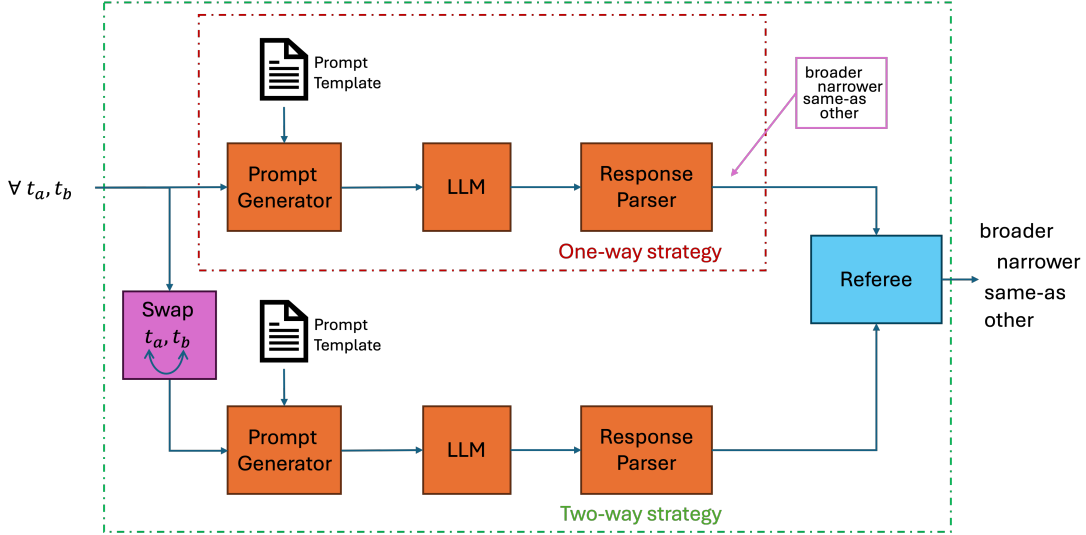


Figure 1: Architecture of our two strategies. The first strategy (red dashed box) determines the relationship between t_a and t_b in one way, whereas the second strategy (green dashed box) determines the relationship between pairs of topics in both ways.

3.2. Gold standard

As gold standard, we selected a balanced sample of 1,000 semantic relationships from the IEEE Thesaurus, with 250 relationships per category. The IEEE Thesaurus includes 11,570 descriptive terms in the field of *Engineering*, and serves as a standardised vocabulary of technical terms for indexing and retrieving content within the IEEE digital library. Specifically, we utilised the IEEE Thesaurus v1.02⁵, which is dated to July 2023, and is available as a PDF file following the ANSI/NISO Z39.4-2021 standard [24].

To generate our gold standard, we first extracted the hierarchical structure and relationships between terms from the original IEEE Thesaurus PDF document and transformed it into RDF

⁵A copy of version 1.02 of the IEEE Thesaurus is available at https://github.com/angelosalatino/ieee-taxonomy-thesaurus-rdf/blob/main/source/ieee-thesaurus_2023.pdf.

format⁶. We represented the various relationships according to the Simple Knowledge Organization System (SKOS) notation: i) `skos:broader`, ii) `skos:narrower`, iii) `skos:altLabel`, iv) `skos:prefLabel`, v) `skos:related`.

We randomly sampled 250 relationships each for the categories *broader*, *narrower*, and *same-as*. For the latter, we used a combination of `skos:altLabel` and `skos:prefLabel`. Finally, we randomly coupled topics to generate 250 *other* relationships, ensuring that these new pairs did not share any of the previously established semantic relationships within the IEEE Thesaurus.

3.3. Large Language Models

For our experiments we selected six quantised LLMs, that had been quantised to 8-bit precision. These included four fine-tuned versions of Mistral-7B, in addition to SOLAR and LLaMa 3.

