Enhancing Scholarly Knowledge Graphs via Domain-Specific Entity Detection and Linking

Nicolau Duran-Silva[1,2], **César Parra-Rojas**[1], **Pablo Accuosto**[1], **Julian Moreno-Schneider**[3], **Georg Rehm**[3]
[1]SIRIS Lab, Research Division of SIRIS Academic, Barcelona, Spain
[2]LaSTUS Lab, TALN Group, Universitat Pompeu Fabra, Barcelona, Spain
[3] Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (DFKI), Berlin, Germany
{nicolau.duransilva, cesar.parra, pablo.accuosto}@sirisacademic.com,
{julian.moreno_schneider, georg.rehm}@dfki.de

Abstract

Navigating scholarly content presents important challenges due to the fragmented and heterogeneous nature of research production and outputs. Scholarly Knowledge Graphs offer an efficient means to integrate diverse data sources and consolidate knowledge across outputs in a structured manner. This representation, combined with the grounding of unstructured textual data to well-defined research-related concepts, has great potential for enhancing knowledge discovery and supporting researchers navigating through vast amounts of scientific information. Knowledge extraction capabilities are commonly limited by the availability of large collections of annotated data supporting named-entity recognition (NER) and linking (EL), and the enormous effort that their elaboration entails for domain experts. Recent advances in natural language processing and generative artificial intelligence provide valuable opportunities to reduce the data annotation toll and produce high-quality NER with minimal expert involvement. Here, we present a pipeline for domain-specific NER and EL, leveraging LLMs and knowledge from experts in a human-in-the-loop approach to streamline the annotation process, along with transformer-based models and few-shot techniques. While the application focuses on showcasing four specific domains, the pipeline is designed to be flexible and domain agnostic for scientific fields.

Keywords: Named Entity Recognition, Entity Linking, Scholarly Document Processing

1. Introduction

Scholarly knowledge is inherently heterogeneous in nature, and the effective integration of its components has become increasingly critical in the current data-intensive research landscape (Manghi, 2024). Research outputs can take different forms—e.g., publications, datasets, software—and be scattered across diverse databases and repositories (Manghi et al., 2019; Peroni and Shotton, 2020; Priem et al., 2022; Hendricks et al., 2020; Wilkinson, 2010; Neylon et al., 2026). In addition to this, each research community has its own terminology and standards, and cataloguing and curation practices usually do not translate to other fields (Friedman et al., 2002). Even within a single field, the conventions adopted by different researchers may be inconsistent and scientific knowledge be often fragmented, unstructured and/or duplicated across outputs under different names, hindering data-informed decision-making by research-funding and research-performing organisations, as well as individual researchers themselves.

In recent years, scholarly knowledge graphs (SKGs) (Jaradeh et al., 2019; Manghi et al., 2019; Priem et al., 2022; Peroni and Shotton, 2020) have emerged as a core infrastructure for metadata-driven discovery. By integrating general and domain-specific knowledge from diverse data sources into interconnected entities and relationships, and linking concepts across different types of research outputs, SKGs transform fragmented and heterogeneous information into a structured and flexible representation that supports knowledge discovery, facilitating the uncovering of hidden patterns and connections between research-related entities (Vergoulis et al., 2019).

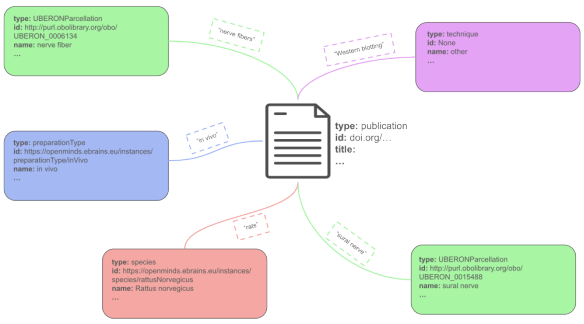

Figure 1: Example of enriching SKG in neuroscience domain with entity detection and linking.

Enriching domain-specific SKGs with relevant entities has the potential to streamline the research process (Leipzig et al., 2021; Wu et al., 2026), as researchers can uncover novel associations arising from the identification and linking of men-

Proceedings of the 3rd Int. Workshop on Natural Scientific Language Processing (NSLP 2026) @ LREC 2026, pages 140–154
12 May 2026. © ELRA Language Resources Association (ELRA), 2026

tions of these entities across different works and/or databases, which can in turn inform potential avenues for further investigation (Taslimi et al., 2025; Liu et al., 2024). This information can originate from curated sources, with some fields boasting a plethora of highly developed platforms and resources for commonly-encountered entities—e.g., diseases, genes, or chemicals in the case of biomedical research—either in the sense of being added as metadata at the moment of creation of the research output or to be employed as training data for automated extraction pipelines via named entity recognition (NER) (Hong et al., 2020; Seow et al., 2025).

Due to the diversity and heterogeneity of research domains, however, these "canonical" or standard collections of domain-relevant terms may not necessarily align with the specific needs of a given subfield. For other fields that lack an extensive range of curated sources, there may be limited or non-existent training data for the identification of relevant entities (Yakimovich et al., 2021; Xu et al., 2023). In this regard, the advent of large language models presents valuable opportunities to enhance the scope of SKGs and increase entity coverage—thus grounding the unstructured text within the graphs in precise concepts—by lowering the barrier for enrichment through high-quality NER and entity linking, achievable with minimal task-specific training, in combination with the ability to generate domain- and context-aware synthetic examples (Shlyk et al., 2024; Zhang et al., 2025b; Xin et al., 2025). This, in turn, lowers curation friction, enabling researchers to navigate domain entities that were previously infeasible due to lack of labelled data.

We posit that high-recall, high-precision named-entity recognition tuned to specific scientific verticals is the keystone for unlocking richer SKGs. To this end, we introduce a flexible pipeline that (i) automatically LLM-generated preliminary annotations (pseudo-labels) for a dataset of full-text article sections sourced from open repositories, (ii) involves domain experts for pseudo-label validation, and (iii) fine-tunes transformer-based NER models for the five pilot domains of the EU-funded SciLake project —Neuroscience, Energy, Maritime Transport, and Connected Cooperative Autonomous Mobility (CCAM). The extracted entities can then be seamlessly fed back into interoperable SKGs. Due to the specific needs and idiosyncrasies of the target domains, the effort is focused towards creating a solution that can accommodate them all individually, in order to provide a tailored experience while being as domain-agnostic as possible. This ensures the enrichment pipeline can be adapted to new use cases in the future with minimal effort (Vergoulis et al., 2019).

