Biochemical Knowledge Extraction through Similarity Search via Finetuned Open-Source Large Language Models

Paulo do Carmo[1], Marcos Gôlo[2], Jonas Gwozdz[3], Edgard Marx[1], Ricardo Marcacini[2]

[1]Institute for Applied Informatics (InfAI)– Leipzig – Germany

[2]University of São Paulo (ICMC/USP) – São Carlos – São Paulo – Brazil

[3]Faculty of Informatics, Leipzig University of Applied Sciences – Leipzig – Germany

{carmo,marx}@infai.org, jonas.gwozdz@htwk-leipzig.de

Abstract. *Natural products have played a central role in drug development, and organizing all these compounds remains a time-consuming, often manual task, motivating efforts to automatically extract biochemical knowledge. Previous approaches based on knowledge graph embeddings and the auxiliary use of Large Language Models (LLMs) have shown promising results. However, these methods still have limitations in distinguishing new from cited properties within the text and rely on proprietary models and constrained prompting strategies. Thus, we propose combining finetuning of open-source LLMs with similarity search. Finetuning adapts the models to the biochemical domain, while similarity search serves as a retrieval layer that maps generated outputs to biochemical candidates. This combination provides an advantage, since newly emerging compounds can be incorporated into the candidate space without requiring new fine-tuning. Results on the NatUKE benchmark show that the proposed strategy achieves state-of-the-art performance compared to open-source models.*

1. Introduction

The search for new drugs remains one of the most important and challenging goals in biochemical research [Atanasov et al. 2021]. In this context, natural products have played a central role in drug development due to their structural diversity and biological relevance, making these compounds promising candidates in drug discovery pipelines [Zhang et al. 2020]. Between 1981 and 2019, approximately 23.4% of all approved drugs were derived from or inspired by natural products [Newman and Cragg 2020], highlighting their continued importance in modern pharmacology. However, identifying and characterizing these compounds into structured databases remains a complex and time-consuming process, largely carried out manually by domain experts [Réda et al. 2020]. This challenge can be formulated as a biochemical knowledge-extraction task in which the input is a scientific article and the output is a set of biochemical properties. Figure 1 illustrates the task of biochemical knowledge extraction.



Figure 1. Illustration of the biochemical knowledge extraction task, where the properties are mapped into five target properties.

Within this perspective, the NatUKE benchmark was introduced to evaluate the automatic extraction of biochemical knowledge from scientific literature through Knowledge Graphs (KGs) [Do Carmo et al. 2023]. Although this structured formulation is an important step toward reproducible evaluation, the task still involves a substantial challenge. In chemistry papers, authors often mention compounds, species, and bioactivities not only to report new findings, but also to discuss prior work, compare extraction procedures, or contextualize known results. As a consequence, identifying whether a passage refers to a newly reported compound-property association or cites a past example from the literature cannot be resolved by structural relations alone. This reveals an important gap: while KGs are effective for organizing entities and candidate relations, using their contextual data alone is insufficient to capture the contextual and discourse-level semantics required for accurate extraction from academic papers.

To address this limitation, Large Language Models (LLMs) have emerged as a promising alternative due to their ability to extract information attached to a context [Zhai 2024, D'Souza et al. 2025]. However, within biochemical knowledge extraction and, more specifically, the NatUKE benchmark, they have mostly been used as auxiliary components. For example, to improve PDF preprocessing, enrich extracted text, or generate embeddings [Fröhlich et al. 2023, Do Carmo et al. 2025]. Although these studies indicate the potential of LLMs, their gains remain uneven across properties, especially for challenging ones. Studies that explore LLMs within NatUKE to extract properties rely on proprietary paid models and explicitly provide candidate biochemical options in the prompt, which limits scalability and practical applicability [Gôlo et al. 2025]. In other domains, finetuning has been successfully applied to enhance extraction performance and reduce hallucinations [Bitelli and Finger 2025, Oliveira et al. 2025].

Therefore, we propose a biochemical knowledge extraction strategy that leverages open-source LLMs, along with finetuning and similarity search, to align vocabularies. The central idea is to place the LLM in a setting that better matches its strengths, namely, text generation and contextual interpretation, while using similarity search as a retrieval layer. Finetuning allows the model to better adapt to the terminology, style, and extraction patterns of biochemical papers. Similarity search performs a mapping stage between the LLM output and the structured entities, improving robustness to textual variation. Additionally, it provides an important practical advantage: when new compounds or candidate values appear, the retrieval layer can incorporate them simply by indexing them, a much cheaper step than finetuning or redesigning the entire prompt. In this sense, our work moves beyond the current state-of-the-art techniques by using finetuned open-source LLMs, combined with similarity search, as the core mechanism for biochemical knowledge extraction on the NatUKE benchmark. In summary, our contributions are:

1. A new biochemical knowledge extraction strategy that combines LLM finetuning with similarity search, and publicly available finetuned models;
2. We present a similarity search, improving robustness against textual variation. Its indexing can accommodate newly emerging compounds immediately and with lower costs than a new finetuning step until enough new resources are collected;
3. Through finetuning experiments, we show that performance improves when we provide more context from the NatUKE benchmark to LLMs, and that this likely is still not fully incorporated into LLM pre-training.