Dolphin-2.1-Mistral-7B: (shortened as *dolphin-mistral*⁷) is a decoder-only model with 7 billion parameters and a token context capacity of 4096. It is based on Mistral-7B and fine-tuned with the Dolphin⁸ dataset, which is an open-source implementation of Microsoft’s Orca [25], with an addition of Airoboros⁹ dataset.

Dolphin-2.6-Mistral-7B-dpo-laser: (shortened as *dolphin-mistral-dpo*¹⁰) is based on Mistral-7B and fine-tuned on top of Dolphin DPO using Layer Selective Rank Reduction (LASER) [26]. It is a decoder-only model with 7 billion parameters, offering a context window of 4096 tokens.

Dolphin2.1-OpenOrca-7B: (shortened as *dolphin-openorca*¹¹) is a model that blends Dolphin-2.1-Mistral-7B and Mistral-7B-OpenOrca¹² models. These models were merged using the “ties merge” [27] technique, keeping the same number of training parameters and token context window size, respectively 7 billion and 4096.

OpenChat-3.5-0106-Gemma: (shortened as *openchat-gemma*¹³) is a model trained on *openchat-3.5-0106*¹⁴ data using Conditioned Reinforcement Learning Fine-Tuning (C-RLFT) framework [28, 29]. This model shares the same properties as *openchat-3.5-0106*, which is a decoder-only model fine-tuned¹⁵ on top of Mistral-7B. It consists of 7 billion parameters with a context window size of 8192 tokens.

SOLAR-10.7B-Instruct-v1.0: (shortened as *solar*¹⁶) is a 10.7 billion parameters model with a Depth Up-Scaling (DUS) architecture, which includes architectural modifications and continued pretraining [30].

Llama-3-8B-Instruct : (shortened as *llama-3*¹⁷) is an auto-regressive model based on transformer architecture. The updated versions of this model utilises supervised fine-tuning (SFT)

⁶The code we employed for converting the IEEE Thesaurus in RDF is available here <https://github.com/angelosalatino/ieee-taxonomy-thesaurus-rdf>.

⁷Dolphin-2.1-Mistral-7B - <https://huggingface.co/TheBloke/dolphin-2.1-mistral-7B-GGUF>

⁸Dolphin Dataset - <https://huggingface.co/datasets/cognitivecomputations/dolphin>

⁹Airoboros - <https://huggingface.co/datasets/jondurbin/airoboros-2.1>

¹⁰Dolphin-2.6-Mistral-7B-dpo-laser - <https://huggingface.co/TheBloke/dolphin-2.6-mistral-7B-dpo-laser-GGUF>

¹¹Dolphin2.1-OpenOrca-7B - <https://huggingface.co/TheBloke/Dolphin2.1-OpenOrca-7B-GGUF>

¹²Mistral-7B-OpenOrca - <https://huggingface.co/Open-Orca/Mistral-7B-OpenOrca>

¹³OpenChat-3.5-0106-Gemma - <https://huggingface.co/gguf/openchat-3.5-0106-gemma-GGUF>

¹⁴openchat-3.5-0106 - <https://huggingface.co/openchat/openchat-3.5-0106>

¹⁵Huggingface - <https://huggingface.co/openchat/openchat-3.5-1210/discussions/4#658288f1168803bdee13d6b3>

¹⁶SOLAR-10.7B-Instruct-v1.0 - <https://huggingface.co/TheBloke/SOLAR-10.7B-Instruct-v1.0-GGUF>

¹⁷Llama-3-8B-Instruct - <https://huggingface.co/lmstudio-community/Meta-Llama-3-8B-Instruct-GGUF>

and reinforcement learning with human feedback (RLHF). This model has 8 billion parameters and a context length of 8k tokens [31].

All these models are openly available on Huggingface. To run them, we used Google Colab, equipped with Nvidia’s V100 and L4 GPU(s), and used KoboldCpp¹⁸, which is a software tool to operate with LLMs. KoboldCpp is a standalone program built on llama.cpp that makes models accessible via an API endpoint.