We release code, models, and dataset under an open license to support research and further work on domain-adaptive NER and EL for scientific content[1].

2. Related Work

Concept recognition, commonly referred to as Named Entity Recognition (NER) and Entity Linking (EL), is a fundamental task in scholarly document processing and is essential for structuring scientific knowledge and enabling downstream applications such as knowledge graph construction, semantic search, and information extraction (Li and Zhang, 2023; Shlyk et al., 2024). Scientific text presents particular challenges due to specialised terminology, long-tail entities, and frequent ambiguity (Zhu et al., 2023). In domains such as neuroscience, transport, and energy research, substantial domain knowledge is embedded in complex and highly specialised concepts, making accurate identification and grounding of entities critical for connecting unstructured text with structured knowledge bases (Mihalcea and Csomai, 2007; Wu et al., 2020a; Xin et al., 2025). A persistent challenge for scientific NER and EL is the scarcity of labelled training data: supervised approaches typically require large annotated corpora, yet annotation is costly and time-consuming, as it demands domain expertise, and model performance is strongly correlated with the amount of available labelled data (Zhu and Jiang, 2021; Yakimovich et al., 2021).

To address domain mismatch, a range of domain-adaptive pre-trained language models have been proposed (Beltagy et al., 2019; Gururangan et al., 2020; Hong et al., 2022), which show that pre-training on scientific corpora substantially improves downstream performance while further emphasising the benefits of targeted pre-training, the latter explicitly leveraging citation links. Other approaches integrate structured or external knowledge sources, including document relations and ontologies, as in SPECTER (Cohan et al., 2020), KeBioLM (Yuan et al., 2021), and DiseaseBERT (He et al., 2020). Despite these advances, adapting such models to fine-grained, domain-specific entity recognition remains challenging, particularly in low-resource settings (Zhu and Jiang, 2021).

Transformer-based architectures dominate NER in scientific text (Seow et al., 2025), supported by scientific datasets, mainly from the biomedical domain, such as GENIA (Kim et al., 2011), BC4CDR (Li et al., 2016), AIONER (Luo et al., 2023) or SciERC (Luan et al., 2018). To reduce reliance on large labelled corpora, label-efficient approaches such as GLiNER (Zaratiana et al., 2024)

[1] Codes and datasets available at https://github.com/sirisacademic/meloner

formulate NER as span classification conditioned on natural-language label descriptions, enabling zero-shot and few-shot recognition, while GLiNER-BioMed (Yazdani et al., 2025) adapts this paradigm to biomedical text. Although these methods improve flexibility and reduce annotation costs, performance remains sensitive to label definitions and contextual coverage, particularly for highly specialised or domain-specific concepts.

Entity Linking (EL) aligns entity mentions in unstructured text with entries in a knowledge base. Most EL systems follow a two-stage pipeline consisting of candidate retrieval and candidate re-ranking (Mihalcea and Csomai, 2007; Wu et al., 2020a; Xu et al., 2023). Candidate retrieval can be based on sparse methods (e.g., TF–IDF or BM25) or dense methods that embed mentions and entity descriptions into a shared vector space (Mustafa et al., 2024). Dense retrieval approaches, such as BLINK (Wu et al., 2020b), employ bi-encoders for scalable candidate retrieval and cross-encoders for re-ranking, achieving strong performance by jointly modelling mention context and entity descriptions. However, most EL methods assume that entities observed at test time were seen during training, which is unrealistic for large and continuously evolving KBs such as Wikidata or UMLS (Ayoola et al., 2022; Zhu et al., 2023).

Some studies explore generative (Xiao et al., 2023), context augmentation (Xin et al., 2025) and retrieval-augmented approaches for entity linking (Shlyk et al., 2024) in order to bypass the training data bottleneck. LLMAEL (Xin et al., 2025) uses generative LLMs as knowledgeable context augmenters, generating mention-centric descriptions that supplement the original text before applying traditional EL models. Rather than performing EL directly with LLMs, this approach enhances the input context while preserving task-specific EL architectures. However, a major challenge in entity linking is determining none of the candidates is correct(Zhou et al., 2024), which can be challenging.

In summary, prior works show the effectiveness of domain-adaptive language models, dense retrieval-based EL architectures, and emerging context augmentation techniques. However, existing approaches often rely on large annotated datasets, focus on canonical entity types, or make assumptions that do not hold in heterogeneous and low-resource scientific domains. Our work builds on these insights by combining weak supervision, expert validation, and modular NER and EL components to support domain-specific entity recognition and linking across four scientific domains.

3. Domains of Interest and their Entities and Knowledge Bases

Each scientific domain considered in this work defines a set of domain-specific entity types of interest and a corresponding reference knowledge base (KB) for entity linking. These entities are not covered by existing NER datasets and were defined in close collaboration with domain experts from each pilot community. Below, we summarise the target domains and their selected entity types; concrete examples and the associated KBs are reported in Appendix 7.

- **Neuroscience**: `technique`, `preparation type`, `parcellation`, `species`, `biological sex`.
- **Energy**: `energy type`, `energy storage`.
- **Maritime Transport**: `vessel type`.
- **Connected Cooperative Autonomous Mobility (CCAM)**: `vehicle type`, `VRU type`, `communication type`, `entity connection type`, `sensor type`, `scenario type`, `level of automation`.

For each domain, entities extracted from text are linked to a domain-appropriate taxonomy or ontology, including openMINDS[2] and UBERON[3] for Neuroscience, IRENA[4] for Energy, AIS vessel classifications[5] for Maritime Transport, and a subset of SINFONICA[6] and FAME[7] for CCAM.