2. Related Work

Non-LLMs methods. Knowledge Graph Embedding (KGE) methods advanced biochemical knowledge extraction by learning representations that explicitly model entities and relations in structured biochemical knowledge bases. This strategy was explored in the NatUKE benchmark, where papers and biochemical properties are represented in a Knowledge Graph (KG), and the extraction task is formulated as link prediction between article nodes and biochemical property nodes [Do Carmo et al. 2023]. The pioneering method, EPHEN, outperformed other KGE baselines, DeepWalk, Node2Vec, and MethaPath2Vec [Do Carmo and Marcacini 2021]. Subsequent works have explored EPHEN with improvements, including BFS-based walk generation for KGE [Zope et al. 2023], named entity recognition as a preprocessing step to improve article representations [Dichte and Mócsy 2023], and a knowledge integration strategy using K-BERT [Chand and Tiwari 2025]. Although these strategies improved specific properties, such as bioactivity, collection sites, and isolation-type extractions, their gains were uneven across tasks, indicating that structured graph representations alone are still insufficient to capture the full semantic richness of scientific papers, thereby motivating the exploration of LLMs as auxiliary components.

LLMs as Auxiliary Component. LLMs are designed to capture contextual and semantic information directly from text on a much broader scale; therefore, they quickly became targets of investigation. In the NatUKE setting, LLMs have already been explored mainly as auxiliary components to improve PDF text extraction or to generate better textual representations before applying KGE methods [Do Carmo et al. 2025, Fröhlich et al. 2023]. For instance, [Fröhlich et al. 2023] used GPT-3.5-turbo to enrich preprocessing and improve the state-of-the-art method of the benchmark (EPHEN), while [Do Carmo et al. 2025] showed that stronger PDF extractors, such as Nougat [Blecher et al. 2023] can substantially affect downstream performance and that BERT embeddings outperform LLM embeddings for EPHEN method. Even so, important gaps remain, particularly for challenging properties such as compound names, species names, and collection sites, as well as in the direct use of LLMs as the main mechanism for detecting biochemical properties in full academic papers. This gap motivates exploring LLMs as core components of the NatUKE benchmark.

LLMs as Core Component. Recent studies have also pointed to the potential of LLMs for direct information extraction [Zhai 2024, D'Souza et al. 2025, Bitelli and Finger 2025]. In this context, a recent study explored various strategies within the Benchmark NatUKE to fully leverage LLMs [Gôlo et al. 2025]. The authors explored the use of LLMs in the NatUKE benchmark and reported promising results for extracting biochemical properties. The study explores zero-shot, few-shot, and finetuning of open-source LLMs. However, its strongest performance was achieved only with proprietary paid models (GPT and Gemini), which limit reproducibility and practical adoption in open and cost-sensitive settings. In addition, the prompts used across all LLM strategies explicitly included all candidate biochemical options from the NatUKE Benchmark (e.g., a classification task), thereby turning the task into a constrained selection problem and reducing its scalability and flexibility when handling new compounds or unseen candidate values. To circumvent paid models and non-scalable prompt engineering, we propose combining finetuning of open-source LLMs with similarity search.

3. Finetuned Large Language Models with Similarity Search for Biochemical Knowledge Extraction

We present a pipeline for finetuning and similarity search (Figure 2) based on open-source LLMs, in which finetuning adapts the model to specific tasks not in focus during pre-training, such as biochemical knowledge extraction. Meanwhile, the similarity search step is responsible for: (i) aligning vocabulary output from the LLM into our structured knowledge base; and (ii) allowing a grounded prediction through an indexed vocabulary updated according to a common embedding model separate from the finetuned LLM, so that new vocabulary updates can be made with little performance overhead.

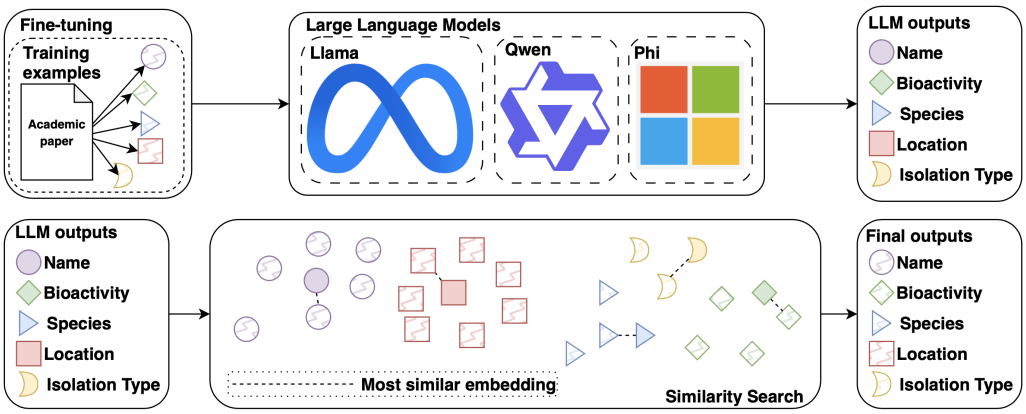


Figure 2. Similarity search via finetuned open-source LLMs for biochemical Knowledge Extraction. Each paper is linked to up to five biochemical properties. The LLMs are finetuned per task, and each task has a different collection of indexed options for similarity search.

