Benchmarking Resource-Efficient LLMs for Research Topic Ontology Generation in the Biomedical Field

Tanay Aggarwal[1,*], Angelo Salatino[1], Francesco Osborne[1,2] and Enrico Motta[1]

[1]Knowledge Media Institute, The Open University, Milton Keynes, UK
[2]Department of Business and Law, University of Milano-Bicocca, Milan, IT

Abstract

Knowledge Organization Systems like Ontologies and taxonomies are fundamental for structuring scientific knowledge, yet their manual curation presents a persistent bottleneck in knowledge management. While Large Language Models (LLMs) offer a scalable mechanism for automated ontology generation, their capacity to classify complex, domain-specific semantics requires systematic evaluation. In this paper, we assess the performance of five small, open-source LLMs (up to 9 billion parameters) in identifying semantic relationships between biomedical concepts. To support this evaluation, we introduce *MeSH-Rel-4K*, a dataset comprising 4K semantic relationships extracted from the Medical Subject Headings (MeSH). We analyse three adaptation strategies: standard prompting, Chain-of-Thought prompting, and fine-tuning. While parameter-constrained models traditionally struggle with the nuances of in-context logic, our results reveal that targeted fine-tuning increases the average F1-score by 34.1 percentage points. These results confirm that direct fine-tuning effectively exceeds the reasoning bottlenecks of smaller LLMs, providing an accurate, automated methodology for the construction and evolution of specialised biomedical ontologies.

Keywords

Large Language Models, Knowledge Organization Systems, Biomedical Ontologies, Fine-Tuning, Ontology Generation, Scientific Knowledge Graphs

1. Introduction

Knowledge Organization Systems (KOSs), such as ontologies and taxonomies, have emerged as essential frameworks for systematically structuring and interlinking information. Within the research ecosystem, these systems are particularly critical for mapping topics, thereby enhancing data retrievability and management in digital libraries [1]. These systems are utilised by major publishers like Springer Nature[1], IEEE[2], and PubMed[3] [2, 3, 4] to categorise research outputs, e.g., research articles, books, monographs. Further, KOSs support smart downstream applications to navigate and interpret academic literature, such as recommender systems [5], classifiers [6], search engines [7], conversational agents [8], and analytics dashboards [9].

However, as shown by Salatino et al. [1] existing ontologies of research areas exhibit uneven coverage and are updated infrequently, causing them to lag behind rapidly emerging research areas. These limitations stem from the creation and maintenance of ontologies that require coordinated manual effort from domain experts over prolonged periods [10]. Previous approaches to automate ontology generation encounter structural limitations when processing the granular complexity of scientific terminologies [11, 12, 10, 13]. Consequently, the development of domain-specific ontologies remains predominantly a manual process.

Workshop on LLM-driven Knowledge Graph and Ontology Engineering (llms4kgoe) 2026 co-located with ESWC 2026
*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)

[1]The SpringerNature Taxonomy to organise their published content
[2]The IEEE Theasurus is used to classify academic articles and research within the IEEE digital library
[3]Medical Subject Headings (MeSH) is used to organise academic content on PubMed

Recent advances in Large Language Models (LLMs) present alternative methodologies for ontology generation [14, 15, 16]. In this paper, we analyse the performance of five small, open-source LLMs (up to 9 billion parameters) in identifying semantic relations between pairs of biomedical research topics. We test three adaptation strategies: standard prompting, Chain-of-Thought (CoT) prompting [17], and parameter-efficient fine-tuning [18]. Our analysis tests the capacity of LLMs to identify three explicit semantic relations: `broader`, `narrower`, and `same-as`, alongside an `other` category for unrelated concepts.

To support this investigation, we extracted 4K semantic relationships from MeSH, introducing *MeSH-Rel-4K*. Our experiments demonstrate that while CoT prompting yields moderate baseline improvements, explicit fine-tuning systematically increases classification accuracy. The fine-tuned `gemma-2-9b-it` achieved an F1-score of 91.6%, while the smallest model (`Llama-3.2-3B-Instruct`) recorded a 60.5 percentage point increase in its F1-score post-fine-tuning.

We release our complete codebase and the *MeSH-Rel-4K* dataset in our GitHub repository[4].

The rest of the paper is structured as follows: Section 2 discusses the related literature. Section 3 defines the task under investigation, describes the *MeSH-Rel-4K* dataset, and details the selected models. Section 4 outlines the experimental methodology and implementation. Section 5 presents the results and an error analysis. Finally, Section 6 concludes this study, outlining areas for future research.

2. Related Work

This section discusses the literature focusing on two key aspects relevant to our work: ontologies of research areas and state-of-the-art methodologies for generating these ontologies.