4. Methods and Materials

4.1. Data

We introduce a collection of scientific papers[8] parsed and segmented by section with GROBID (Lopez, 2009), primarily designed for research in the development and evaluation of NLP models. This dataset contains 1,000 full-text papers from various scientific domains, including Neuroscience, Transport, and Energy, in addition to Cancer, along with 1,000 other random papers from general scientific domains (Pàmies et al., 2023). The list has been curated so that all papers in it are published using licenses that allow for legal reuse, specifically CC-BY and Public Domain. The papers

[2] openminds.docs.om-i.org
[3] ebi.ac.uk
[4] irena.org
[5] api.vtexplorer.com
[6] sinfonica.eu
[7] taxonomy.connectedautomateddriving.eu
[8] Available at https://huggingface.co/datasets/SIRIS-Lab/scilake-fulltext-corpus

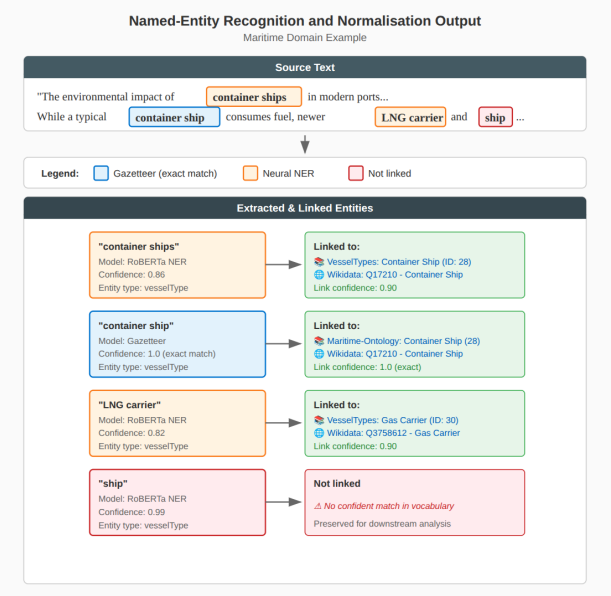

Figure 2: Example output showing entities identified and linked to domain KBs.

included in this dataset were sourced through OpenAIRE (Manghi et al., 2019), with random selection to ensure diverse content. The license information was verified by cross-referencing the publisher's landing pages, metadata from OpenAIRE, and the Unpaywall (Chawla, 2017).

4.2. Annotation

4.2.1. NER

To minimise manual annotation effort while retaining domain accuracy, we adopted a two-step annotation strategy combining LLM-based pseudo-labelling with expert refinement. For each pilot domain, we sampled 150–200 document excerpts from the corpus, consisting of titles, abstracts, and randomly selected body sections per paper (Table 1). In the first step, pseudo-labels were generated using a large language model (Claude 3.7 Sonnet)[9], prompted with general extraction instructions and domain- and entity-specific guidelines (Appendix B). The model produced candidate entity spans, and their corresponding entity types. In the second step, the pseudo-labelled data was reviewed and refined by domain experts using the Argilla (Daniel and Francisco, 2023).[10] Annotators validated entity spans and types, corrected errors, and added missing entities. This human-in-the-loop process ensured both domain relevance and annotation consistency, while substantially reducing the overall labelling effort compared to fully manual annotation.

[9] Anthropic Claude Sonnet 3.7 documentation
[10] https://docs.argilla.io/

Dataset	# Entities	# Samples
Neuroscience	5	182
Energy	2	147
CCAM	7	191
Maritime transport	1	189

Table 1: Overview of the initial annotated datasets, number of entity types defined for the domain and number of samples.

After validation, the annotated documents were segmented at the sentence level and split into training and test sets, reserving 20% of the data for evaluation in each domain.

4.2.2. Entity Linking

For each domain, we have a domain-specific target knowledge base (KB) to support entity normalisation. As summarised in Table 2, KB sizes vary substantially across domains, ranging from 22 entities in Maritime Transport to 2,565 in Neuroscience, reflecting differences in domain scope and conceptual granularity. For each domain, between 700-1,000 entity mentions were manually annotated by domain experts with links to KB entities when a suitable match existed.

Domain	# KB Entities	# Samples	NoC (%)
Neuroscience	2,565	1,000	47.9%
Energy	267	964	20.9%
CCAM	176	734	35.0%
Maritime	22	1,000	20.6%

Table 2: Statistics of the entity linking datasets. None-of-the-Candidates (NoC) denotes the number of mentions for which no suitable candidate existed in the target knowledge base.

To improve coverage, KB entries were enriched with GENRE model (Cao et al., 2021) to Wikipedia identifiers when available, enabling the inclusion of alternative labels, aliases, and acronyms. Each KB is stored in a structured format including a `unique identifier, preferred name, (optional) hierarchy, (optional) type, description, and (optional) Wikidata metadata (id, variants).` Entity linking candidates were generated from NER-extracted mentions and a gazetteer built from KB labels and aliases. During annotation, mentions were either linked to the most appropriate KB entry or marked as `UNLINKED` when no corresponding concept existed in the KB.

Example 1 (CCAM). *"Since Distributed Acoustic Sensing constitutes a (curvi-)linear array of single-component **seismic sensors**, it is*

amenable to beamforming analysis."
→ linked to `SINFONICA::Sensor`

Example 2 (Maritime Transport). *"Both approaches are tested for two case studies, a bulk carrier and a **small cruise ship**."*
→ linked to `Passenger Ship`

4.3. Models & Training

4.3.1. NER

We fine-tune transformer encoder models for token classification on each domain-specific dataset, as suggested by (Devlin et al., 2019), splitting 80/20 across domains. Our model selection considers general-purpose language models (RoBERTa-base and RoBERTa-large (Liu et al., 2019)), and scientific language models pre-trained on academic documents (SPECTER2 (Cohan et al., 2020; Singh et al., 2022) and ScholarBERT (Hong et al., 2022)). We train all four general-purpose and scientific backbones from scratch on each domain's labelled data. In addition, across all four domains we fine-tune a GliNER model (Zaratiana et al., 2024), a generalist entity-recognition framework that accepts arbitrary natural-language label, providing a complementary zero-shot baseline. Full hyperparameters and training details are provided in Appendix C.

4.3.2. Entity Linking

We frame entity linking as a two-stage retrieval and reranking pipeline, systematically evaluating combinations of retrievers, rerankers, and an optional LLM-based context augmentation inspired from (Xin et al., 2025) step across our four domains. We conduct an ablation study of different retriever and reranking models, with focus on NoC. However, most models used are not explicitly trained for entity-level retrieval and ranking, they are optimized for longer text passages, which limits their effectiveness when only minimal contextual information is available. We split each domain dataset as 50% for train and 50% for test. We use TF-IDF retrieval combined with a threshold-based decision rule as a baseline for entity linking, representing a competitive baseline approach in low-resource settings.

Retrieval. Given a mention in context, the retriever returns the top-k candidates (k=10) from a domain-specific knowledge base constructed from each taxonomy's concepts, descriptions, and aliases. Each entity is encoded by concatenating the concept label, its description, Wikidata aliases, and synonyms (separated by `[SEP]` tokens). Mentions are encoded as the entity surface form followed by its surrounding sentence context (formatted as `"{mention}. Context: {sentence}"`). Following (Mustafa et al., 2024), we compare a sparse baseline (character n-gram TF-IDF) with six dense bi-encoder retrievers spanning general-purpose models (BGE-M3 (Chen et al., 2024), E5-base and E5-large-instruct (Wang et al., 2024), Jina Embeddings v3 (Sturua et al., 2024)), a biomedical model (SapBERT (Liu et al., 2021)), and a scientific model (SPECTER (Cohan et al., 2020)). We additionally include our fine-tuned variant of E5-base, trained on the entity linking pairs from our training split.