Finetuning is a well-established technique for enhancing performance in specific tasks using pre-trained LLMs. We developed our finetuning step to better guide the LLM toward specific biochemical knowledge-extraction tasks. All the training data comes from the NatUKE benchmark [Do Carmo et al. 2023] splits. This benchmark was initially designed to evaluate graph embedding-based systems, and all splits were randomly selected for each extraction task. Since each task has its own split, we cannot guarantee that only the training data is used for finetuning unless we use a different model for each task. Therefore, we chose to finetune using single-task models [Gozzi and Di Maio 2024]. The primary advantage is that, by focusing on a single task, it can better adjust the weights, thereby improving performance in consistently extracting a single property.

In single-task models, each finetuning step must be executed separately and requires specific task prompts. Figure 3 presents the system, user, and assistant prompt templates for finetuning. The academic paper text becomes the user prompt, and the known extractions from the training set are provided as a list in the assistant prompt. Within the system prompt template, the *[task explanation]* variable is substituted with one of the sentences for each task, as follows:

- **Compound name**: For each paper, you have to analyze the content (text) to identify the Compound name. It can be more than one compound name.
- **Biological activity**: For each paper, you have to analyze the content (text) to identify the Biological Activity. It can be more than one biological activity.

- **Collection species**: For each paper, you have to analyze the content (text) to identify the Collection Species, i.e., the Species from which natural products were extracted. Provide the scientific name, binomial form. The family name can be provided. For example, Tithonia diversifolia, Styrax camporum (Styracaceae), or Colletotrichum gloeosporioides (Phyllachoraceae).
- **Collection sites**: For each paper, you have to analyze the content (text) to identify the collection Site, i.e., the place of the collection.
- **Isolation type**: For each paper, you have to analyze the content (text) to identify the Collection Type, i.e., Collection type of the species.

System prompt:
You are a scientist trained in chemistry. You must extract information from scientific papers identifying relevant properties associated with each natural product discussed in the academic publication. For each paper, *[task explanation]*. Your output should be a list with the names. Return only the list, without any additional information.

User:
[paper text extracted with Nougat]

Assistant:
[Example extracted 1, Example extracted 2]

Figure 3. Prompt template for finetuning. The system prompt provides basic directions to the LLM for biochemical knowledge extraction within a task. The user prompt contains the text of the academic paper. And the assistant prompt contains manually extracted training data for that paper.

We use the *unsloth* library[1] for all finetuning tasks. It is an open-source framework focused on memory- and time-efficient finetuning. Unsloth requires a message stream containing system data, created by completing each message using the data for a specific paper from the previously presented template. We then apply the Quantized Low-Rank Adaptation (QLoRA) [Dettmers et al. 2023] finetuning strategy, which is an even more optimized and memory-efficient version of LoRA [Hu et al. 2022]. LoRA allows finetuning without changing the original weights of the LLM, and QLoRA quantizes these weights to 4 bits instead of 16, further reducing memory requirements.

The Similarity Search module is responsible for two critical requirements from our proposal: (i) vocabulary alignment from the LLM into structured data; and (ii) it grounds the prediction with an indexed vocabulary in a common embedding model to allow new predictions to be made without updating the finetuned LLM. The available vocabulary for prediction is indexed in a common embedding, enabling nearest-neighbor cosine-similarity search over any LLM output, as required in (i). Using similarity search within an indexed vocabulary may be seen strictly as a limitation. However, we argue that both pre-trained LLMs and finetuning require training data on known entities. At the same time, indexing the new vocabulary is much more efficient than pre-training or finetuning. Therefore, embedding generation within a common space can also be executed dynamically for only new entities that are then immediately available for similarity search, as required in (ii). Using these steps when necessary aims to make our pipeline efficient in terms of time and performance in different scenarios.

[1] `https://docs.unsloth.ai`

4. Experimental Evaluation

This section presents the experimental evaluation, covering the benchmark, experimental settings, and results. Our goal is to show that our proposal outperforms baseline methods on the NatUKE benchmark for extracting biochemical properties from academic literature. All experiment and model codes are available at our GitHub repository[2].