2.1. Ontologies of Research Areas

Ontologies of research areas are essential for organising and retrieving research outputs, such as publications and datasets, within digital libraries and repositories [19]. By delineating subfields and their semantic relationships, ontologies provide structured maps of academic domains. Notable examples include domain-specific ontologies, such as Medical Subject Headings (MeSH) [3], Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT) [20], and the Gene Ontology (GO) [21] within the biomedical domain. Other academic disciplines rely on resources such as the ACM Computing Classification System (CCS) [4], the Computer Science Ontology (CSO) [22], the IEEE Thesaurus[5], the Mathematical Subject Classification (MSC), and Physics Subject Headings (PhySH) [23]. In Biomedicine, MeSH is a comprehensive ontology comprising over 30K concepts, developed and maintained by the National Library of Medicine [3]. Widely adopted across the medical and health sciences, it is updated annually. On the other hand, SNOMED CT encompasses over 376K distinct medical concepts, functions as a highly granular clinical terminology for electronic health records to standardise clinical documentation and reporting [24]. Finally, the Gene Ontology includes more than 43K concepts to classify molecular functions, biological processes, and cellular components across various species and genomic databases [25]. In Computer Science, the ACM CCS[6] is a taxonomy of research topics within the field of computer science, encompassing approximately 2K research topics. Curated by the Association for Computing Machinery (ACM), its most recent update occurred in 2012. The IEEE Thesaurus primarily covers the field of engineering while incorporating various concepts relevant to computer science. It is developed by the Institute of Electrical and Electronics Engineers[7], contains 5.6K concepts and 24K semantic relationships, and is updated annually. The MSC is a taxonomy with more than 6.5K topics, encompassing a wide range of mathematical disciplines, from pure mathematics to statistics, jointly curated by the American Mathematical Society and zbMATH. It is updated decennially [2]. PhySH features approximately 3.7K concepts related to physics and is primarily employed to index literature in

[4]Our code and dataset - https://github.com/ImTanay/Biomedical-Ontology-Generation
[5]IEEE Thesaurus - https://www.ieee.org/content/dam/ieee-org/ieee/web/org/pubs/ieee-thesaurus.pdf
[6]The ACM CCS – http://www.acm.org/publications/class-2012
[7]IEEE Taxonomy - https://www.ieee.org/content/dam/ieee-org/ieee/web/org/pubs/ieee-taxonomy.pdf

Physical Review and on arXiv [23]. Maintained by the American Physical Society, it is updated annually, with the most recent release occurring in December 2025.

Furthermore, multi-disciplinary ontologies play a crucial role across broader academic landscapes. The UNESCO Thesaurus, curated by UNESCO, encompasses 4.4K subjects and is predominantly utilised for indexing and retrieving resources within UNESCO's document repository. It undergoes frequent minor revisions, typically every two to three months. Similarly, the ANZSRC Fields of Research (FoR) was developed by the Australian and New Zealand research councils. It covers 4.4K topics across diverse disciplines, though it is updated less frequently (~10 years). Digital Science employs this taxonomy to organise research content within its primary datasets, Dimensions.ai and Figshare. OpenAlex Topics is another multi-disciplinary taxonomy comprising 4.8K topics. It is primarily utilised within OpenAlex, a free and open catalogue of scholarly articles, and is currently in its initial release [26].

A systematic survey [1] of 45 widely adopted ontologies that describe research areas revealed that none of them is simultaneously openly available, sufficiently fine-grained, comprehensively describes all academic disciplines, and is frequently maintained. Part of this gap is due to the high costs of manual curation. Indeed, 82% of them are manually curated, demanding substantial time and financial resources. We argue that automated frameworks have emerged as a highly viable alternative to bypass these resource bottlenecks and ensure the sustainable evolution of scientific knowledge bases.

2.2. Automatic Generation of Research Area Ontologies

Early semi-automatic ontology generation progressed from combining expert input with statistical tools [27] to utilising machine learning frameworks (e.g., Text2Onto [28], OntoLearn [29]) to minimise manual effort. More recently, deep learning models such as BERT have significantly advanced concept extraction and hierarchical relationship modelling [30, 31, 32].

LLMs have further demonstrated strong potential for ontology generation [14]. However, persistent accuracy and consistency issues dictate that frameworks such as NeOn-GPT [33] and LLMs4Life [34] function best as assistive tools requiring context-rich prompting and human-in-the-loop verification [16, 35, 36]. Consequently, LLMs currently struggle to autonomously produce expert-level ontologies for specialised tasks like scientific publication classification [37].

Specifically within the domain of research topic ontologies, prior semi-automated and automated approaches have successfully employed subsumption logic, co-occurrence data, and citation clustering to construct large-scale frameworks. Notable examples include the CSO [22, 10], Microsoft's Field of Science [38], and taxonomies for OpenAlex [26] and an ANZSRC extension [39], alongside methods designed to enrich existing ontologies [40, 41].

Although LLMs are increasingly utilised for broader academic tasks, including paper discovery [42], citation prediction [43], scientific question answering [44], and literature review generation [45], their specific application in generating research topic ontologies remains underexplored. This study addresses this gap by systematically evaluating the capacity of recent LLMs to infer semantic relationships between research topics.

3. Background

This section outlines the task under investigation (Section 3.1), details the *MeSH-Rel-4K* dataset (Section 3.2), and provides an overview of LLMs utilised in this study (Section 3.3).

3.1. Task Definition

The task is defined as identifying the semantic relationship between a pair of research topics (t_A, t_B). It is formulated as a single-label, multi-class classification problem, where each pair is assigned to one of the following four categories:

- `broader`: t_A is a parent topic of t_B. For example, *plants* is broader than *agricultural crops*.