4. Results and Discussion

We assessed the performance of the LLMs using precision, recall, and F1-score. Table 1 reports the results of the one-way strategy. All models excel in precision (≥ 0.85) for the “same-as” relationships, although most struggle with recall, except for the solar model (recall = 0.784). The models also demonstrate high precision for “narrower” and high recall for “broader”. For the “other” relationship, llama-3 and openchat-gemma exhibit high precision, dolphin-mistral and dolphin-mistral-dpo exhibit high recall, whereas dolphin-openorca excels in both precision and recall.

Notably, dolphin-openorca emerges as the top performer with an average precision of 0.780, recall of 0.733, and F1-score of 0.724. Indeed, dolphin-openorca has high precision for “narrower”, “same-as”, and “other”, and high recall for “broader” and “other”.

Table 1

Performance values of our experiments when employing the one-way strategy. BR are the performance associated with *broader* relationships, NA to *narrower*, SA to *same-as*, OT to *others*, and AVG are average performance considering the four types of relationships. In **bold** are the best performing models for a given class.

MODEL	PRECISION					RECALL					F1-SCORE				
	AVG	BR	NA	SA	OT	AVG	BR	NA	SA	OT	AVG	BR	NA	SA	OT
dolphin-mistral	0.676	0.613	0.736	0.901	0.455	0.599	0.652	0.412	0.508	0.824	0.599	0.632	0.528	0.650	0.586
dolphin-mistral-dpo	0.734	0.613	0.868	0.960	0.497	0.603	0.988	0.236	0.284	0.904	0.552	0.757	0.371	0.438	0.641
dolphin-openorca	0.780	0.612	0.811	0.960	0.737	0.733	0.952	0.496	0.576	0.908	0.724	0.745	0.615	0.720	0.814
openchat-gemma	0.679	0.396	0.471	1.000	0.849	0.523	0.968	0.296	0.268	0.560	0.506	0.562	0.364	0.423	0.675
solar	0.723	0.642	0.877	0.850	0.523	0.657	0.916	0.544	0.784	0.384	0.646	0.755	0.672	0.529	0.627
llama-3	0.715	0.560	0.446	1.000	0.853	0.582	0.732	0.808	0.256	0.532	0.568	0.634	0.575	0.408	0.655

The two-way strategy yielded dramatic improvements in average precision, recall, and F1-score, as reported in Table 2. Specifically, the values of precision for “broader” and recall for “narrower” are significantly higher, effectively addressing the weaknesses of the one-way strategy. Once again, dolphin-openorca emerges as the top performer with an impressive average F1-score of 0.853, due to consistently high scores across all relationship types: 0.841 (broader), 0.829 (narrower), 0.845 (same-as), and 0.897 (other).

Table 3 demonstrates that by exploiting the symmetric nature of the analysed semantic relationships, the two-way strategy leads to a considerable improvement in F1-scores for most models. This improvement is approximately 7% for solar and openchat-gemma, whereas for the dolphin family models it exceeds 10%, with dolphin-mistral-dpo reaching a notable 24.3%

¹⁸KoboldCpp - <https://github.com/LostRuins/koboldcpp>

Table 2

Performance values of our experiments when employing the one-way strategy. BR are the performance associated with *broader* relationships, NA to *narrower*, SA to *same-as*, OT to *others*, and AVG are average performance considering the four types of relationships. In **bold** are the best performing models for a given class.