4.1. NatUKE Benchmark

The NatUKE benchmark addresses the extraction of natural product knowledge from academic literature. Built from more than two thousand expert-annotated biochemical property instances, it was originally used to evaluate unsupervised graph embedding methods such as DeepWalk, Node2Vec, Metapath2Vec, and EPHEN [Do Carmo et al. 2023]. The annotations cover five biochemical properties from $NUBBE_{db}$ [Pilon et al. 2017]: compound name, bioactivity, species, collection site, and isolation type. This benchmark was used in two editions of the International Biochemical Knowledge Extraction Challenge [Marx et al. 2023, Marx et al. 2025], in which new methods were evaluated on the same dataset. After curation, the dataset contains 143 papers with valid properties and serves as a benchmark for predicting one or more property values from academic papers in PDF format. The Benchmark has 448 unique compound names, 33 unique bioactivities, 115 unique species, 51 unique sites, and 5 unique isolation types.

4.2. Experimental Settings

We compare our approach against nine baselines from the NatUKE benchmark. Four unsupervised graph embedding methods, EPHEN, Metapath2Vec, DeepWalk, and Node2Vec, were evaluated in the original benchmark. They used a DOI-centered knowledge graph as input and treated biochemical property extraction as a similarity-based link prediction task [Do Carmo et al. 2023]. We also compare our method with five methods that improved results in NatUKE: a BFS-based knowledge graph embedding strategy [Zope et al. 2023], a preprocessing pipeline combining ChatGPT 3.5-turbo before using the entire KG pipeline with EPHEN [Fröhlich et al. 2023], an NER-based preprocessing strategy also combined with EPHEN [Dichte and Mócsy 2023], K-BERT knowledge integration strategy combined with a regularization model of EPHEN [Chand and Tiwari 2025], and EPHEN^2, an improved version of EPHEN that achieved its best results with Nougat for PDF extraction and BERT embeddings [Do Carmo et al. 2025]. We do not directly compare our approach with the method presented in [Gôlo et al. 2025]. It relies on proprietary paid LLMs and, most importantly, uses a prompting strategy that explicitly provides all candidate properties, essentially transforming the task into a multi-class classification task, which is out of scope for our proposal. However, we put the results from [Gôlo et al. 2025] as the Topline.

We experiment with the open-source models `LLaMa 3.1` with 8 billion parameters [Meta 2024], `Phi 4` with 14 billion parameters [Abdin et al. 2024], and `Qwen 2.5` with 14 billion parameters [Alibaba 2024]. All models were finetuned with a learning rate of $5 \cdot 10^{-4}$ for 25 steps. Regarding the embedding model used for the similarity search step, we compared performance with five trending sentence embedding models on Hugging Face and reported the best result per LLM. To load the papers' textual

[2] `https://github.com/chemnet-io/ballena`

data, originally stored as PDFs, we chose Nougat [Blecher et al. 2023]. Nougat is a transformer-based OCR extraction tool designed for academic papers and was directly responsible for performance gains in biochemical knowledge extraction reported in do Carmo et al. [Do Carmo et al. 2025]. We selected models from Hugging Face that could be loaded on an Nvidia A2 GPU to explore similarity search and reported the best performer for each model: (1) *BAAI/bge-m3* - `Qwen`; (2) *sentence-transformers/all-miniLM-L6-v2* - `Phi`; and (3) *intfloat/multilingual-e5-large* - `LLaMa`. The embedding models *Qwen/Qwen3-Embedding-0.6B* and *sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2* were also tested for all models.

We strictly follow the $hits@k$ metric and evaluation protocol defined in the original NatUKE benchmark [Do Carmo et al. 2023]. This includes train/test splits and the k values defined in the original benchmark. NatUKE defines four train/test splits (20/80%, 40/60%, 60/40%, and 80/20%), with 10 holdouts per stage, each using a different seed. Since this yields 200 training scenarios across five prediction tasks, our finetuned single-task LLMs are trained only on the first-stage split (20% of the data) to reduce computational burden. To assess the impact of this simplification, we also conducted an ablation study with a single LLM that was finetuned using the different stages, i.e., with 20%, 40%, 60%, and finally 80% of the data from a one-holdout from ten. The NatUKE benchmark also defines the k values as follows: compound name ($k = 50$), bioactivity ($k = 5$), species ($k = 50$), collection site ($k = 20$), and isolation type ($k = 1$).

4.3. Results and Discussion

Based on our experiments and the NatUKE benchmark, we present Table 1, which includes prior results and our LLM experiments. The table is divided into three blocks related to different types of experiments: (i) the original NatUKE benchmark graph embedding results; (ii) the extended results of state-of-the-art methods; (iii) finetuned LLMs with similarity search (FSS, i.e., our proposal); and (iv) a topline model that relies on proprietary paid LLMs and a prompting strategy that transforms the task into multi-class classification making it non-scalable. Overall, the different LLMs in our proposal achieve the best and second-best performance across all scenarios. Furthermore, we present an in-depth analysis and an ablation study of the open-source LLMs.