- `narrower`: t_A is a child topic of t_B. For example, *pneumocystis infections* is contained within *fungal diseases*. This constitutes the inverse relationship to `broader`.
- `same-as`: t_A and t_B can be used interchangeably to represent the same research concept. For example, *taste dysfunction* and *taste disorders*.
- `other`: t_A and t_B are unrelated or when none of the above applies. In contrast to the previous three categories, this does not define a semantic relationship. It rather provides a mechanism for the classifier to label negative examples.

In alignment with our previous study [15], we considered only `broader`, `narrower`, and `same-as` as semantic relationships, since these are fundamental for constructing ontologies that effectively map academic disciplines.

3.2. The MeSH-Rel-4K Dataset

To evaluate the efficacy of fine-tuned LLMs in inferring semantic relationships between research topics compared to different prompting strategies, we sampled 4,000 semantic relationships from MeSH[8]. MeSH employs a custom schema[9] (`mesh:`), wherein each topic (`mesh:TopicalDescriptor`) is delineated by a set of properties. These topics are organised hierarchically via the `mesh:broaderDescriptor` property, while synonymous concepts are denoted using the `mesh:relatedConcept` property. Utilising the January 2025 release of MeSH, we randomly sampled 1,000 relationship pairs for both the `broader` and `narrower` categories, deriving these from the `mesh:broaderDescriptor` property. Furthermore, we extracted 1,000 `same-as` instances based on the `mesh:relatedConcept` property, alongside 1,000 pairs of semantically disjoint topics to populate the `other` category. The *MeSH-Rel-4K* dataset was subsequently partitioned into training (2,800 semantic relationships), validation (400), and test (800) subsets, adhering to a 7:1:2 ratio. Moreover, research concepts were strictly isolated within their respective sets to prevent node leakage. The complete dataset and the source code used for its construction are openly available in our GitHub repository[10].

3.3. Large Language Models

Building upon insights from our study in the engineering domain [15], we fine-tuned a selection of five high-performing, open-source LLMs: `mistral-7b` [46], `llama-3b` [47], `gemma-9b` [48], `phi-3` [49], and `zephyr-7b` [50]. All models are 4-bit quantised and are available on HuggingFace. Table 1 provides an overview of these models, detailing their full names, the abbreviated aliases adopted throughout this paper, parameters count, context window sizes, and the specific Low-Rank Adaptation (LoRA) [51] hyperparameters applied during fine-tuning: the rank of the update matrices (**r**[11]) and the LoRA scaling factor (**alpha**). Initial selections for **r** and **alpha** followed the recommendations established by [52]. We then conducted multiple training iterations to empirically adjust these hyperparameters, and Table 1 presents only the optimal configurations for each model.

4. Methodology

This section details the procedures employed for the prompting (Section 4.1) and fine-tuning (Section 4.2) experiments, alongside the technical specifications (Section 4.3).

4.1. Prompting Strategies

We utilised two distinct prompting techniques: standard and CoT prompting [17, 53]. In **standard prompting**, prompts are generated for each pair of research topics using a predefined template[12]. This

[8]Medical Subject Headings (MeSH) - https://id.nlm.nih.gov/mesh/
[9]MeSH Schema - http://id.nlm.nih.gov/mesh/vocab
[10]Our code - https://github.com/ImTanay/Biomedical-Ontology-Generation
[11]**r** - The rank of the update matrices (where lower values yield smaller update matrices with fewer trainable parameters)
[12]Our prompt template is available at: https://github.com/ImTanay/Biomedical-Ontology-Generation

Table 1
Overview of the five LLMs used in this study. The table includes the **Model** name, the **Alias** adopted in this paper, the trainable **Parameters** count, the size of context **Window**, the rank (**r**) and scaling factor (**alpha**) employed in LoRA.

Model	Alias	Parameters	Window	r	alpha
mistral-7b-instruct-v0.3	mistral-7b	7.25B	32K	16	16
Llama-3.2-3B-Instruct	llama-3b	3.21B	128K	256	128
gemma-2-9b-it	gemma-9b	9.24B	8K	256	128
Phi-3.5-mini-instruct	phi-3	3.82B	128K	256	128
zephyr-sft	zephyr-7b	7.24B	8K	16	16

template outlines the task, defines the four target relationships (as detailed in Section 3.1), and instructs the LLMs to produce outputs in a structured format to facilitate automated parsing.

The **CoT, two-way prompting** strategy builds upon our previous findings [15, 54], which demonstrate the efficacy of this method in classifying semantic relationships. This approach employs a two-stage sequential prompting mechanism. The initial prompt instructs the model to define both topics, construct a sentence incorporating them, and reflect upon their potential semantic relationship. The subsequent prompt, in a chain-of-thought fashion, concatenates this generated output with specific classification instructions. To ensure robustness, the two topics are then swapped, and the entire process is repeated. Finally, a predefined set of empirical rules is applied to resolve any discrepancies between the bidirectional outcomes. To ensure a fair and consistent comparison, identical prompt templates were administered across all evaluated models. Comprehensive details regarding the prompt templates and the empirical referee rules are described in [15].

4.2. Fine-Tuning

We fine-tuned five open-source LLMs on the *MeSH-Rel-4K*, employing identical training and validation sets (as detailed in Section 3.2) to ensure comparability across results and to maintain a standardised, prompt-based fine-tuning methodology. The base prompt template was structured as follows:

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

Within this template, `'[TOPIC-A]'` and `'[TOPIC-B]'` serve as placeholders, which were dynamically replaced with specific pairs of research topics during the fine-tuning process. For instance:

```
user: Classify the relationship between 'acinic cells' and 'cells'
model: relationship: 'narrower'
```

During both training and validation, the expected output was explicitly formulated as `relationship: [RELATIONSHIP-TYPE]`. This formatting conditioned the models to generate structured, easily parsable responses, thereby facilitating the automated extraction of the predicted relationships during evaluation.

