Crossing Domains without Labels: Distant Supervision for Term Extraction

Elena Senger[1,2] Yuri Campbell[2] Rob van der Goot[3] Barbara Plank[1]

[1]MaiNLP, Center for Information and Language Processing, LMU Munich, Germany
[2]Fraunhofer Center for International Management and Knowledge Economy IMW, Germany
[3]Department of Computer Science, IT University of Copenhagen, Denmark
elena.senger@cis.lmu.de, yuri.campbell@imw.fraunhofer.de
robv@itu.dk, b.plank@lmu.de

Abstract

Automatic Term Extraction (ATE) is a critical component in downstream NLP tasks such as document tagging, ontology construction and patent analysis. Current state-of-the-art methods require expensive human annotation and struggle with domain transfer, limiting their practical deployment. This highlights the need for more robust, scalable solutions and realistic evaluation settings. To address this, we introduce a comprehensive benchmark spanning seven diverse domains, enabling performance evaluation at both the document- and corpus-levels. Furthermore, we propose a robust LLM-based model that outperforms both supervised cross-domain encoder models and few-shot learning baselines and performs competitively with its GPT-4o teacher on this benchmark. The first step of our approach is generating pseudo-labels with this black-box LLM on general and scientific domains to ensure generalizability. Building on this data, we fine-tune the first LLMs for ATE. To further enhance document-level consistency, oftentimes needed for downstream tasks, we introduce lightweight post-hoc heuristics. Our approach exceeds previous approaches on 5/7 domains with an average improvement of 10 percentage points. We release our dataset and fine-tuned models to support future research in this area.[1]

1 Introduction

Automatic Term Extraction (ATE) is a crucial component of many NLP systems, with applications in information retrieval, machine translation, topic detection, and sentiment analysis (Tran et al., 2023; Xu et al., 2025). Traditional rule-based or frequency-based ATE systems, as well as state-of-the-art (SOTA) methods with pretrained models, rely heavily on fine-tuning with human-annotated datasets, which are typically available for only a handful of domains. Recent surveys explicitly highlight this dependence as a key limitation, noting that multi-domain ATE scenarios remain an unsolved challenge for current SOTA approaches (Tran et al., 2023; Xu et al., 2025). Large language models (LLMs), with their massive pretraining across diverse corpora, offer a promising path toward generalizable ATE. Yet, early applications of LLMs in term extraction remain limited and typically perform worse compared to supervised SOTA methods (Tran et al., 2023). Moreover, proprietary black-box LLMs incur high API costs and pose privacy risks when handling sensitive or confidential data.

To address these limitations, we introduce a novel ATE framework: DiSTER (**Di**stant **S**upervision for **T**erm **E**xtraction with **R**obustness), that leverages LLMs with distant supervision. Our approach trains smaller, open models using synthetic data generated via pseudo-labels from a black-box LLM, thereby removing the need for human annotation and enabling cross-domain scalability. To enhance consistency within and across documents, we incorporate simple post-hoc consistency heuristics. These heuristics significantly improve F1 scores and oftentimes lead to more balanced precision and recall.

Moreover, we perform a comprehensive empirical study, spanning the seven following domains: biomedicine, corruption, dressage, heart failure, coastal geography, computational linguistics and wind energy. Combining these established datasets makes this the largest and most diverse multi-domain ATE evaluation to date. We assess models under both corpus-level and document-level setups to better reflect real-world extraction needs. Our results demonstrate that training on distantly supervised data leads to notable improvements in cross-domain robustness and that post-hoc consistency enforcement yields further gains, boosting

[1] Dataset: https://huggingface.co/datasets/ElenaSenger/SynTerm; Model: https://huggingface.co/ElenaSenger/DiSTER-Llama-3-8B-Instruct

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: Industry Track, pages 1366–1378
November 4-9, 2025 ©2025 Association for Computational Linguistics

document-level F1 scores by up to 55 percentage points. Our contributions are:

- We propose DiSTER, a novel distantly supervised ATE framework that combines synthetic data generation, LLM fine-tuning, and lightweight post-hoc consistency heuristics for robust and scalable term extraction without human annotation.
- We demonstrate that a strategically constructed, domain-diverse synthetic training corpus significantly enhances cross-domain generalization.
- We conduct the most comprehensive cross-domain ATE evaluation to date, spanning seven diverse domains and evaluating both corpus-level and document-level performance.