Reranking. Candidates are then rescored by one of five rerankers representing different architectural families: a cross-encoder baseline (MS MARCO MiniLM (Reimers and Gurevych, 2019)), Jina Reranker v3 (Wang et al., 2025), Qwen3-Reranker-0.6B (Zhang et al., 2025a) using yes/no logit scoring, and our fine-tuned variant of the Qwen3 reranker trained on our entity linking data, and two generative models (Qwen3-1.7B and Qwen3-8B (Team, 2025)), suggested by (Shlyk et al., 2024), that are asked to output either a candidate index or `REJECT` for NoC prediction.

Context augmentation. Following Xin et al. (2025), we optionally augment mention context with LLM-generated entity descriptions before reranking. A Qwen3-1.7B model (Team, 2025) generates a short technical description of the mention using domain-specific few-shot prompts, described in Appendix D, which is appended to the original context before being passed to the reranker.

NoC prediction and threshold tuning. As described by (Zhou et al., 2024), a key challenge in zero-shot entity linking re-ranking settings is that mentions may not correspond to any concept in the taxonomy (NoC entities). Generative rerankers handle this natively via the `REJECT` option. For score-based retrievers and rerankers, we use the training split (50%) to tune a decision threshold: mentions whose top-candidate score falls below the threshold are predicted as NoC. All reported metrics are computed on the held-out test split, focusing on evaluating the None-of-the-Candidates (NoC).

Fine-tuning. To assess importance of task adaptation given that most retriever and rerankers are trained on textual passages, not on entity linking. Two components are fine-tuned on the training split: an E5-base retriever (Wang et al., 2024), adapted via contrastive learning on (mention, gold concept) pairs, and a Qwen3-Reranker-0.6B (Zhang et al., 2025a), fine-tuned on candidate lists, using one positive and nine negative candidates per mention, where negatives are drawn from the top candidates

retrieved by a TF-IDF retriever. Hyperparameters and training details are provided in Appendix C.

5. Results & Discussion

5.1. Evaluation

We evaluate NER and Entity Linking (EL) separately using test splits for each domain. For NER, annotated sentence-level data is split into training and test sets, reserving 20% for test, and performance is measured using standard micro-average span-based precision, recall, and F1-score. For EL, annotated mentions are split evenly into training and test sets (50/50), where the training split is used for threshold tuning or model fine-tuning, when applicable. EL performance is evaluated using Accuracy@1 on the test set, complemented by Recall@K to assess retrieval quality and by precision and recall for None-of-the-Candidates (NoC) prediction.

5.2. NER

Table 3 reports micro-averaged F1-score across the four scientific domains. Zero-shot GLiNER models perform poorly in all settings, confirming that domain-specific supervision is essential for extracting specialised scientific and technical entities. Fine-tuning leads to substantial improvements for all models.

Model	CCAM	Mar.	Ener.	Neuro.
ZERO SHOT				
Gliner-large	.21	.37	.21	.29
Gliner-medium	.16	.24	.19	.24
Gliner-small	.16	.15	.28	.23
FINE-TUNED				
Gliner-large	.765	.356	.359	.660
Gliner-medium	.798	.398	.418	.680
Gliner-small	.738	.468	.476	.636
Roberta-base	.739	.745	.652	.697
Roberta-large	.804	.861	.652	.785
ScholarBERT	.775	.798	.670	.721
Specter2	.762	.822	.647	.763

Table 3: Named Entity Recognition performance across domains, reported using micro-averaged F1-score on held-out test split.

Among fine-tuned approaches, token classification models consistently outperform prompt-based methods, with RoBERTa-large achieving the strongest overall performance. Scientific pre-trained models (ScholarBERT and SPECTER2) provide stable gains in domains such as Neuroscience, CCAM and Maritime transport. Fine-tuned GLiNER models remain competitive in some domains but show higher variance, suggesting sensitivity to label semantics and prompt formulation. To improve GIINER, we note that despite the fine-tuning with the specific entity label names, more descriptive label—e.g., "neuroanatomical region" instead of `UBERONParcellation`, or "ship" or "vessel" replacing `vesselType`, to list some straightforward examples—may further improve prompt-based extraction. We leave this for future work.

Finally, the small size of the annotated training and test data remains a key limitation across domains. While fine-tuning yields strong gains, performance on sparse or semantically broad entity types remains lower. Future work could explore data augmentation strategies, such as synthetic example generation, as well as alternative few-shot and instruction-based information extraction methods, to further reduce annotation costs while maintaining domain accuracy.

5.3. Entity Linking

We evaluate entity linking as a two-stage retrieval–reranking steps and report results of each component to isolate its contribution and capacities. We first assess candidate retrieval, reranking, and the effect of an intermediate LLM-based context augmentation step.

Table 4 reports retrieval-only performance using Recall@K, which represents an upper bound on downstream linking reranking. Across domains, some dense retrievers generally outperform the TF-IDF lexical baseline, although it remains very competitive, specially for Neuroscience and Maritime. Retrieval-oriented encoders (E5, BGE-M3) outperform document-encoder models (SPECTER, JINA), indicating that task alignment for short-text retrieval is more challenging due to the lack of enough context, being SPECTER trained on scholarly documents. Instruction-tuned retrieval (E5-INSTRUCT) shows strong gains in CCAM and Maritime but degrades in Neuroscience and Energy, suggesting domain-sensitive instruction effects. Fine-tuning E5 on (mention, gold concept) pairs leads to consistent and substantial gains, with Recall@10 approaching saturation in all domains and Recall@1 improving by more than 30 points on average. This suggests that most errors in retrievers could stem from domain mismatch and most models are not trained for entity-level retrieval, rather than inherent ambiguity in the candidate set. Despite high Recall@10, Recall@1 remains comparatively low for non-fine-tuned retrievers, motivating the use of reranking.