4.3. Experimental Setup

To fine-tune and interact with the LLMs detailed in Table 1, we utilised two open-source libraries: KoboldAI[13] and Unsloth [55].

KoboldAI was employed to interact with the LLMs within a zero-shot prompting configuration. Built upon llama.cpp[14], this platform facilitates API-based access to locally hosted models.

Unsloth was utilised for the fine-tuning process. This library is constructed upon the HuggingFace Transformers [56], and Parameter-Efficient Fine-Tuning (PEFT) [18] frameworks. It supports 4-bit quantised training via BitsAndBytes[15], substantially reducing memory consumption and permitting

[13]KoboldAI - https://github.com/KoboldAI/KoboldAI-Client
[14]llama.cpp - https://github.com/ggml-org/llama.cpp
[15]BitsAndBytes - https://github.com/bitsandbytes-foundation/bitsandbytes

training on consumer-grade GPUs. Furthermore, Unsloth integrates LoRA [51], enabling PEFT by injecting lightweight, trainable adapter layers into the frozen pre-trained model architecture.

All experiments were executed within Google Colaboratory environments (using NVIDIA A100 and L4 GPUs). To facilitate reproducibility, the complete codebase is openly accessible in our GitHub repository[16]. This repository contains all necessary scripts for both the prompting and fine-tuning phases, alongside the precise configuration parameters applied within Unsloth and KoboldAI.

5. Results and Discussion

This section presents the results of our experiments, beginning with the prompting strategies (Section 5.1), followed by the outcomes of the fine-tuning (Section 5.2), and concluding with a comprehensive analysis of all LLMs fine-tuned on the *MeSH-Rel-4K* (Section 5.3).

5.1. Prompting Strategies

We evaluated the LLMs using precision, recall, and F1-score. Table 2 reports the results for standard prompting and Table 3 for the CoT, two-way approach.

When comparing standard prompting against the CoT, two-way approach. The CoT, two-way approach shows consistently better performance, yielding an average F1-score increase of 5.4 percentage points (standard deviation ±3.5). `gemma-9b` recorded the highest absolute metrics across both configurations (F1: 71.6% with CoT, two-way and F1: 66.9% with standard prompting). The largest relative gains occurred in `mistral-7b` (F1 improved by 10.0 percentage points) and `zephyr-7b` (F1 improved by 7.1 percentage points). However, the lowest-parameter model (`llama-3b`) registered a minimal 0.5 percentage point increase. At the category level, metrics indicate that under standard prompting, recall for `same-as` and `narrower` falls strictly below that of `broader` and `other`. The CoT, two-way approach resolves these specific boundary deficits. For instance, `gemma-9b` recorded the highest overall recall (72.5%) and precision (97.7%) for `same-as`. `mistral-7b` also registered precision increases for both `broader` (76.6%) and `narrower` (66.5%). These data confirm that structured reasoning improves the performance of LLMs, particularly for smaller and cost-effective models.

Table 2
Precision, Recall, and F1-score for standard prompting on the test set (800 semantic relationships) of *MeSH-Rel-4K*. BR refers to performance on `broader` relations, NA to `narrower`, OT to `other`, and SA to `same-as`. AVG indicates the average performance across the four categories. The best-performing scores for each metric are highlighted in **bold & underlined**. Due to space constraints, the leading zero has been omitted from all values.

MODEL	F1-SCORE					PRECISION					RECALL				
	AVG	BR	NR	OT	SA	AVG	BR	NR	OT	SA	AVG	BR	NR	OT	SA
mistral-7b	.603	**.706**	.568	.686	.452	.684	**.692**	.641	.532	.871	.625	.720	.510	**.965**	.305
llama-3b	.252	.398	.137	.417	.057	.413	.265	.244	.598	.545	.311	.800	.095	.320	.030
gemma-9b	**.669**	.705	**.722**	**.811**	.438	**.766**	.571	**.705**	.788	**1.000**	**.694**	**.920**	**.740**	.835	.280
phi-3	.529	.578	.559	.718	.263	.651	.480	.468	**.794**	.861	.557	.725	.695	.655	.155
zephyr-7b	.415	.452	.245	.491	**.472**	.476	.538	.442	.359	.566	.436	.390	.170	.780	**.405**
AVG	.494	.568	.446	**.625**	.336	.598	.509	.500	.614	**.769**	.525	**.711**	.442	**.711**	.235

5.2. Fine-Tuning

Fine-tuning on the *MeSH-Rel-4K* substantially improved performance across all models compared to prompting strategies, as reported in Table 4. When evaluated against the CoT, two-way approach, fine-tuning yielded a significant average F1-score improvement of 34.1 percentage points (61.6% relative performance increase). While `gemma-9b` retained overall dominance (F1: 91.6%), the smallest model

[16]Our code - https://github.com/ImTanay/Biomedical-Ontology-Generation

Table 3
Precision, Recall, and F1-score for CoT, two-way on Testset (800 semantic relationships) of *MeSH-Rel-4K*. BR refers to performance on `broader` relations, NA to `narrower`, OT to `other`, and SA to `same-as`. AVG indicates the average performance across the four categories. The best-performing scores for each relation are highlighted in **bold & underlined**. Due to space constraints, the leading zero has been omitted from all values.