2 Related work

2.1 Automatic Term Extraction

Traditional ATE methods typically follow a two-step pipeline: (1) extracting candidate terms using linguistic and statistical features, and (2) ranking them based on termhood and unithood scores (Xu et al., 2025). Supervised machine learning approaches enhance this process using manually designed features and classifiers like SVMs, Random Forests, or CRFs (Tran et al., 2023). With deep learning, models such as BiLSTMs, CNNs, and Transformers have been adopted for token classification and embedding-based approaches, showing SOTA results across languages and domains (Tran et al., 2023, 2022a,b). Recently, LLMs have entered the field. Giguere (2023) showed that GPT-4 (OpenAI et al., 2024b) performs well in zero-shot settings across legal, technical, and medical domains, outperforming statistical baselines on small test sets. Meanwhile, Tran et al. (2024) explored few-shot prompting with `LLaMA` and `GPT-3.5-Turbo` for the ACTER heart-failure dataset, though results still lag behind cross-domain sequence labeling with XLM-R (Conneau et al., 2020).

2.2 Distillation and Pseudo-Labeling

Knowledge Distillation (KD) transfers knowledge from larger teacher models to smaller student models. In black-box KD, often used with proprietary LLMs, only the teacher's outputs are available (Yang et al., 2024), hence, the student model learns by mimicking the teacher's generated sequences instead of its internal states. Specifically, *Labeling Knowledge*, when a LLM labels a set of examples based on an instruction (with or without demonstrations), is widely considered effective for transferring specific LLM skills (Xu et al., 2024). This approach has proven effective across diverse NLP tasks (Li et al., 2025). In the information extraction domain, several studies have shown promising results: Hsu and Roberts (2024) applied LLM weak supervision improving performance on medical entity extraction with minimal human annotation. Similarly, UniversalNER (Zhou et al., 2024) successfully distilled open-domain named-entity recognition (NER) capabilities from `GPT-3.5-Turbo-0301` into smaller models that ultimately outperformed their teacher. For more complex extraction tasks, MetaIE (Peng et al., 2024) employed distillation as a meta-learning framework to create task-flexible information extraction systems capable of adapting to various relation and entity types.

3 DiSTER

Our approach DiSTER first creates a distantly supervised dataset, then fine-tunes an LLM on that data to generate candidate terms, and lastly selectively applies post-hoc heuristics. This pipeline is illustrated in Figure 1. We cover the details of each component in the following sections.

3.1 Dataset Creation

To create our synthetic dataset *SynTerm* for model fine-tuning, we used the dataset of Zhou et al. (2024) as a basis. Their synthetic NER dataset was generated by prompting `gpt-3.5-turbo-0301` to identify named entities within text snippets taken from *The Pile* (Gao et al., 2020). We utilize their annotations as a starting point, but apply a filtering step (as described next) to focus only on relevant entity types and added labeled data from *arXiv* for broader domain coverage.

Entity Type Filtering In order to filter, two authors manually annotated the 130 most common entity types covering approximately 74% of all extracted entities, labeling them as either *terms* or *non-terms* (e.g., person names, locations, etc.) following the ACTER annotation guidelines (Rigouts Terryn, Ayla, 2021). This manual annotation process resulted in a Cohen's kappa coefficient of 0.723, indicating substantial inter-annotator agree-