FSS achieves the best overall performance across all tasks except the collection site and the 4th evaluation for compound name extraction. SPHOTA [Chand and Tiwari 2025] is a method that combines Knowledge Injection with embedding propagation, directly using a knowledge graph structure during the extraction. Since the collection site represents locations of the extracted plants, standardized by the annotators, we believe that, for this task, having direct contextual information from the graph improves performance. The topline is a pure LLM-based model that achieves the highest performance; however, as we mentioned, this method presents all options directly in the prompt. On the other hand, FSS-LLaMa and Qwen achieved the next-best results, surpassing EPHEN[2] [Do Carmo et al. 2025], which uses graph regularization with textual embeddings, demonstrating that our method can sample context through finetuning and similarity search without presenting direct context.

Moreover, FSS achieves the best overall performance across compound name, species, bioactivity, and isolation type extractions compared to the baselines. Bioactivity and isolation type in the NatUKE benchmark also aim to extract properties that are

Table 1. Performance of graph embedding methods, methods that improved EPHEN's performance, our proposed finetuned LLMs with similarity search (FSS), and a topline method based on proprietary LLMs. We present $hits@k$ across all stages and properties. k value was chosen based on the NatUKE benchmark. Best results are in bold and second-best underlined.

	Name ($k = 50$)				**Bioact.** ($k = 5$)				**Specie** ($k = 50$)				**Site** ($k = 20$)				**Isolation** ($k = 1$)			
	1st	**2nd**	**3rd**	**4th**	**1st**	**2nd**	**3rd**	**4th**	**1st**	**2nd**	**3rd**	**4th**	**1st**	**2nd**	**3rd**	**4th**	**1st**	**2nd**	**3rd**	**4th**
DeepWalk	.08	.00	.00	.00	.41	.12	.10	.07	.37	.24	.27	.25	.56	.41	.38	.29	.25	.14	.14	.09
Node2Vec	.08	.00	.00	.00	.41	.07	.03	.03	.36	.22	.25	.24	.57	.36	.28	.23	.10	.07	.05	.01
Metapath2Vec	.10	.08	.09	.20	.27	.17	.13	.12	.40	.41	.42	.44	.40	.42	.42	.40	.28	.22	.19	.18
[Do Carmo et al. 2023]	.09	.02	.03	.04	.55	.57	.60	.64	.36	.24	.29	.30	.53	.52	.55	.55	.71	.66	.75	.75
[Zope et al. 2023]	.09	.19	.35	**.83**	.37	.11	.11	.10	.47	.65	.75	.81	.32	.31	.36	.41	.03	.04	.05	.12
[Fröhlich et al. 2023]	.09	.00	.00	.00	.62	.69	.69	.69	.35	.24	.30	.30	.56	.60	.62	.64	.76	.77	.80	.81
[Dichte and Mócsy 2023]	.09	.00	.00	.00	.60	.65	.65	.69	.34	.23	.27	.29	.55	.58	.60	.57	.73	.74	.77	.76
[Do Carmo et al. 2025]	.09	.00	.00	.00	.59	.66	.69	.71	.34	.23	.29	.30	.56	.62	.63	.65	.78	.78	.78	.80
[Chand and Tiwari 2025]	.09	.00	.00	.00	.59	.65	.71	**.87**	.44	.29	.32	.32	**.67**	**.70**	**.69**	**.74**	.78	.78	.76	.71
FSS-LLaMa	**.70**	**.72**	**.73**	.74	**.74**	**.74**	**.76**	.77	.86	.88	.88	.86	.63	.63	.64	.65	.91	**.92**	**.93**	**.94**
FSS-Qwen	.62	.64	.65	.67	.71	.73	.75	.76	**.89**	**.91**	**.92**	**.92**	.62	.63	.64	.66	**.92**	**.92**	**.93**	**.94**
FSS-Phi	.65	.68	.69	.70	.73	.74	.75	.74	.83	.84	.86	.87	.61	.61	.63	.64	.91	**.92**	**.93**	**.94**
Topline [Gôlo et al. 2025]	.92	.92	.90	.90	.81	.82	.83	.83	.97	.97	.97	.95	.76	.76	.77	.77	.96	.96	.96	.97

not necessarily present in the paper; these properties are more likely to be mentioned in the text with the same spelling, and LLM-based approaches achieved the highest performance compared to other baselines. Additionally, the performance of compound name and collection species extraction shows the largest increase compared to baselines, providing further evidence of the benefits of LLMs' general knowledge. These two tasks aim to extract standardized names from numerous knowledge bases across the internet, and this performance increase indicates that LLMs were able to better distinguish between actual innovations and citations within the papers.

Another highlight is isolation-type extraction, as FSS achieved essentially identical results with all three evaluated LLMs and an approximately 0.13-point increase in $hits@1$ across all evaluation stages compared to the baselines. It is also the task where the open-source models achieved the closest performance to the topline ceiling, where it has the strongest advantage with only 5 options for classification. The 5 unique answers were also standardized by the annotators. FSS achieved essentially identical results across three different LLM families, demonstrating that our finetuning and similarity-search pipeline can capture context and improve performance in biochemical knowledge extraction. It also indicates that the NatUKE benchmark was likely not a primary data source during pre-training for the LLMs we experimented with. To further investigate improvements in the pipeline steps, we conducted an ablation study presented in Table 2, including results from: zero-shot, finetuned, similarity search on zero-shot, and FSS.