MODEL	F1-SCORE					PRECISION					RECALL				
	AVG	BR	NR	OT	SA	AVG	BR	NR	OT	SA	AVG	BR	NR	OT	SA
mistral-7b	.703	.761	.680	.785	.587	.732	**.766**	**.665**	.669	.827	.714	.755	.695	**.950**	.455
llama-3b	.257	.395	.319	.204	.109	.459	.300	.249	.686	.600	.301	.580	.445	.120	.060
gemma-9b	**.716**	**.763**	**.715**	**.787**	**.597**	**.775**	.662	.633	.829	**.977**	**.725**	**.900**	**.820**	.750	.430
phi-3	.576	.627	.577	.675	.423	.702	.518	.473	**.900**	.917	.588	.795	.740	.540	.275
zephyr-7b	.486	.448	.407	.530	.557	.516	.607	.496	.429	.532	.495	.355	.345	.695	**.585**
AVG	.548	**.599**	.540	.596	.455	.637	.571	.503	.703	**.771**	.565	**.677**	.609	.611	.361

(`llama-3b`) demonstrated an improvement of 60.5 percentage point (from F1: 25.7% with CoT, two-way to F1: 86.2% after fine-tuning). Nevertheless, other models exhibit distinct, metric-specific strengths. For instance, `mistral-7b` achieves the highest precision for the `other` category (96.8%), whilst `zephyr-7b` attains the highest recall for `narrower` (88.0%), and `llama-3b` for `broader` (95.0%).

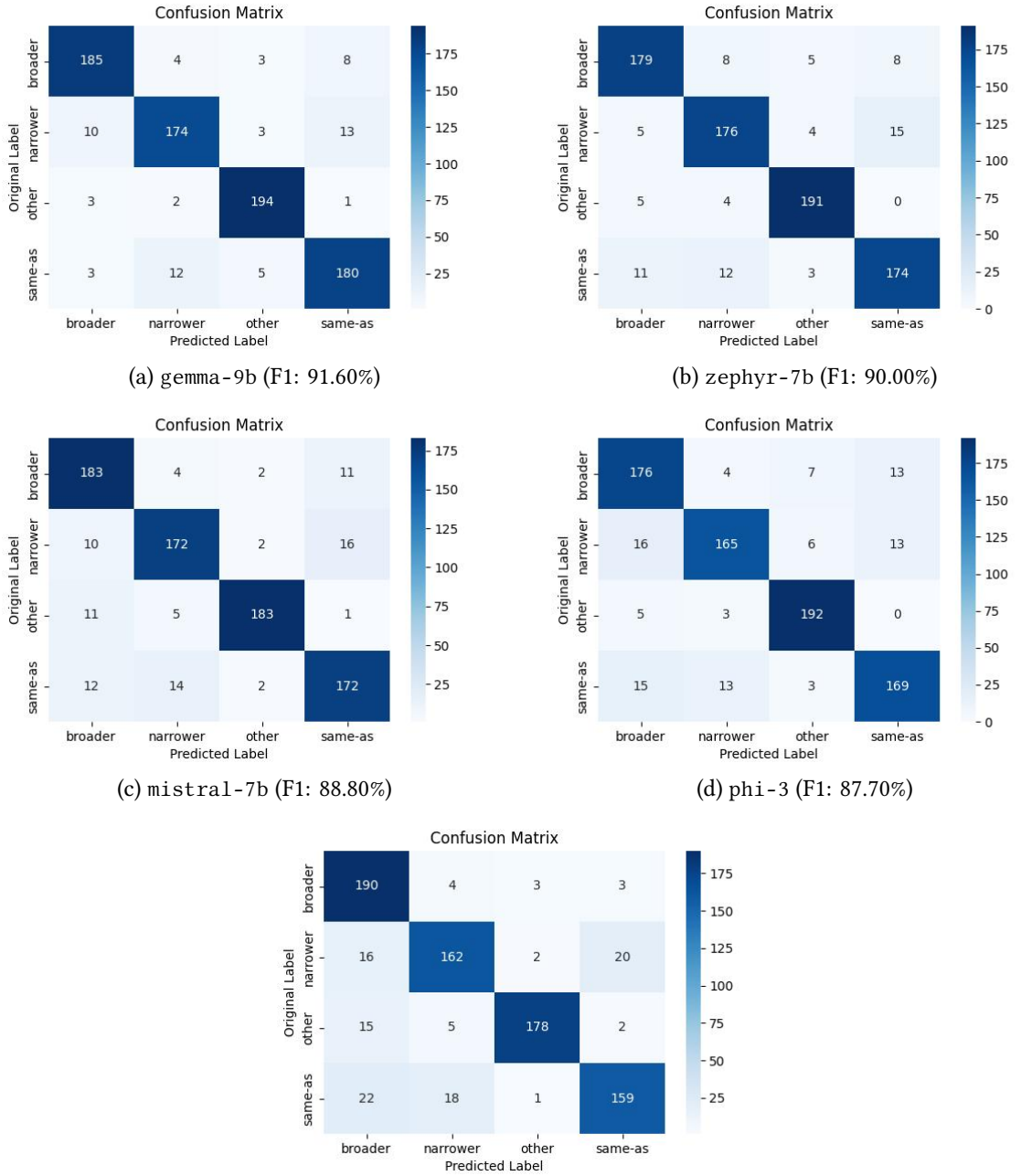
Across all evaluated models, the four categories achieve comparable average F1-scores ranging from 86.6% to 94.3%. The `other` category consistently yields the highest average scores across all metrics (F1: 94.3%, precision: 94.9%, and recall: 93.8%). These exceptionally high results demonstrate that fine-tuned LLMs efficiently identify and isolate unrelated topics. This is important for ontology generation, as incorrectly hallucinated relations can compromise the structural integrity of an ontology by introducing loops [10]. Conversely, `same-as` proves to be the most challenging relation, recording the lowest average F1-score (86.3%) and precision (87.3%) across the cohort. Followed by `narrower` exhibiting the lowest average recall (84.9%). These findings confirm that fine-tuning effectively bridges the reasoning deficits of smaller models, yielding near-state-of-the-art models.