Table 5 reports macro-averaged results across domains for each retriever–reranker combination, including Accuracy@1 and precision/recall for None-of-the-Candidates (NoC) prediction. Values

Retriever	R@1	R@5	R@10
TFIDF			
Neuroscience	.421	.633	.734
CCAM	.377	.701	.734
Energy	.373	.686	.743
Maritime	.546	.744	.876
SAPBERT			
Neuroscience	.270	.544	.633
CCAM	.270	.467	.537
Energy	.267	.478	.589
Maritime	.355	.672	.806
SPECTER			
Neuroscience	.286	.471	.556
CCAM	.295	.520	.627
Energy	.298	.488	.607
Maritime	.318	.648	.854
BGE-M3			
Neuroscience	.351	.653	.730
CCAM	.410	.664	.770
Energy	.424	.661	.740
Maritime	.447	.769	.901
E5			
Neuroscience	.351	.645	.741
CCAM	.402	.709	.811
Energy	.398	.702	.787
Maritime	.496	.836	.931
E5-INSTRUCT			
Neuroscience	.154	.282	.344
CCAM	.512	.803	.902
Energy	.201	.504	.697
Maritime	.543	.896	.965
JINA			
Neuroscience	.189	.398	.471
CCAM	.299	.541	.639
Energy	.265	.530	.630
Maritime	.397	.767	.908
E5 (FINE-TUNED)			
Neuroscience	.784	.938	.946
CCAM	.758	.939	.963
Energy	.640	.846	.923
Maritime	.739	.928	.978

Table 4: Retriever-only entity linking performance across domains. Recall@K indicates the upper bound for reranking accuracy.

Reranker	Acc@1	NoC P.	NoC R.
TFIDF (avg R@10 = .764)			
Threshold baseline	.500	0.555	.803
Cross-encoder	.440 (-0.10)	0.435	.773
Jina Reranker	.620 (+0.03)	0.543	.849
Qwen Reranker	.589 (+0.03)	0.546	.898
Generative (Qwen 1.7B)	.446 (+0.05)	0.715	.353
Generative (Qwen 8B)	.595 (-0.01)	0.495	.866
Qwen Reranker (fine-tuned)	.671 (-0.02)	0.555	.852
BGE-M3 (avg R@10 = .785)			
Threshold baseline	.437	0.420	.690
Cross-encoder	.414 (-0.11)	0.400	.779
Jina Reranker	.564 (+0.01)	0.490	.795
Qwen Reranker	.535 (+0.05)	0.484	.846
Generative (Qwen 1.7B)	.417 (+0.05)	0.566	.251
Generative (Qwen 8B)	.589 (-0.01)	0.495	.832
Qwen Reranker (fine-tuned)	.644 (-0.02)	0.527	.875
E5 (avg R@10 = .817)			
Threshold baseline	.445	0.442	.693
Cross-encoder	.413 (-0.10)	0.391	.799
Jina Reranker	.565 (+0.02)	0.466	.851
Qwen Reranker	.539 (+0.04)	0.483	.861
Generative (Qwen 1.7B)	.417 (+0.04)	0.562	.257
Generative (Qwen 8B)	.589 (-0.01)	0.502	.835
Qwen Reranker (fine-tuned)	.657 (-0.02)	0.558	.815
E5-INSTRUCT (avg R@10 = .727)			
Threshold baseline	.419	0.404	.806
Cross-encoder	.398 (-0.10)	0.386	.793
Jina Reranker	.522 (+0.02)	0.438	.854
Qwen Reranker	.521 (+0.02)	0.464	.869
Generative (Qwen 1.7B)	.338 (+0.03)	0.483	.106
Generative (Qwen 8B)	.545 (-0.00)	0.525	.775
Qwen Reranker (fine-tuned)	.617 (-0.03)	0.515	.854
JINA (avg R@10 = .662)			
Threshold baseline	.389	0.389	.818
Cross-encoder	.407 (-0.09)	0.382	.785
Jina Reranker	.526 (+0.01)	0.446	.815
Qwen Reranker	.513 (+0.04)	0.463	.862
Generative (Qwen 1.7B)	.408 (+0.05)	0.503	.398
Generative (Qwen 8B)	.527 (+0.00)	0.436	.859
Qwen Reranker (fine-tuned)	.589 (-0.01)	0.458	.886
E5-FINETUNED (avg R@10 = .952)			
Threshold baseline	.689	0.617	.778
Cross-encoder	.422 (-0.12)	0.410	.737
Jina Reranker	.598 (+0.02)	0.519	.778
Qwen Reranker	.573 (+0.04)	0.532	.863
Generative (Qwen 1.7B)	.419 (+0.08)	0.603	.218
Generative (Qwen 8B)	.598 (-0.01)	0.504	.839
Qwen Reranker (fine-tuned)	.681 (-0.00)	0.577	.821

Table 5: Entity linking performance of reranking models using a fixed retriever (with R@10 ceiling), macro-averaged across domains. Values in parentheses indicate the change in Acc@1 obtained by augmenting mention context with LLM-generated descriptions. Thresholds are fixed on the training split.

in parentheses indicate the average change in Acc@1 when augmenting mention context with LLM-generated descriptions.

Across all retrievers, the threshold-based baseline provides a strong and competitive reference point, particularly for NoC detection. Its high NoC recall indicates that simple score calibration is already effective at rejecting spurious candidates, and most rerankers do not substantially improve this aspect. This highlights that NoC prediction is largely constrained by retrieval confidence rather than by ranking capacity.

Generative rerankers show mixed behaviour, while they occasionally improve accuracy, they tend to trade off NoC recall for precision, making them less reliable in settings where rejection is critical. Fine-tuning the Qwen reranker leads to consistent gains over its base version, but these improvements remain moderate compared to the substantial gains achieved through retriever fine-tuning. This suggests that reranking performance is still bottlenecked by candidate quality and by the lim-

ited amount of task-specific supervision available for reranker adaptation.

The impact of LLM-based context augmentation remains inconsistent across domains. While small improvements are observed in some configurations, augmentation does not systematically improve reranking performance and can introduce noise or redundant information in some cases. This suggests that current prompt-based augmentation strategies may lack the level of control required for fine-grained entity-level disambiguation. Retrieval quality remains the dominant factor in the overall performance, with context augmentation providing only secondary gains.

Domain	Acc@1	NoC P.	NoC R.
CCAM	.576 (+0.14)	.539	.671
Energy	.661 (+0.08)	.503	.866
Maritime	.646 (+0.11)	.422	.851
Neuroscience	.830 (+0.15)	.771	.963

Table 6: Entity linking performance across domains using a fine-tuned E5 retriever and fine-tuned Qwen reranker. Values in parentheses indicate the absolute change in Acc@1 compared to the non-finetuned configuration.