Across all tasks, the zero-shot LLMs achieved lower performance than the original NatUKE benchmark baselines, showing that open-source LLMs' general knowledge is not sufficient for direct knowledge extraction to a single domain knowledge base at the moment. In the compound name and species extraction tasks, zero-shot LLMs achieved their best results relative to other tasks and baselines. At the same time, our proposal

Table 2. Ablation with zero-shot, finetuned, zero-shot plus similarity search (S. Search), and finetuned LLMs with similarity search (our Proposal). Each block contains the experiments for one out of the three evaluated models (LLaMa, Qwen, and Phi). We present $hits@k$ across all stages and properties. k was chosen based on the NatUKE benchmark. Best results *per model (i.e., block)* are in bold.

	Name (k=50)				Bioact. (k=5)				Specie (k=50)				Site (k=20)				Isolation (k=1)			
	1st	2nd	3rd	4th	1st	2nd	3rd	4th	1st	2nd	3rd	4th	1st	2nd	3rd	4th	1st	2nd	3rd	4th
LLaMa																				
Zero-Shot	.05	.05	.06	.07	.00	.00	.00	.00	.10	.09	.09	.09	.01	.01	.01	.01	.00	.00	.00	.00
Finetuned	.33	.34	.32	.30	.72	.73	.75	.75	.49	.50	.51	.46	.55	.54	.55	.56	**.91**	**.92**	**.93**	**.94**
S. Search	.49	.51	.50	.50	.44	.45	.44	.45	.52	.52	.50	.49	.30	.31	.32	.31	.81	.82	.83	.84
Proposal	**.70**	**.72**	**.73**	**.74**	**.74**	**.74**	**.76**	**.77**	**.86**	**.88**	**.88**	**.86**	**.63**	**.63**	**.64**	**.65**	**.91**	**.92**	**.93**	**.94**
Qwen																				
Zero-Shot	.09	.09	.08	.10	.06	.06	.05	.06	.37	.39	.41	.38	.01	.01	.01	.01	.00	.00	.00	.00
Finetuned	.26	.27	.25	.25	.70	.72	.73	.74	.44	.45	.46	.43	.55	.55	.56	.59	**.92**	**.92**	**.93**	**.94**
S. Search	**.81**	**.82**	**.81**	**.81**	**.71**	.72	.74	.72	.86	.86	.86	.86	.61	.62	**.64**	**.66**	.79	.80	.81	.84
Proposal	.62	.64	.65	.67	**.71**	**.73**	**.75**	**.76**	**.89**	**.91**	**.92**	**.92**	**.62**	**.63**	**.64**	**.66**	**.92**	**.92**	**.93**	**.94**
Phi																				
Zero-shot	.08	.08	.08	.09	.02	.02	.02	.01	.22	.23	.24	.23	.01	.01	.01	.01	.00	.00	.00	.00
Finetuned	.32	.33	.31	.29	.72	.73	.74	.73	.52	.53	.53	.49	.55	.55	.56	.58	**.91**	.91	**.93**	.93
S. Search	**.85**	**.87**	**.87**	**.89**	.72	**.74**	**.75**	**.74**	**.90**	**.90**	**.90**	**.89**	**.61**	**.63**	**.64**	**.66**	.66	.65	.66	.65
Proposal	.65	.68	.69	.70	**.73**	**.74**	**.75**	**.74**	.83	.84	.86	.87	**.61**	.61	.63	.64	**.91**	**.92**	**.93**	**.94**

significantly improved performance. However, when comparing our proposal to the pre-trained and similarity-search experiments for Qwen and Phi, we see another performance increase that causes it to surpass even our proposal with LLaMA. We believe the main reason for Qwen and Phi's decrease in performance during finetuning for compound name extraction is that only 20% of NatUKE's data is available due to the benchmark's train/test split. For more challenging tasks, where pre-training includes a large amount of context data, using similarity search for vocabulary alignment alongside pre-trained LLMs can achieve sufficient performance with lower computational overhead. Furthermore, when comparing the zero-shot and finetuned versions for the bioactivity, collection site, and isolation extraction tasks, we see that finetuning is likely the biggest factor in the full proposal's performance gain relative to benchmarks, indicating that finetuning is important whenever the extraction goals are less covered by pre-training.