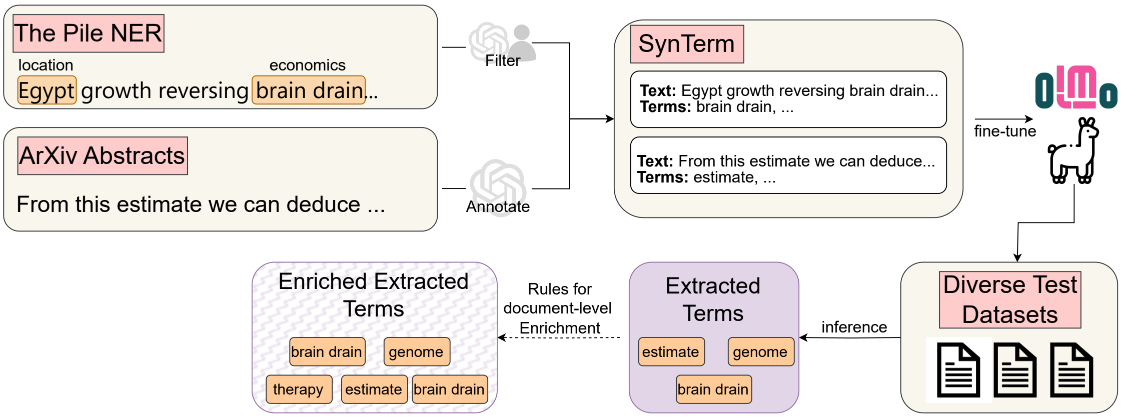

Figure 1: An overview of the key components of our DiSTER approach.

ment. These human-provided labels were used for all entities within the 130 most frequent types. For the remaining 26% of entities, which belong to less common types, we relied on GPT-4o (OpenAI et al., 2024a) to classify whether phrases of a given entity type represent terms (see Appendix C). To assess consistency, we compared the model's labels with the human annotations on the 130 most common types, which yielded a lower Cohen's kappa of 0.45. Notably, the LLM tended to label fewer entity types as terms than the annotators (109 vs. 127 and 136, respectively). For instance, GPT-4o labeled entity types such as 'event,' 'nationality,' falsely as terms and instances of 'medical conditions' as non-terms, whereas the annotators considered them valid terms. Overall, the potential noise introduced by the LLM is expected to be limited, as most extracted entities fall within the 130 most common types and were annotated manually. Manual annotation of all entity types would have been prohibitively expensive, as the NER dataset contains a total of 13,020 entity types. By focusing on the most common types where noise reduction matters most, we ensured high-quality filtering for the majority of data while using GPT-4o only for the long-tail of rare entity types.

Human:
Text: From this estimate we can deduce that the spatially homogeneous Boltzmann equation is well posed in a class of measure-valued processes. We also prove in an appendix a basic lemma on the total variation of time-integrals of time-dependent signed measures.

GPT:
I've read this text.

Human:
Please extract the terms from the text that are relevant to the domain: Probability, Analysis of PDEs.

GPT:
["spatially homogeneous Boltzmann equation", "measure-valued processes", "total variation", "time-integrals", "time-dependent signed measures"]

Figure 2: Conversation example showing extraction of domain-specific terms from an *arXiv* text. For data points without specific domain, like the ones coming from the Pile, we substitute the domain by "General".

Domain-Aware Data Augmentation To increase domain diversity, we synthesized labeled examples from two-sentence snippets from *arXiv* abstracts using GPT-4o. The two-sentence snippets were chosen to closely match the length of data points from both *The Pile* and the test datasets. As in the few-shot setup of Tran et al. (2024), we included domain information in the extraction prompt (see Appendix D). The domain information was derived from the *arXiv* categories associated with each abstract. Given the domain-dependent nature of termhood (Xu et al., 2025), we expect this to improve the relevance of extracted terms. We also incorporated domain labels into the final data points used for fine-tuning, an example is shown in Figure 2. In contrast, for data points based on *The Pile*, where no domain labels are available and thus not integrated in the data points, we prompted the model during training to extract terms without specifying a domain.

In summary, we created two data subsets: *TermPile* (45,432 instances), the filtered NER dataset based on *The Pile* and, *TermArXiv* (37,829 instances) the newly synthesized *arXiv* data points. Combining the two yields our final *SynTerm* dataset.