Table 4
Precision, Recall, and F1-score for models fine-tuned and evaluated on the *MeSH-Rel-4K*. BR refers to performance on `broader` relations, NR to `narrower`, OT to `other`, and SA to `same-as`. AVG indicates the average performance across the four categories. The best-performing scores for each relation are highlighted in **bold & underlined**. Due to space constraints, the leading zero has been omitted from all values.

MODEL	F1-SCORE					PRECISION					RECALL				
	AVG	BR	NR	OT	SA	AVG	BR	NR	OT	SA	AVG	BR	NR	OT	SA
mistral-7b	.888	.880	.871	.941	.860	.889	.847	.882	**.968**	.860	.887	.915	.860	.915	.860
llama-3b	.862	.858	.833	.927	.828	.868	.782	.857	.967	.864	.861	**.950**	.810	.890	.795
gemma-9b	**.916**	**.923**	**.888**	**.958**	**.895**	**.916**	**.920**	**.906**	.946	**.891**	**.916**	.925	.870	**.970**	**.900**
phi-3	.877	.854	.857	.941	.856	.878	.830	.892	.923	.867	.877	.880	.825	.960	.845
zephyr-7b	.900	.895	.880	.948	.877	.900	.895	.880	.941	.883	.900	.895	**.880**	.955	.870
AVG	.889	.882	.866	**.943**	.863	.890	.855	.883	**.949**	.873	.888	.913	.849	**.938**	.854

5.3. Analysis of LLMs on MeSH-Rel-4K

The confusion matrices, presented in Fig. 1, indicate that all fine-tuned models achieve high classification accuracy, with F1-scores ranging from 86.2% to 91.6%. The `other` category consistently yields the highest true positive rates across the cohort (194 out of 200 instances for `gemma-9b`, see Fig. 1a). This demonstrates that differentiating unrelated topics presents minimal difficulty for fine-tuned LLMs. The primary classification bottleneck across all models is the disambiguation of equivalence (`same-as`) from hierarchical relations (`broader`/`narrower`). The frequency of these boundary errors scales inversely with model parameter count. While `gemma-9b` minimises this confusion, the smallest model (`llama-3b`, see Fig. 1e) frequently predicts false hierarchical relations, misclassifying 40 `same-as` instances as



(a) `gemma-9b` (F1: 91.60%)

(b) `zephyr-7b` (F1: 90.00%)

(c) `mistral-7b` (F1: 88.80%)

(d) `phi-3` (F1: 87.70%)

(e) `llama-3b` (F1: 86.20%)

Figure 1: Confusion matrices for LLMs fine-tuned and evaluated on *MeSH-Rel-4K*

`broader` (22) and `narrower` (18). The mid-sized models (`zephyr-7b`, `mistral-7b`, and `phi-3`) exhibit symmetric confusion, misclassifying hierarchical and equivalence relationships at comparable rates. For instance, `zephyr-7b` (see Fig. 1b) misclassifies 23 hierarchical instances (8 `broader`, 15 `narrower`) as `same-as`, whilst predicting exactly 23 `same-as` instances as `broader` (11) and `narrower` (12). This bidirectional error distribution is mirrored by `mistral-7b` (see Fig. 1c), which predicts 27 hierarchical topics (11 `broader`, 16 `narrower`) as `same-as` against 26 `same-as` instances classified as `broader` (12) and `narrower` (14). `phi-3` (see Fig. 1d) replicates this pattern, misattributing 26 hierarchical instances (13 `broader`, 13 `narrower`) to `same-as`, alongside 28 `same-as` instances to hierarchical categories (15 `broader`, 13 `narrower`).

Analysis of these misclassifications indicates that this error pattern correlates with dense lexical overlap within the MeSH. For example, “alitretinoin” (9-cis-retinoic acid) is formally defined as a

narrower topic relative to the foundational compound "tretinoin". However, all five fine-tuned models unanimously classified this relationship as same-as.