Figure 4 shows an ablation study to assess the impact of including all training data from each evaluation stage on finetuning LLaMa. For the previous finetuning results, only 20% of the data from the first evaluation stage was used due to the increased processing time and resources required to evaluate finetuning across all evaluation stages for every LLM and fold in NatUKE. LLaMa was chosen for this ablation study since it is the only model that achieves its best performance with finetuning across all tasks. More often than not, there is an increase in performance when training on the extra data present in the fourth evaluation stage across all extraction tasks. In particular, on the collection species extraction task, there is a jump in performance from 0.53 to 0.65 from the second to the

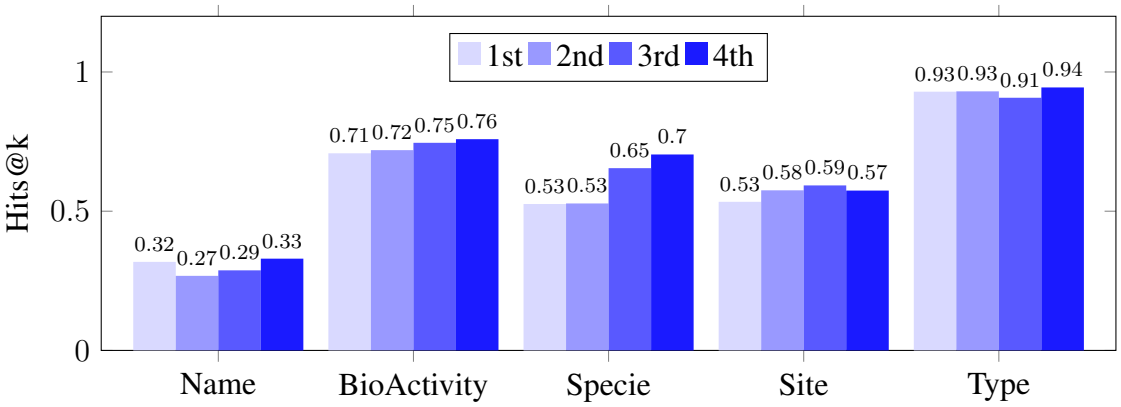


Figure 4. Finetuning of LLaMa in all stages in one fold to evaluate the power of data for biochemical properties prediction.

third stage, and another 0.05 $hits@20$ gain when using the fourth evaluation-stage training split for finetuning. For bioactivity and collection site extraction tasks, the gains are not as significant at 0.05 $hits@5$ and 0.04 $hits@20$ respectively; however, more data still increases performance. For compound name and isolation type, there is no notable gain from using more training data during finetuning. However, these two tasks also achieve the best performance with the other families of pre-trained LLMs, so it is expected that finetuning provides the least gains for LLaMa as well.

5. Conclusion and Future Work

In this work, we presented a novel approach to biochemical knowledge extraction utilizing Large Language Models (LLMs) and similarity search, demonstrating significant performance gains on the NatUKE benchmark. Our Finetuned LLMs with Similarity Search (FSS) consistently achieved the best and second-best results across various experimental scenarios, surpassing scalable state-of-the-art methods. Our experiments show that similarity search improves robustness to textual variation and, when accompanied by LLMs' general knowledge, can accommodate newly emerging properties without requiring a new finetuning cycle. We were also able to show that the NatUKE benchmark is still unlikely to be fully incorporated into LLM pre-training, since extraction tasks in which annotator preferences play a major role saw significant performance gains after finetuning.

Despite the promising results, some limitations remain. The current pipeline requires separate finetuning runs for each extraction task, which increases computational cost and training time. A broader study about the impacts of LLM pre-training on existing benchmarks was not explored. Finally, the sole use of pre-trained or fine-tuned LLMs without knowledge bases to ground them in reality might introduce hallucinations at the first step of the pipeline. Future work will focus on developing a multitask model to extract all five biochemical properties. This work must be done in conjunction with the exploration of other knowledge bases to better evaluate performance. It also increases the available training data for finetuning a multi-task model and the evaluation data, enabling a broader study of the effects of benchmark memorization by LLMs. Simultaneously, the use of context-seeking AI agents with access to grounded data within a Neurosymbolic AI approach can generate better extractions before any finetuning or similarity-search steps.

Acknowledgments

This study was partially funded by FAPEC (23104.003594/2025-12), CNPq (317002/2026-0), FAPESP (2025/07160-0, 2024/08485-8, 2024/15430-5), and SMARTY Project (CHIPS Joint Undertaken as part of the European Union's Horizon Europe research and innovation programme, 101140087).

References

Abdin, M., Aneja, J., Behl, H., Bubeck, S., Eldan, R., Gunasekar, S., Harrison, M., Hewett, R. J., Javaheripi, M., Kauffmann, P., et al. (2024). Phi-4 technical report.

Alibaba, T. (2024). Qwen2 technical report.

Atanasov, A. G., Zotchev, S. B., Dirsch, V. M., and Supuran, C. T. (2021). Natural products in drug discovery: advances and opportunities. *Nature reviews Drug discovery*.

Bitelli, B. and Finger, M. (2025). Numerical information extraction in legal texts using open and closed Large Language Models. In *Anais do XXII Encontro Nacional de Inteligência Artificial e Computacional*, pages 1197–1208, Porto Alegre, RS, Brasil.