Domain	Example Sentence
corp (Corruption)	The first criterion creates a link between the **offence** and the **legal person**.
equi (Dressage / Equity)	They might go from a **lengthened stride** and **half halt** back to a **working trot**.
htfl (Heart Failure)	**Heart failure** risk among **patients** with **rheumatoid arthritis** starting a **TNF antagonist**.
wind (Wind Energy)	**Wind turbine technology** has developed rapidly in recent years and Europe is at the hub of this hightech **industry**.
coast (Coastal Science)	**Coastal communities** are prone to a **natural disaster** such as **tsunami**.
genia (Biomedical)	**HB24** is likely to have an important role in **lymphocytes** as well as in certain developing tissues.
acl (Computational Linguistics)	**Word Identification** has been an important and active issue in **Chinese Natural Language Processing**.

Table 1: Example sentences from each of the seven domains used in our experiments. Terms are bold.

3.2 LLM Fine-Tuning

After constructing the dataset for fine-tuning we perform standard instruction tuning in order to transfer the ATE skill to a smaller model. Precisely, we fine-tune two smaller instruction tuned models, Llama-3-8B-Instruct (LLaMA) and Olmo-7B-Instruct (Olmo), with standard next-token prediction objective and conversation-style chat templates.[2] In both cases, we use only completions in order to compute the loss, that is, only the tokens generated by the language model after the last "Assistant"-marker. The models were trained for 3 epochs, with learning rate of $2e-4$ and batch size of 8. Only the final checkpoints were taken for further analysis.

3.3 Post-hoc Consistency Enforcement

To address the inconsistent extraction behavior of LLMs we introduce two lightweight post-hoc heuristics for enforcing consistency. The first, document-level consistency (DC) enforcement, aims to correct the LLM's tendency to return only one instance of each extracted term per document, even when multiple mentions occur. To remedy this, we identify all exact string matches of each LLM-extracted term within the document. This approach is conceptually aligned with prior work in NER that enforces intra-document label agreement for repeated spans (Krishnan and Manning, 2006; Gui et al., 2020).

The second rule, corpus-level consistency (CC) enforcement, promotes any term extracted in at least 50% of the documents it appears in to all such documents—addressing inconsistencies in LLM output. This simple heuristic mirrors the Term Re-extraction Model (TREM) by Vu et al. (2008), which reintroduced globally validated terms into individual documents, and aligns with frequency-based termhood estimation methods (Kageura and Umino, 1996). Such reinforcement mechanisms help align model outputs more closely with span-level gold annotations and improve both recall and consistency in document-level evaluations.

4 SOTA Approaches

To better evaluate DiSTER, we compare it against two strong baseline approaches that represent the current SOTA methodologies in ATE.

4.1 Sequence Labeling Approach

We adopt the approach introduced by Tran et al. (2022a), which frames ATE as a sequence classification task using the IOB tagging scheme. This approach employs an XLM-R-based token classifier with standard hyper-parameters and has been shown to achieve SOTA in term extraction. It remains a strong baseline, as even recent few-shot LLM-based methods could not consistently outperform it (Tran et al., 2024).

4.2 Few-Shot Approach

We re-implement the few-shot in-context learning approach of Tran et al. (2024), but use cross-domain few-shot samples instead of samples from the target test data. Following Tran et al. (2024), each prompt contains three examples. For every target test domain, we select examples showing the highest semantic similarity to the domain name, as measured by the paraphrase-multilingual-mpnet-base-v2 embedding model.[3] Each example is structured as a demonstration pairing a source sentence with its corresponding extracted terms. In alignment with Tran et al. (2024), we employ Direct Term Extraction rather than IOB tagging, the LLM explicitly outputs the identified terms. We enriched the prompt templates for extraction as introduced in Tran et al. (2024), with complete specifications of these templates provided in Appendix B.

[2] LLaMA model: https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct, Olmo model: https://huggingface.co/allenai/OLMo-7B-Instruct