MODEL	PRECISION					RECALL					F1-SCORE				
	AVG	BR	NA	SA	OT	AVG	BR	NA	SA	OT	AVG	BR	NA	SA	OT
dolphin-mistral	0.718	0.686	0.673	0.884	0.629	0.704	0.708	0.732	0.704	0.672	0.706	0.697	0.701	0.764	0.664
dolphin-mistral-dpo	0.834	0.759	0.764	0.968	0.844	0.808	0.956	0.932	0.476	0.868	0.795	0.846	0.840	0.638	0.856
dolphin-openorca	0.862	0.799	0.774	0.928	0.947	0.852	0.888	0.892	0.776	0.852	0.853	0.841	0.829	0.845	0.897
openchat-gemma	0.734	0.484	0.512	1.000	0.939	0.590	0.804	0.792	0.332	0.432	0.579	0.605	0.622	0.499	0.592
solar	0.723	0.731	0.753	0.784	0.623	0.715	0.836	0.816	0.508	0.700	0.710	0.780	0.783	0.617	0.659
llama-3	0.746	0.476	0.544	1.000	0.964	0.586	0.856	0.836	0.328	0.324	0.562	0.611	0.659	0.494	0.485

Table 3

Summary of the average F1-scores of the two strategies. In **bold** are the best performing models for a given class.

MODEL	Two-way strategy	One-way strategy	Diff. from one-way to two-way
dolphin-mistral	0.706	0.598	0.108
dolphin-mistral-dpo	0.794	0.551	0.243
dolphin-openorca	0.853	0.723	0.130
openchat-gemma	0.579	0.505	0.074
solar	0.709	0.645	0.064
llama-3	0.562	0.568	-0.006

increase. In contrast, llama-3 appears to be the only model not significantly impacted by the choice of strategy.

5. Conclusions and Future Work

In this paper, we evaluated the ability of six LLMs in identifying the semantic relationships between research concepts, comparing their performance to a gold standard of 1,000 relationships from the IEEE Thesaurus. Our results demonstrate that state-of-the-art models like dolphin-openorca achieve excellent zero-shot reasoning performance (0.853 of F1-score).

Our future work will focus on several directions. Currently, we are expanding our analysis to include additional models, such as non-quantised models, models with larger numbers of parameters (e.g., LLaMa 3 70B), and proprietary models like ChatGPT 4.0 and the Claude family. Furthermore, we plan to investigate fine-tuning some models (e.g., Google’s Gemma), incorporate a reasoner into our pipeline to identify and resolve inconsistencies, and extend our analysis to other scientific fields like Material Science, Medicine, and Physics.

Acknowledgments

We would like to thank Alessia Pisu, PhD Student from the University of Cagliari (IT) who helped us in converting the IEEE Thesaurus from PDF to RDF.