Table 6 reports per-domain results using the fully fine-tuned configuration (E5 retriever + Qwen reranker). Performance varies considerably across domains, with the strongest gains observed in Neuroscience and the weakest in CCAM and Maritime. Interestingly, this variation correlates more strongly with size of their KBs. Domains with large, fine-grained taxonomies and lexically specific entities (e.g., neuroanatomical regions) benefit from closer surface-form alignment between mentions and concepts. In contrast, the maritime KB involve an smaller taxonomy but require higher levels of abstraction (e.g., mapping “small cruise ship” to “passenger ship”’), increasing semantic distance between mentions and KB entries and making linking more challenging. However, under specific configurations, context augmentation can provide benefits across domains with both large and small KBs, as observed in Neuroscience and CCAM/Maritime in Table 6. This suggests that its effectiveness is not only dependent on KB size, but also on the interaction between retrieval quality and the generated contextual information.

These results highlight that cross-domain variability is not only a consequence of model performance, but is strongly influenced by properties of the target knowledge bases and domains. In particular, domains with larger and more fine-grained taxonomies benefit from improved lexical alignment, while smaller or more abstract taxonomies require higher levels of semantic generalisation. This suggests that entity linking performance in scientific domains is inherently tied to knowledge base properties, and that domain-agnostic pipelines must account for differences in conceptual granularity and coverage when transferring across areas.

Despite the substantial gains achieved through fine-tuning and modular evaluation, named entity recognition and entity linking remain challenging in scientific domains with limited and specialised training data. Domain mismatch, abstract terminology, and sparse supervision continue to affect both retrieval and reranking, particularly for low-frequency or weakly lexicalised entities, specially dealing with low-quality KBs.

Moreover, the enrichment of large-scale SKGs introduces constraints on efficiency and scalability. While large language models enable richer context modelling and semantic abstraction, their computational cost limits large-scale deployment. Future work must therefore balance enrichment quality with throughput, exploring entity-oriented reranking, favouring lightweight or hybrid pipelines that apply expensive components selectively. In this setting, human-in-the-loop approaches remain essential in low-resource and highly-specialised domains. Future work should also explore more robust rejection mechanisms and tighter integration between retrieval and reranking for improved NoC prediction, as handling these cases remains a persistent challenge.

6. Conclusions

In this work, we presented a modular evaluation of named entity recognition and entity linking across four scientific domains, explicitly separating retrieval, reranking, and context augmentation, motivated by the fragmented and heterogeneous nature of scholarly content and the need to ground unstructured text to well-defined concepts in Scholarly Knowledge Graphs and domain-specific knowledge bases. Our results suggest that domain-specific fine-tuning is useful for both tasks, with retrieval quality emerging as the primary driver of end-to-end entity linking performance. Reranking and context augmentation provide consistent but secondary improvements, particularly in ambiguous cases. While the observed performance improvements are incremental, our contribution lies in providing a modular and domain-adaptive pipeline, together with a systematic evaluation across different scientific domains, which remains underexplored in prior work. Overall, our findings highlight the importance of efficient, domain-adaptive, and human-guided approaches for scalable knowledge graph enrichment in low-resource domains.

7. Acknowledgements

Supported by the Industrial Doctorates Plan of the Department of Research and Universities of the Generalitat de Catalunya, by Departament de Recerca i Universitats de la Generalitat de Catalunya (grant reference 2022/DI /00017). This work was co-funded by the EU HORIZON project SciLake (Grant Agreement 101058573) and the consortium NFDI for Data Science and Artificial Intelligence (NFDI4DS)[11] as part of the non-profit association National Research Data Infrastructure (NFDI e. V.). The consortium is funded by the Federal Republic of Germany and its states through the German Research Foundation (DFG) project NFDI4DS (no. 460234259). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the funders. Neither the European Union nor the granting authority can be held responsible for them.

We thank the anonymous reviewers for their constructive feedback and suggestions, which improved the clarity and quality of this work.

8. Bibliographical References

Tom Ayoola, Shubhi Tyagi, Joseph Fisher, Christos Christodoulopoulos, and Andrea Pierleoni. 2022. Refined: An efficient zero-shot-capable approach to end-to-end entity linking. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track*, pages 209–220.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive entity retrieval.

Dalmeet Singh Chawla. 2017. Unpaywall finds free versions of paywalled papers. *Nature*.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. 2020. Specter: Document-level representation learning using citation-informed transformers.

Vila-Suero Daniel and Aranda Francisco. 2023. Argilla - Open-source framework for data-centric NLP.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Carol Friedman, Pauline Kra, and Andrey Rzhetsky. 2002. Two biomedical sublanguages: a description based on the theories of zellig harris. *Journal of biomedical informatics*, 35(4):222–235.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of ACL*.

Yun He, Ziwei Zhu, Yin Zhang, Qin Chen, and James Caverlee. 2020. Infusing Disease Knowledge into BERT for Health Question Answering, Medical Inference and Disease Name Recognition. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4604–4614, Online. Association for Computational Linguistics.

Ginny Hendricks, Dominika Tkaczyk, Jennifer Lin, and Patricia Feeney. 2020. Crossref: The sustainable source of community-owned scholarly metadata. *Quantitative Science Studies*, 1(1):414–427.

Zhi Hong, Aswathy Ajith, Gregory Pauloski, Eamon Duede, Carl Malamud, Roger Magoulas, Kyle Chard, and Ian T. Foster. 2022. Scholarbert: Bigger is not always better. *ArXiv*, abs/2205.11342.

Zhi Hong, Roselyne Tchoua, Kyle Chard, and Ian Foster. 2020. Sciner: extracting named entities from scientific literature. In *International Conference on Computational Science*, pages 308–321. Springer.

Mohamad Yaser Jaradeh, Allard Oelen, Kheir Eddine Farfar, Manuel Prinz, Jennifer D'Souza, Gábor Kismihók, Markus Stocker, and Sören Auer.

[11] https://www.nfdi4datascience.de

2019. Open research knowledge graph: Next generation infrastructure for semantic scholarly knowledge. In *Proceedings of the 10th international conference on knowledge capture*, pages 243–246.

Jin-Dong Kim, Yue Wang, Toshihisa Takagi, and Akinori Yonezawa. 2011. Overview of genia event task in bionlp shared task 2011. In *Proceedings of BioNLP shared task 2011 workshop*, pages 7–15.