[3] https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2

Model	***corp* F1**	***equi* F1**	***htfl* F1**	***wind* F1**	***coast* F1**	***genia* F1**	***acl* F1**	**Avg F1**
IOB sequence labeling (cross-domain)	31.35	41.64	14.60	**44.26**	14.14	18.20	10.55	24.96
IOB sequence labeling *SynTerm*	29.89	32.52	42.74	31.17	56.49	42.33	56.78	41.70
`LLaMA` few-shot (cross domain)	10.47	41.74	**51.29**	33.39	42.71	45.33	40.93	37.98
`LLaMA` few-shot *SynTerm*	32.64	43.53	41.71	38.47	36.67	46.99	51.62	41.94
`Olmo` few-shot (cross domain)	27.09	37.13	49.87	34.54	49.00	46.40	31.23	39.32
`LLaMA` fine-tuned *SynTerm* (DiSTER)	**37.93**	**45.54**	50.10	42.93	**66.96**	**51.80**	**65.37**	**51.52**
`Olmo` fine-tuned *SynTerm* (DiSTER)	32.03	34.37	45.10	35.28	57.48	44.29	57.73	43.75
Teacher Model zero-shot	38.18	50.96	51.54	45.83	62.58	50.19	62.66	51.70
Teacher Model few-shot (cross domain)	41.19	40.66	53.57	36.40	62.89	48.21	60.82	49.10

Table 2: Corpus-level F1 scores across seven datasets. Best result per dataset is marked in bold, second best results are underlined. Teacher model performance is included for comparison.

5 Benchmark Datasets

We make use of several established benchmarks, resulting in seven datasets. They are topically quite distant, which leads to substantially different types of terms, often including domain-specific jargon (see Table 1). All datasets have predefined splits, except for the ACTER datasets, which were introduced in a cross-domain setting. While most original studies report in-domain performance, we only use cross-domain test splits in this work.

We employ the ACTER dataset introduced by the TermEval 2020 Shared Task (Rigouts Terryn et al., 2020), which includes four subsets with each a different domain: heart failure (*htfl*), wind energy (*wind*), dressage (equity) (*equi*), and corruption (*corp*). Secondly, we use the CoastTerm (*coast*) dataset (Delaunay et al., 2024) consisting of scientific abstracts focused on coastal regions. Due to the inherently interdisciplinary nature of coastal studies, the texts include a wide range of specialized terms spanning domains such as environmental science, geography, ecology, and sociology. We also incorporate the *genia* dataset, a standard benchmark for biomedical term extraction (Kim et al., 2011). Finally, we use the ACL-RD-TEC 2.0 *acl* dataset, which contains abstracts from the ACL Anthology from the domain of *computational linguistics* (QasemiZadeh and Schumann, 2016).

6 Experimental Setup

6.1 Evaluation Strategies

We employ both *corpus-level* and *document-level* evaluation. The *corpus-level* approach aggregates predictions and gold annotations across the entire dataset before computing metrics, while the *document-level* strategy calculates metrics per document and then averages results.

6.2 Model Configurations

We evaluate three distinct model categories. First, for our method, DiSTER, which relies on **fine-tuned models**, we use two instruction tuned LLMs on our synthetic data: Llama-3-8B-Instruct (`LLaMA`) and Olmo-7B-Instruct (`Olmo`). Regarding the **few-shot prompted models**, we evaluate the same LLM architectures (`LLaMA` and `Olmo`) in a few-shot setting, using cross-domain demonstrations from the remaining datasets. For each target domain, we construct prompts with semantically similar examples from other domains. Additionally, we evaluate the better-performing model `LLaMA` using demonstrations from our dataset *SynTerm*. For the **IOB sequence labeling approaches**, we implement two training configurations. Firstly, we implement a leave-one-out approach where for each test domain, we train on five domains and validate on the sixth. We consistently use *wind* for validation (given its STEM domain alignment) except when testing on *wind* itself, where we validate on *htfl*. In our second configuration, we train on our *SynTerm* dataset to enable direct comparison with the fine-tuned LLMs (the full DiSTER methodology).

This experimental design allows us to systematically evaluate the impact of model architecture, training methodology, and data composition on cross-domain generalization in ATE.