References

- [1] A. Salatino, T. Aggarwal, A. Mannocci, F. Osborne, E. Motta, A survey on knowledge organization systems of research fields: Resources and challenges, arXiv preprint arXiv:2409.04432 (2024).
- [2] E. Dunne, K. Hulek, Mathematics subject classification 2020, EMS Newsletter 2020–3 (2020) 5–6. URL: <http://dx.doi.org/10.4171/NEWS/115/2>. doi:10.4171/news/115/2.
- [3] C. E. Lipscomb, Medical subject headings (mesh), Bulletin of the Medical Library Association 88 (2000) 265.
- [4] B. Rous, Major update to acm’s computing classification system, Communications of the ACM 55 (2012) 12–12.
- [5] J. Beel, B. Gipp, S. Langer, C. Breitingner, Paper recommender systems: a literature survey, International Journal on Digital Libraries 17 (2016) 305–338.
- [6] M. Gusenbauer, N. R. Haddaway, Which academic search systems are suitable for systematic reviews or meta-analyses? evaluating retrieval qualities of google scholar, pubmed, and 26 other resources, Research synthesis methods 11 (2020) 181–217.
- [7] A. Meloni, S. Angioni, A. Salatino, F. Osborne, D. R. Recupero, E. Motta, Integrating conversational agents and knowledge graphs within the scholarly domain, Ieee Access 11 (2023) 22468–22489.
- [8] S. Angioni, A. Salatino, F. Osborne, D. R. Recupero, E. Motta, The aida dashboard: a web application for assessing and comparing scientific conferences, IEEE Access 10 (2022) 39471–39486.
- [9] J. W. Goodell, S. Kumar, W. M. Lim, D. Pattnaik, Artificial intelligence and machine learning in finance: Identifying foundations, themes, and research clusters from bibliometric analysis, Journal of Behavioral and Experimental Finance 32 (2021) 100577.
- [10] A. Salatino, S. Angioni, F. Osborne, D. R. Recupero, E. Motta, Diversity of expertise is key to scientific impact: a large-scale analysis in the field of computer science, arXiv preprint arXiv:2306.15344 (2023).
- [11] L. Bornmann, R. Mutz, Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references, Journal of the Association for Information Science and Technology 66 (2015) 2215–2222. doi:<https://doi.org/10.1002/asi.23329>.
- [12] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners, 2023. URL: <https://arxiv.org/abs/2205.11916>. arXiv:2205.11916.
- [13] A. A. Salatino, T. Thanapalasingam, A. Mannocci, F. Osborne, E. Motta, The computer science ontology: a large-scale taxonomy of research areas, in: The Semantic Web–ISWC 2018: 17th International Semantic Web Conference, Monterey, CA, USA, October 8–12, 2018, Proceedings, Part II 17, Springer, 2018, pp. 187–205.
- [14] A. Smith, Physics subject headings (physh), KO KNOWLEDGE ORGANIZATION 47 (2020) 257–266.
- [15] F. Osborne, E. Motta, Klink-2: Integrating multiple web sources to generate semantic topic networks, in: M. Arenas, O. Corcho, E. Simperl, M. Strohmaier, M. d’Aquin, K. Srinivas, P. Groth, M. Dumontier, J. Heflin, K. Thirunarayan, K. Thirunarayan, S. Staab (Eds.), The Semantic Web - ISWC 2015, Springer International Publishing, Cham, 2015, pp. 408–424.
- [16] P. Cimiano, J. Völker, Text2onto, in: A. Montoyo, R. Muñoz, E. Métais (Eds.), Natural Lan-

- guage Processing and Information Systems, Springer Berlin Heidelberg, Berlin, Heidelberg, 2005, pp. 227–238.
- [17] M. Le, S. Roller, L. Papaxanthos, D. Kiela, M. Nickel, Inferring concept hierarchies from text corpora via hyperbolic embeddings, in: A. Korhonen, D. Traum, L. Màrquez (Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 3231–3241. doi:10.18653/v1/P19-1313.
 - [18] Z. Shen, H. Ma, K. Wang, A web-scale system for scientific knowledge exploration, in: F. Liu, T. Solorio (Eds.), Proceedings of ACL 2018, System Demonstrations, Association for Computational Linguistics, Melbourne, Australia, 2018, pp. 87–92. doi:10.18653/v1/P18-4015.
 - [19] OpenAlex, Openalex: End-to-end process for topic classification, 2024. URL: <https://docs.google.com/document/d/1bDopkhuGieQ4F8gGNj7sEc8WSE8mvLZS>.
 - [20] G. Wohlgenannt, A. Weichselbraun, A. Scharl, M. Sabou, Dynamic integration of multiple evidence sources for ontology learning, *Journal of Information and Data Management* 3 (2012) 243–254.
 - [21] J. Mortensen, M. Musen, N. Noy, Crowdsourcing the verification of relationships in biomedical ontologies, *Annual Symposium proceedings / AMIA Symposium. AMIA Symposium 2013* (2013) 1020–9.
 - [22] B. P. Allen, L. Stork, P. Groth, Knowledge engineering using large language models, *arXiv preprint arXiv:2310.00637* (2023).
 - [23] C. Chen, K. Lin, D. Klein, Constructing taxonomies from pretrained language models, in: North American Chapter of the Association for Computational Linguistics, 2020. URL: <https://api.semanticscholar.org/CorpusID:233992529>.
 - [24] Ansi/niso z39.4-2021, criteria for indexes, 2021. URL: <http://dx.doi.org/10.3789/ansi.niso.z39.4-2021>. doi:10.3789/ansi.niso.z39.4-2021.
 - [25] S. Mukherjee, A. Mitra, G. Jawahar, S. Agarwal, H. Palangi, A. Awadallah, Orca: Progressive learning from complex explanation traces of gpt-4, *arXiv preprint arXiv:2306.02707* (2023).
 - [26] P. Sharma, J. T. Ash, D. Misra, The truth is in there: Improving reasoning in language models with layer-selective rank reduction, *arXiv preprint arXiv:2312.13558* (2023).
 - [27] P. Yadav, D. Tam, L. Choshen, C. A. Raffel, M. Bansal, Ties-merging: Resolving interference when merging models, *Advances in Neural Information Processing Systems* 36 (2024).
 - [28] T. de Bruin, J. Kober, K. Tuyls, R. Babuška, Fine-tuning deep rl with gradient-free optimization, *IFAC-PapersOnLine* 53 (2020) 8049–8056.
 - [29] G. Wang, S. Cheng, X. Zhan, X. Li, S. Song, Y. Liu, Openchat: Advancing open-source language models with mixed-quality data, *arXiv preprint arXiv:2309.11235* (2023).
 - [30] D. Kim, C. Park, S. Kim, W. Lee, W. Song, Y. Kim, H. Kim, Y. Kim, H. Lee, J. Kim, C. Ahn, S. Yang, S. Lee, H. Park, G. Gim, M. Cha, H. Lee, S. Kim, Solar 10.7b: Scaling large language models with simple yet effective depth up-scaling, 2023. *arXiv:2312.15166*.
 - [31] AI@Meta, Llama 3 model card (2024). URL: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