Jeremy Leipzig, Daniel Nüst, Charles Tapley Hoyt, Karthik Ram, and Jane Greenberg. 2021. The role of metadata in reproducible computational research. *Patterns*, 2(9):100322.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Mingchen Li and Rui Zhang. 2023. How far is language model from 100% few-shot named entity recognition in medical domain. *arXiv preprint arXiv:2307.00186*.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238, Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Yiren Liu, Si Chen, Haocong Cheng, Mengxia Yu, Xiao Ran, Andrew Mo, Yiliu Tang, and Yun Huang. 2024. How ai processing delays foster creativity: Exploring research question co-creation with an llm-based agent. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Patrice Lopez. 2009. Grobid: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In *International conference on theory and practice of digital libraries*, pages 473–474. Springer.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium. Association for Computational Linguistics.

Ling Luo, Chih-Hsuan Wei, Po-Ting Lai, Robert Leaman, Qingyu Chen, and Zhiyong Lu. 2023. Aioner: all-in-one scheme-based biomedical named entity recognition using deep learning. *Bioinformatics*, 39(5):btad310.

Paolo Manghi. 2024. Challenges in building scholarly knowledge graphs for research assessment in open science. *Quantitative Science Studies*, 5(4):991–1021.

Paolo Manghi, Alessia Bardi, Claudio Atzori, Miriam Baglioni, Natalia Manola, Jochen Schirrwagen, Pedro Principe, Michele Artini, Amelie Becker, Michele De Bonis, et al. 2019. The openaire research graph data model. *Zenodo*.

Rada Mihalcea and Andras Csomai. 2007. Wikify! linking documents to encyclopedic knowledge. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, page 233–242, New York, NY, USA. Association for Computing Machinery.

Faizan E Mustafa, Corina Dima, Juan Ochoa, and Steffen Staab. 2024. Leveraging Wikidata for biomedical entity linking in a low-resource setting: A case study for German. In *Proceedings of the 6th Clinical Natural Language Processing Workshop*, pages 202–207, Mexico City, Mexico. Association for Computational Linguistics.

Cameron Neylon, Bianca Kramer, Alysson Fernandes Mazoni, Rodrigo Costas, Najko Jahn, and Nees Jan van Eck. 2026. Sharing the load: Building a collective to support open research information online. *Upstream*.

Marc Pàmies, Joan Llop, Francesco Multari, Nicolau Duran-Silva, César Parra-Rojas, Aitor Gonzalez-Agirre, Francesco Alessandro Massucci, and Marta Villegas. 2023. A weakly supervised textual entailment approach to zero-shot text classification. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 286–296.

Silvio Peroni and David Shotton. 2020. Opencitations, an infrastructure organization for open scholarship. *Quantitative Science Studies*, 1(1):428–444.

Jason Priem, Heather A. Piwowar, and Richard Orr. 2022. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *ArXiv*, abs/2205.01833.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Wei Liang Seow, Iti Chaturvedi, Amber Hogarth, Rui Mao, and Erik Cambria. 2025. A review of named entity recognition: from learning methods to modelling paradigms and tasks. *Artificial Intelligence Review*, 58(10):315.

Darya Shlyk, Tudor Groza, Marco Mesiti, Stefano Montanelli, and Emanuele Cavalleri. 2024. REAL: A retrieval-augmented entity linking approach for biomedical concept recognition. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 380–389, Bangkok, Thailand. Association for Computational Linguistics.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. 2022. Scirepeval: A multi-format benchmark for scientific document representations. In *Conference on Empirical Methods in Natural Language Processing*.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Andreas Koukounas, Nan Wang, and Han Xiao. 2024. jina-embeddings-v3: Multilingual embeddings with task lora.

Sina Taslimi, Artemis Capari, Hosein Azarbonyad, Zi Long Zhu, Zubair Afzal, Evangelos Kanoulas, and George Tsatsaronis. 2025. Extracting, detecting, and generating research questions for scientific articles. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8573–8588.

Qwen Team. 2025. Qwen3 technical report.

Thanasis Vergoulis, Serafeim Chatzopoulos, Ilias Kanellos, Panagiotis Deligiannis, Christos Tryfonopoulos, and Theodore Dalamagas. 2019. Bip! finder: Facilitating scientific literature search by exploiting impact-based ranking. In *Proceedings of the 28th ACM international conference on information and knowledge management*, pages 2937–2940.

Feng Wang, Yuqing Li, and Han Xiao. 2025. jina-reranker-v3: Last but not late interaction for document reranking.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Max Wilkinson. 2010. Datacite: The international data citation initiative: Datasets programme.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020a. Scalable zero-shot entity linking with dense entity retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6397–6407, Online. Association for Computational Linguistics.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020b. Scalable zero-shot entity linking with dense entity retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6397–6407, Online. Association for Computational Linguistics.

Mingfang Wu, Felicitas Löffler, Brigitte Mathiak, Fotis Psomopoulos, Uwe Schindler, Amir Aryani, Jordi Bodera Sempere, Antica Culina, Andreas Czerniak, Chris Erdmann, et al. 2026. Bridging the data discovery gap: User-centric recommendations for research data repositories. *Data Science Journal*, 25(6).

Zilin Xiao, Ming Gong, Jie Wu, Xingyao Zhang, Linjun Shou, and Daxin Jiang. 2023. Instructed language models with retrievers are powerful entity linkers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2267–2282, Singapore. Association for Computational Linguistics.

Amy Xin, Yunjia Qi, Zijun Yao, Fangwei Zhu, Kaisheng Zeng, Bin Xu, Lei Hou, and Juanzi Li. 2025. Llmael: Large language models are good context augmenters for entity linking. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*, pages 3550–3559.

Zhenran Xu, Yulin Chen, Baotian Hu, and Min Zhang. 2023. A read-and-select framework for zero-shot entity linking. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13657–13666, Singapore. Association for Computational Linguistics.

Artur Yakimovich, Anaël Beaugnon, Yi Huang, and Elif Ozkirimli. 2021. Labels in a haystack: Approaches beyond supervised learning in biomedical applications. *Patterns*, 2(12):100383.

Anthony Yazdani, Ihor Stepanov, and Douglas Teodoro. 2025. Gliner-biomed: A suite of efficient models for open biomedical named entity recognition. *arXiv preprint arXiv:2504.00676*.

Zheng Yuan, Yijia Liu, Chuanqi Tan, Songfang Huang, and Fei Huang. 2021. Improving biomedical pretrained language models with knowledge. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 180–190, Online. Association for Computational Linguistics.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. GLiNER: Generalist model for named entity recognition using bidirectional transformer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376, Mexico City, Mexico. Association for Computational Linguistics.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025a. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

Yunyi Zhang, Ruozhen Yang, Xueqiang Xu, Rui Li, Jinfeng Xiao, Jiaming Shen, and Jiawei Han. 2025b. Teleclass: Taxonomy enrichment and llm-enhanced hierarchical text classification with minimal supervision. In *Proceedings of the ACM on Web Conference 2025*, pages 2032–2042.