6. Conclusion and Future Work

This paper evaluated the performance of five small, open-source LLMs in identifying semantic relationships between research topics within the biomedical domain. To support this analysis, we introduced *MeSH-Rel-4K* (comprising 4K semantic relationships) extracted from the MeSH. Our experiments demonstrated that while advanced prompting strategies (such as CoT, two-way [15]) yielded moderate baseline improvements, explicit fine-tuning on *MeSH-Rel-4K* significantly increased classification accuracy. The fine-tuned gemma-9b achieved the highest overall performance (F1: 91.6%). Furthermore, llama-3b exhibited a 60.5 percentage point increase in its F1-score following fine-tuning.

Our future work will focus on broadening our analysis across various academic disciplines. Beginning with STEM fields, we will expand into the Humanities, Social Sciences, and Economics. Additionally, we plan to develop a comprehensive pipeline for ontology matching and evolution. To identify semantic links between existing ontologies, this pipeline will use an LLM-based classifier. It will extract relevant metadata[17] from scholarly articles and identify emerging research topics. These topics will be systematically integrated into a self-evolving, multidisciplinary ontology of research topics.

Declaration on Generative AI

During the preparation of this work, the author(s) used Grammarly, and Gemini to check grammar and spelling. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the content of the publication.

References

[1] A. Salatino, T. Aggarwal, A. Mannocci, F. Osborne, E. Motta, A survey of knowledge organization systems of research fields: Resources and challenges, Quantitative Science Studies 6 (2025) 567–610.

[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] E. Palumbo, D. Monti, G. Rizzo, R. Troncy, E. Baralis, entity2rec: Property-specific knowledge graph embeddings for item recommendation, Expert Systems with Applications 151 (2020) 113235.

[6] A. Cadeddu, A. Chessa, V. De Leo, G. Fenu, E. Motta, F. Osborne, D. R. Recupero, A. Salatino, L. Secchi, A comparative analysis of knowledge injection strategies for large language models in the scholarly domain, Engineering Applications of Artificial Intelligence 133 (2024) 108166.

[7] 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.

[8] 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.

[9] S. Angioni, A. Salatino, F. Osborne, D. R. Recupero, E. Motta, Aida: A knowledge graph about research dynamics in academia and industry, Quantitative Science Studies 2 (2021) 1356–1398.

[17] Metadata: Title, Abstract, Keywords, Introduction, and Conclusion

[10] F. Osborne, E. Motta, Klink-2: integrating multiple web sources to generate semantic topic networks, in: The Semantic Web-ISWC 2015: 14th International Semantic Web Conference, Bethlehem, PA, USA, October 11-15, 2015, Proceedings, Part I 14, Springer, 2015, pp. 408–424.
[11] K. Han, P. Yang, S. Mishra, J. Diesner, Wikicssh: extracting computer science subject headings from wikipedia, in: ADBIS, TPDL and EDA 2020 Common Workshops and Doctoral Consortium: International Workshops: DOING, MADEISD, SKG, BBIGAP, SIMPDA, AIMinScience 2020 and Doctoral Consortium, Lyon, France, August 25–27, 2020, Proceedings 24, Springer, 2020, pp. 207–218.
[12] F. Osborne, E. Motta, Mining semantic relations between research areas, in: The Semantic Web–ISWC 2012: 11th International Semantic Web Conference, Boston, MA, USA, November 11-15, 2012, Proceedings, Part I 11, Springer, 2012, pp. 410–426.
[13] M. Sanderson, B. Croft, Deriving concept hierarchies from text, in: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, 1999, pp. 206–213.
[14] H. Babaei Giglou, J. D'Souza, S. Auer, Llms4ol: Large language models for ontology learning, in: International Semantic Web Conference, Springer, 2023, pp. 408–427.
[15] T. Aggarwal, A. Salatino, F. Osborne, E. Motta, Large language models for scholarly ontology generation: An extensive analysis in the engineering field, Information Processing & Management 63 (2026) 104262.
[16] A. S. Lippolis, M. J. Saeedizade, R. Keskisärkkä, S. Zuppiroli, M. Ceriani, A. Gangemi, E. Blomqvist, A. G. Nuzzolese, Ontology generation using large language models, arXiv preprint arXiv:2503.05388 (2025).
[17] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, 2023. URL: https://arxiv.org/abs/2201.11903. arXiv:2201.11903.
[18] S. Mangrulkar, S. Gugger, L. Debut, Y. Belkada, S. Paul, B. Bossan, Peft: State-of-the-art parameter-efficient fine-tuning methods, https://github.com/huggingface/peft, 2022.
[19] M. L. Zeng, Knowledge organization systems (kos), KO Knowledge Organization 35 (2008) 160–182.
[20] D. Lee, N. de Keizer, F. Lau, R. Cornet, Literature review of snomed ct use, Journal of the American Medical Informatics Association 21 (2014) e11–e19.
[21] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, et al., Gene ontology: tool for the unification of biology, Nature genetics 25 (2000) 25–29.
[22] 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.
[23] A. Smith, Physics subject headings (physh), KO KNOWLEDGE ORGANIZATION 47 (2020) 257–266.
[24] E. Chang, J. Mostafa, The use of snomed ct, 2013-2020: a literature review, Journal of the American Medical Informatics Association 28 (2021) 2017–2026.
[25] G. O. Consortium, The gene ontology resource: 20 years and still going strong, Nucleic acids research 47 (2019) D330–D338.
[26] OpenAlex, Openalex: End-to-end process for topic classification, 2024.
[27] A. Maedche, S. Staab, Learning ontologies for the semantic web., in: SemWeb, 2001.
[28] P. Cimiano, J. Völker, Text2onto: A framework for ontology learning and data-driven change discovery, in: International conference on application of natural language to information systems, Springer, 2005, pp. 227–238.
[29] P. Velardi, S. Faralli, R. Navigli, Ontolearn reloaded: A graph-based algorithm for taxonomy induction, Computational Linguistics 39 (2013) 665–707.
[30] C. Chen, K. Lin, D. Klein, Constructing taxonomies from pretrained language models, arXiv preprint arXiv:2010.12813 (2020).