A. Prompt

For consistency we applied the same prompt across all the models. We engineered this prompt through various refinements, to ensure optimal comprehension of the task and accurate responses.

Below is the template of our prompt, customised for each topic pair by substituting [TOPIC-A] for the first topic and [TOPIC-B] for the second.

Classify the relationship between '[TOPIC-A]' and '[TOPIC-B]' by applying

→ the following relationship definitions:

1. '[TOPIC-A]' is-broader-than '[TOPIC-B]' if '[TOPIC-A]' is a
→ super-category of '[TOPIC-B]', that is '[TOPIC-B]' is a type, a branch,
→ or a specialized aspect of '[TOPIC-A]' or that '[TOPIC-B]' is a tool or
→ a methodology mostly used in the context of '[TOPIC-A]' (e.g., car
→ is-broader-than wheel).
2. '[TOPIC-A]' is-narrower-than '[TOPIC-B]' if '[TOPIC-A]' is a sub-category
→ of '[TOPIC-B]', that is '[TOPIC-A]' is a type, a branch, or a
→ specialized aspect of '[TOPIC-B]' or that '[TOPIC-A]' is a tool or a
→ methodology mostly used in the context of '[TOPIC-B]' (e.g., wheel
→ is-narrower-than car).
3. '[TOPIC-A]' is-same-as-than '[TOPIC-B]' if '[TOPIC-A]' and '[TOPIC-B]'
→ are synonymous terms denoting an identical concept (e.g., beautiful
→ is-same-as-than attractive), including when one is the plural form of
→ the other (e.g., cat is-same-as-than cats).
4. '[TOPIC-A]' is-other-than '[TOPIC-B]' if '[TOPIC-A]' and '[TOPIC-B]'
→ either have no direct relationship or share a different kind of
→ relationship that does not fit into the other defined relationships.

Given the previous definitions, determine which one of the following

→ statements is correct:

1. '[TOPIC-A]' is-broader-than '[TOPIC-B]'
2. '[TOPIC-B]' is-narrower-than '[TOPIC-A]'
3. '[TOPIC-A]' is-narrower-than '[TOPIC-B]'
4. '[TOPIC-B]' is-broader-than '[TOPIC-A]'
5. '[TOPIC-A]' is-same-as-than '[TOPIC-B]'
6. '[TOPIC-A]' is-other-than '[TOPIC-B]'

Answer by only stating the correct statement and its number.