Kang Zhou, Yuepei Li, Qing Wang, Qiao Qiao, and Qi Li. 2024. GenDecider: Integrating "none of the candidates" judgments in zero-shot entity linking re-ranking. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 239–245, Mexico City, Mexico. Association for Computational Linguistics.

Minghao Zhu and Keyuan Jiang. 2021. Semi-supervised language models for identification of personal health experiential from Twitter data: A case for medication effects. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 228–237, Online. Association for Computational Linguistics.

Tiantian Zhu, Yang Qin, Qingcai Chen, Xin Mu, Changlong Yu, and Yang Xiang. 2023. Controllable contrastive generation for multilingual biomedical entity linking. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5742–5753, Singapore. Association for Computational Linguistics.

A. Table of Entities and Target KGs

Table 7 summarises the entities of interest for each of the pilots, along with their corresponding taxonomy of choice for linking and integration within the corresponding SKGs.

B. Pseudo-labelling instructions

General Instructions

```
Extract and categorize these [DOMAIN_NAME]
     entities, focusing on precise entity
     identification.

Entity types:

1. ...
2. ...
3. ...

Output Format

For each document, produce a json output of
     the form

{
    "entities": [
        {
            "entity": the EXACT entity text,
            "label": the entity type
        }
    ]
}

The 'entity' field must contain the exact
     text as it appears in the document (
     maintaining capitalization). When an
     acronym is introduced with its full
     form (e.g., 'electroencephalogram (EEG)
     '), extract the complete text including
      both the full form and the acronym.
     When only the acronym is used
     subsequently (e.g., just 'EEG'),
     extract only the acronym. Never expand
     acronyms that aren't explicitly
     expanded in the text, and never modify
     the entity text in any way from how it
     originally appears.

Only map entities to taxonomy categories
     when there is explicit textual evidence
      supporting that specific
     categorization. Do not make assumptions
      about entities based on common
     associations if not stated in the text.
      Maintain a high standard of evidence---
     if in doubt, do not extract the entity.

Do not ignore documents without extracted
     entities. Provide outputs for all
     documents, even if empty
```

C. Experimental setup

We provide experimental details of our fine-tuned models. For *fine-tuning* NER, GLiNER models, re-

Domain	Entity	Examples	Relevant Taxonomy
Neuroscience	`technique`	DNA sequencing; sodium MRI; voltage clamp	openMINDs
	`preparationType`	in vitro; in vivo	openMINDs
	`UBERONParcellation`	spinal cord; pineal tract; medial orbital frontal cortex	UBERON, restricted to Central Nervous System
	`species`	mice; human; bovine	openMINDS
	`biologicalSex`	male; female	openMINDS
Energy	`energyType`	biogas; photovoltaic; peat	IRENA
	`energyStorage`	pumped hydro storage; thermal storage; ultracapacitors	IRENA
Maritime	`vesselType`	bulk carrier; ferry; yacht	Ad-hoc based on AIS ship types
CCAM	`vehicleType`	car; truck; shuttle	Ad-hoc based on SINFONICA and FAME
	`VRUType`	pedestrian; scooter; bicycle	Ad-hoc based on SINFONICA and FAME
	`communicationType`	5G; cellular communication; DSRC	Ad-hoc based on SINFONICA and FAME
	`entityConnectionType`	vehicle-to-pedestrian; I2X; V2V	Ad-hoc based on SINFONICA and FAME
	`sensorType`	camera; LIDAR; odometer	Ad-hoc based on SINFONICA and FAME
	`scenarioType`	platooning; traffic scenario; test scenario	Ad-hoc based on SINFONICA and FAME
	`levelOfAutomation`	electronic stability control; lane centering; park assist	Ad-hoc based on SINFONICA and FAME

Table 7: List of research communities and their entities of interest, with the correspondent taxonomy of choice.

Hyperparameter	Value
Optimizer	AdamW
Learning rate	$2e^{-5}$
Learning rate schedule	Linear
Warm-up steps	100
Weight decay	0.01
Training batch size	16
Evaluation batch size	16
Maximum epochs	6

Table 8: Hyperparameters used for fine-tuning NER models across all domains.

triever and reranker models, we make use of the `huggingface` library. Training was run (using 1x NVIDIA A100 GPU) for NER models with hyperparameter defined Table 8, GLiNER models in Table 9, E5 retriever model and for Qwen Reranker in Table 10.

Hyperparameter	Value
Training objective	Focal loss
Learning rate	$5e^{-6}$
Learning rate (non-backbone)	$1e^{-5}$
Weight decay	0.01
Batch size	8
Warm-up ratio	0.1
Learning rate scheduler	Linear
Prediction threshold	0.5
Evaluation metric	seqeval F_1

Table 9: Hyperparameters used for fine-tuning GLiNER models across domains.

Hyperparameter	Value
Training objective	MNRLoss
Optimizer	AdamW
Learning rate	$2e^{-5}$
Batch size	32
Epochs	3
Warm-up ratio	0.1
Negative sampling	In-batch negatives
Hard negatives	10 (TF–IDF)
Mixed precision (AMP)	Enabled
Random seed	42

Table 10: Hyperparameters for fine-tuning the E5 model for entity linking retrieval and reranking.

D. Prompt Templates for Context Augmentation

For context augmentation, we employ a few-shot prompting strategy inspired by LLMaEL (Xin et al., 2025), using 2–3 domain-specific examples per domain (3 for maritime transport, 2 for neuroscience, energy, and CCAM), followed by the target mention in context, as described in the fllowing template.

Context Augmentation Template for Entity Linking

```
Example i. Consider the following text from
     a scientific or technical document.
Text: {EXAMPLE_TEXT_WITH_MARKED_ENTITY}
Please provide more descriptive information
     about { {EXAMPLE_MENTION} } from the
     text above.
Make sure to include {EXAMPLE_MENTION} in
     your description.
Answer:
{EXAMPLE_DESCRIPTION}

Now consider the following text from a
     scientific or technical document.
Text: {LEFT_CONTEXT} { {MENTION} } {
     RIGHT_CONTEXT}
Please provide more descriptive information
     about { {MENTION} } from the text above.

Make sure to include {MENTION} in your
     description.
Answer:
```