[31] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.
[32] M. Grootendorst, Bertopic: Neural topic modeling with a class-based tf-idf procedure, arXiv preprint arXiv:2203.05794 (2022).
[33] N. Fathallah, A. Das, S. D. Giorgis, A. Poltronieri, P. Haase, L. Kovriguina, Neon-gpt: a large language model-powered pipeline for ontology learning, in: European Semantic Web Conference, Springer, 2024, pp. 36–50.
[34] N. Fathallah, S. Staab, A. Algergawy, Llms4life: Large language models for ontology learning in life sciences, 2024. URL: https://arxiv.org/abs/2412.02035. arXiv:2412.02035.
[35] M. J. Saeedizade, E. Blomqvist, Navigating ontology development with large language models, in: European Semantic Web Conference, Springer, 2024, pp. 143–161.
[36] S. Tsaneva, D. Dessì, F. Osborne, M. Sabou, Knowledge graph validation by integrating llms and human-in-the-loop, Information Processing & Management 62 (2025) 104145.
[37] Y. Sun, H. Xin, K. Sun, Y. E. Xu, X. Yang, X. L. Dong, N. Tang, L. Chen, Are large language models a good replacement of taxonomies?, arXiv preprint arXiv:2406.11131 (2024).
[38] K. Wang, Z. Shen, C. Huang, C.-H. Wu, Y. Dong, A. Kanakia, Microsoft academic graph: When experts are not enough, Quantitative Science Studies 1 (2020) 396–413.
[39] G. B. Jenset, P. J. Bevan, A. Jain, A large-scale, granular topic classification system for scientific documents (2025).
[40] K. I. Kotis, G. A. Vouros, D. Spiliotopoulos, Ontology engineering methodologies for the evolution of living and reused ontologies: status, trends, findings and recommendations, The Knowledge Engineering Review 35 (2020) e4.
[41] F. Osborne, E. Motta, Pragmatic ontology evolution: reconciling user requirements and application performance, in: International Semantic Web Conference, Springer, 2018, pp. 495–512.
[42] S. Chow, L. Guo, J. Chow, C. Chia, S. Li, D.-Y. Huang, Semantic search using llm-aided topic generation on knowledge graphs for paper discovery, in: 2024 IEEE 14th International Symposium on Chinese Spoken Language Processing (ISCSLP), IEEE, 2024, pp. 353–357.
[43] D. Buscaldi, D. Dessí, E. Motta, M. Murgia, F. Osborne, D. R. Recupero, Citation prediction by leveraging transformers and natural language processing heuristics, Information Processing & Management 61 (2024) 103583.
[44] S. Auer, D. A. Barone, C. Bartz, E. G. Cortes, M. Y. Jaradeh, O. Karras, M. Koubarakis, D. Mouromtsev, D. Pliukhin, D. Radyush, et al., The sciqa scientific question answering benchmark for scholarly knowledge, Scientific Reports 13 (2023) 7240.
[45] F. Bolanos, A. Salatino, F. Osborne, E. Motta, Artificial intelligence for literature reviews: Opportunities and challenges, Artificial Intelligence Review 57 (2024) 259.
[46] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al., Mistral 7b, arXiv preprint arXiv:2310.06825 (2023).
[47] AI@Meta, Llama 3 model card (2024). URL: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.
[48] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love, et al., Gemma: Open models based on gemini research and technology, arXiv preprint arXiv:2403.08295 (2024).
[49] M. Abdin, J. Aneja, H. Awadalla, A. Awadallah, A. A. Awan, N. Bach, A. Bahree, A. Bakhtiari, J. Bao, H. Behl, et al., Phi-3 technical report: A highly capable language model locally on your phone, 2024, URL https://arxiv. org/abs/2404.14219 (2024).
[50] D. Han, Unsloth/zephyr-sft · hugging face, 2024. URL: https://huggingface.co/unsloth/zephyr-sft.
[51] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models. arxiv 2021, arXiv preprint arXiv:2106.09685 (2021).
[52] LoRA fine-tuning Hyperparameters Guide | Unsloth Documentation — unsloth.ai, https://unsloth.ai/docs/get-started/fine-tuning-llms-guide/lora-hyperparameters-guide#lora-alpha-and-rank-relationship, 2025.
[53] 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.

[54] T. Aggarwal, A. Salatino, F. Osborne, E. Motta, et al., Identifying semantic relationships between research topics using large language models in a zero-shot learning setting, in: CEUR Workshop Proceedings, volume 3780, CEUR-WS, 2024.

[55] M. H. Daniel Han, U. team, Unsloth, 2023. URL: https://github.com/unslothai/unsloth.

[56] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: https://www.aclweb.org/anthology/2020.emnlp-demos.6.