7 Results

Table 2 presents the corpus-level F1 scores. Notably, our models achieve the highest F1 scores in most domains, surpassing both sequence-labeling and few-shot prompting methods. The fine-tuned `LLaMA` model reaches the best overall performance. `Olmo` also performs consistently well across domains, demonstrating the effectiveness of our approach even for open-data models. While few-shot prompted models show competitive perfor-

Val/Test	*corp* F1	*equi* F1	*htfl* F1	*wind* F1
corp	–	45.32	40.98	33.74
equi	6.10	–	46.68	30.78
htfl	6.40	51.80*	–	42.57
wind	7.89	31.57	31.68	–

Table 3: F1 scores across ACTER domains using a leave-one-out setup: one domain for testing, one for validation, and the remaining two for training. The *htfl* score is reproduced from Tran et al. (2022a) under the original Shared Task setting.

mance in specific domains (e.g., LLaMA on *htfl*), their performance remains fundamentally inconsistent, with significant variability across different domains such as *corp*. Using our *SynTerm* dataset for demonstrations yields a four percentage point average improvement compared to cross-domain few-shot with LLaMA. Notably, the performance gains for *corp* and *wind*, the two domains with the lowest cross-domain few-shot scores, suggest that *SynTerm* helps achieve more robust overall few-shot performance. The supervised cross-domain IOB models reveal inherent limitations, particularly struggling with recall under domain shift. Even when trained on our *SynTerm* dataset, these models consistently underperform compared to our fine-tuned LLM-based approaches. Remarkably, Table 2 also shows that the fine-tuned LLaMA student surpasses the teacher in corpus-level F1 on three of seven test corpora and stays on-par on the macro average. These findings provide strong support for the generalizability and use of DiSTER as a pivotal strategy for the development of more domain-robust and light-weight ATE systems.

8 Analysis

In this section, we begin by comparing the student to its black-box teacher. Then we inspect the error sources causing the few-shot and IOB model failure. Finally, we present a discussion about the distinctions and use cases of document-level evaluation and the impact of our post-hoc heuristics.

8.1 The Student Rivals Its Teacher

Table 2 shows that DiSTER effectively distills ATE capabilities from the teacher into the much lighter LLaMA student. The student outperforms the teacher on three of seven domains and comes within 1.5 F1 points on two more, demonstrating strong competitive performance despite its smaller size. The largest gains appear in domains semantically aligned with the synthetic corpus (*acl*, *coast*, *genia*). Where overlap is weaker, the student tends to lag further behind, suggesting that broader domain coverage in the pseudo-labeled training data could close the remaining gaps. See Appendix §J for an analysis of training data overlap.

We hypothesize that distillation works due to two complementary factors: (i) sequence-level supervision encourages the student to mimic the teacher's span predictions exactly, reducing prompt sensitivity and reinforcing task-specific patterns; and (ii) domain cues in the synthetic data act as scaffolding: smaller models benefit from domain-specific regularities, aiding generalization in overlapping domains. This suggests that strong cross-domain performance still depends on diverse fine-tuning data, making automated labeling approaches like DiSTER a cost-effective path to improving ATE generalizability in smaller, deployable models.

8.2 Where IOB and Few-Shot Fail

Two systematic error sources (*extraction count* and *term-span length*) explain much of the underperformance observed in the baseline systems, as thoroughly discussed in Appendix H. In brief, in the few-shot setting, models often return few or no candidates. On the *corp* subset, the median number of predicted terms is zero. Even when terms are extracted, their median length far exceeds the gold standard, which severely depresses recall (Table 9).

The supervised IOB model exhibits the converse pathology. When trained on *SynTerm*, the model assigns the term label to overly long spans that often include stop words, thereby inflating recall while harming precision. DiSTER's LLM fine-tuning better addresses this challenge by producing extraction counts and term spans that are closer to the gold standard. Therefore, we posit that the underlying limitation is architectural. These findings are consistent with the instability patterns observed under cross-domain training (Table 3; see Appendix K for a detailed analysis). Even within the original four ACTER domains, F1 scores can drop by over 30 points depending on the validation split. Together, these analyses highlight the need for a more flexible architecture and targeted fine-tuning to best leverage the distillation data *SynTerm* provides.