Blecher, L., Cucurull, G., Scialom, T., and Stojnic, R. (2023). Nougat: Neural optical understanding for academic documents.

Chand, B. and Tiwari, S. (2025). Sphota: Knowledge graph structure prediction with a hybrid orientation of textual alignment using k-bert. In *Proceedings of the Second Biochemical Knowledge Extraction challenge with Text2KG at ESWC*.

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. (2023). Qlora: Efficient finetuning of quantized llms. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

Dichte, S. S. and Mócsy, I. J. (2023). Improving natural product automatic extraction with named entity recognition. In *Proceedings of the First Biochemical Knowledge Extraction (BiKE) challenge co-located with Text2KG at ESWC*, pages 226–234. CEUR-WS Proceedings.

Do Carmo, P. and Marcacini, R. (2021). Embedding propagation over heterogeneous event networks for link prediction. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 4812–4821. IEEE.

Do Carmo, P. V., Marx, E., Marcacini, R., Valli, M., Silva e Silva, J. V., and Pilon, A. (2023). Natuke: A benchmark for natural product knowledge extraction from academic literature. In *2023 IEEE 17th International Conference on Semantic Computing (ICSC)*, pages 199–203.

Do Carmo, P. V., Silva Gôlo, M. P., Gwozdz, J., Marx, E., and Marcondes Marcacini, R. (2025). Improving natural product knowledge extraction from academic literature with enhanced pdf text extraction and large language models. In *Proceedings of the 40th ACM/SIGAPP Symposium on Applied Computing*, pages 980–987.

D'Souza, J., Laubach, Z., Mustafa, T. A., Zarrieß, S., Frühstückl, R., and Illari, P. (2025). Mining for species, locations, habitats, and ecosystems from scientific papers in invasion biology: A large-scale exploratory study with large language models. *arXiv preprint arXiv:2501.18287*.

Fröhlich, P., Gwozdz, J., and Jooß, M. (2023). Leveraging chatgpt api for enhanced data preprocessing in natuke. In *Proceedings of the First Biochemical Knowledge Extraction (BiKE) challenge co-located with Text2KG at ESWC*, pages 244–255. CEUR-WS Proceedings.

Gôlo, M. P. S., Medeiros Júnior, J. G. B. d., Vilas Boas, G. S., Lobato, F. M. F., Silva, D. F., and Marcacini, R. M. (2025). Large language models ensemble for biochemical properties detection in scientific articles. In *Proceedings of the Second Biochemical Knowledge Extraction (BiKE) challenge co-located with Text2KG at ESWC*.

Gozzi, M. and Di Maio, F. (2024). Comparative analysis of prompt strategies for large language models: Single-task vs. multitask prompts. *Electronics*, 13(23):4712.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. (2022). Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Marx, E., Valli, M., da Silva e Silva, J., and do Carmo, P. V. (2025). Preface of the second international biochemical knowledge extraction challenge (bike). In *LLM-Integrated Knowledge Graph Generation From Text and International BiKE Challenge*, volume Vol-4020 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Marx, E., Valli, M., da Silva e Silva, J., Tiwari, S., and do Carmo, P. (2023). Preface of the first international biochemical knowledge extraction challenge (bike). In *Joint Proceedings of TEXT2KG 2023 and BiKE 2023*, volume Vol-3447 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Meta (2024). The llama 3 herd of models.

Newman, D. J. and Cragg, G. M. (2020). Natural products as sources of new drugs over the nearly four decades from 01/1981 to 09/2019. *Journal of natural products*.

Oliveira, M. S., Pio, P. B., and Garcia, L. P. (2025). Generative slms meet brazilian legal documents: Efficient ner via lora fine-tuning. In *Brazilian Conference on Intelligent Systems*, pages 586–599. Springer.

Pilon, A. C., Valli, M., Dametto, A. C., Pinto, M. E. F., Freire, R. T., Castro-Gamboa, I., Andricopulo, A. D., and Bolzani, V. S. (2017). Nubbedb: an updated database to uncover chemical and biological information from brazilian biodiversity. *Scientific Reports*, 7(1):7215.

Réda, C., Kaufmann, E., and Delahaye-Duriez, A. (2020). Machine learning applications in drug development. *Computational and structural biotechnology journal*, 18.

Zhai, C. (2024). Large language models and future of information retrieval: opportunities and challenges. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 481–490.

Zhang, L., Song, J., Kong, L., Yuan, T., Li, W., Zhang, W., Hou, B., Lu, Y., and Du, G. (2020). The strategies and techniques of drug discovery from natural products. *Pharmacology & Therapeutics*, 216.

Zope, B., Mishra, S., and Tiwari, S. (2023). Enhancing biochemical extraction with bfs-driven knowledge graph embedding approach. In *Proceedings of the First Biochemical Knowledge Extraction (BiKE) challenge co-located with Text2KG at ESWC*. CEUR-WS Proceedings.