8.3 Document-Level Evaluation

While corpus-level ATE has been the primary focus in prior evaluations, document-level evaluation better reflects downstream tasks like computer-assisted translation and information retrieval (Šajatović et al., 2019). As shown in Table 4, document-

Model	***corp* F1**	***equi* F1**	***htfl* F1**	***wind* F1**	***coast* F1**	***genia* F1**	***acl* F1**	**Avg F1**
IOB sequence labeling (cross-domain)	25.66	46.91	8.01	**49.90**	11.41	13.49	8.72	23.44
IOB sequence labeling *SynTerm*	21.65	33.81	39.63	26.52	42.11	28.27	39.37	33.05
LLaMA few-shot (cross domain)	5.82	**58.50**	43.65	37.49	35.41	42.88	32.05	36.54
LLaMA few-shot *SynTerm*	29.22	38.57	25.23	41.25	30.23	45.77	42.20	36.07
Olmo few-shot (cross domain)	14.52	48.41	47.28	37.94	44.14	44.44	23.28	37.14
LLaMA fine-tuned *SynTerm*	**38.25**	51.60	**49.03**	49.87	**63.19**	**51.10**	**56.31**	**51.34**
Olmo fine-tuned *SynTerm*	35.63	43.77	47.51	39.12	55.60	38.55	50.07	44.32

Table 4: Document-level F1 scores across seven datasets. Best result per dataset is marked in bold, second best results are underlined.

Model	*corp*		*equi*		*htfl*		*wind*		*coast*		*genia*		*acl*		**Mean F1**	**Mean |R-P|**
	F1	|R-P|	F1	|R-P|	F1	|R-P|	F1	|R-P|	F1	|R-P|	F1	|R-P|	F1	|R-P|	(%)	(%)
LLaMA fine-tuned	38.25	7.74	51.60	15.33	49.03	10.27	49.87	4.11	63.19	19.92	51.10	18.55	56.31	22.53	51.91	13.63
LLaMA fine-tuned DC	39.87	10.78	54.27	12.77	51.36	8.21	53.06	9.01	64.25	18.73	52.87	21.88	**65.14**	12.72	54.40	13.73
LLaMA fine-tuned CC	40.05	14.89	54.79	5.68	52.01	1.92	51.30	11.89	63.74	12.47	51.76	27.51	55.98	17.84	52.80	13.74
LLaMA fine-tuned DC + CC	**41.25**	19.32	**57.14**	1.82	**54.22**	0.80	**54.55**	18.68	**64.53**	10.72	**53.74**	32.00	63.98	5.40	**55.92**	12.68

Table 5: F1 scores (bolded for highest score per dataset), absolute precision-recall gaps, and mean F1 score across datasets. Showing the influence of the document consistency (DC) and corpus consistency (CC) heuristics.

level F1 scores are generally lower, reflecting the greater challenge of consistently extracting terms within individual documents. In particular, recall is significantly lower and, in most cases, falls below precision, reversing the trend observed at the corpus-level (see Appendix E). At the corpus level, terms from all documents are pooled, so consistency within individual documents or term repetition matters less.

As shown in Table 5, applying consistency enforcement heuristics improves document-level F1 scores. Especially, the *acl* dataset, with many term repetitions per document, sees a nine-point F1 gain. Since the heuristics target different patterns, within-document (DC) and cross-corpus (CC) term repetition, their effects are complementary. When combined, they yield the highest F1 scores in six of seven evaluated datasets. Additionally, when precision exceeds recall, these heuristics narrow the precision-recall gap, resulting in a more balanced performance.

9 Conclusion

We introduced DiSTER, a scalable and robust ATE framework combining synthetic data generation, LLM fine-tuning, and post-hoc consistency heuristics. By using pseudo-labels from a black-box LLM, we built the diverse *SynTerm* corpus to support cross-domain generalization. The fine-tuned LLMs, especially LLaMA, outperform both supervised sequence labeling and few-shot prompting and perform competitively with the GPT-4o teacher model, despite the size gap. Our results highlight the importance of data composition in cross-domain ATE and show that our approach generalizes well even to less related domains. We also show that document-level evaluation reveals important limitations in consistency, which can be effectively addressed using simple heuristics. Crucially, DiSTER eliminates the need for domain-specific training, making ATE more scalable and practical.

Limitations

Our approach, while effective, has several limitations. First, the pseudo-labels used for training are derived from a black-box LLM and may contain noise, especially for rare or ambiguous terms. Second, while our models perform well even in domains with limited overlap in the training data, generalization to entirely unseen or underrepresented domains cannot be guaranteed. Expanding the diversity of synthetic data could further strengthen cross-domain robustness. Third, the post-hoc consistency heuristics are simple heuristics and do not handle paraphrases or semantic variants, which could limit precision. Fourth, while we use relatively lightweight LLMs, fine-tuning still demands substantial computational resources, potentially limiting accessibility for low-resource settings. Lastly, our experiments are conducted only in English, and computational cost of fine-tuning open-weight models may hinder adoption in low-resource settings.

References

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Julien Delaunay, Hanh Thi Hong Tran, Carlos-Emiliano González-Gallardo, Georgeta Bordea, Mathilde Ducos, Nicolas Sidere, Antoine Doucet, Senja Pollak, and Olivier De Viron. 2024. Coastterm: a corpus for multidisciplinary term extraction in coastal scientific literature. *Preprint*, arXiv:2406.09128.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Julie Giguere. 2023. Leveraging large language models to extract terminology. In *Proceedings of the First Workshop on NLP Tools and Resources for Translation and Interpreting Applications*, pages 57–60, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Tao Gui, Jiacheng Ye, Qi Zhang, Yaqian Zhou, Yeyun Gong, and Xuanjing Huang. 2020. Leveraging document-level label consistency for named entity recognition. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 3976–3982. International Joint Conferences on Artificial Intelligence Organization. Main track.

Enshuo Hsu and Kirk Roberts. 2024. Leveraging large language models for knowledge-free weak supervision in clinical natural language processing. *Scientific Reports*, 15.

Kyo Kageura and Bin Umino. 1996. Methods of automatic term recognition — a review —. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 3.

Gauri Kambhatla, Thuy Nguyen, and Eunsol Choi. 2023. Quantifying train-evaluation overlap with nearest neighbors. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 2905–2920, Toronto, Canada. Association for Computational Linguistics.

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, Portland, Oregon, USA. Association for Computational Linguistics.

Vijay Krishnan and Christopher D. Manning. 2006. An effective two-stage model for exploiting non-local dependencies in named entity recognition. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 1121–1128, Sydney, Australia. Association for Computational Linguistics.

Juanhui Li, Sreyashi Nag, Hui Liu, Xianfeng Tang, Sheikh Sarwar, Limeng Cui, Hansu Gu, Suhang Wang, Qi He, and Jiliang Tang. 2025. Learning with less: Knowledge distillation from large language models via unlabeled data. *Preprint*, arXiv:2411.08028.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024a. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024b. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Letian Peng, Zilong Wang, Feng Yao, Zihan Wang, and Jingbo Shang. 2024. Metaie: Distilling a meta model from llm for all kinds of information extraction tasks. *Preprint*, arXiv:2404.00457.

Behrang QasemiZadeh and Anne-Kathrin Schumann. 2016. The ACL RD-TEC 2.0: A language resource for evaluating term extraction and entity recognition methods. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1862–1868, Portorož, Slovenia. European Language Resources Association (ELRA).

Ayla Rigouts Terryn, Veronique Hoste, Patrick Drouin, and Els Lefever. 2020. TermEval 2020: Shared task on automatic term extraction using the annotated corpora for term extraction research (ACTER) dataset. In *Proceedings of the 6th International Workshop on Computational Terminology*, pages 85–94, Marseille, France. European Language Resources Association.

Rigouts Terryn, Ayla. 2021. ACTER terminology annotation guidelines.

Antonio Šajatović, Maja Buljan, Jan Šnajder, and Bojana Dalbelo Bašić. 2019. Evaluating automatic term extraction methods on individual documents. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, pages 149–154, Florence, Italy. Association for Computational Linguistics.

Hanh Thi Hong Tran, Carlos-Emiliano González-Gallardo, Julien Delaunay, Antoine Doucet, and Senja Pollak. 2024. Is prompting what term extraction needs? In *Text, Speech, and Dialogue: 27th International Conference, TSD 2024, Brno, Czech Republic, September 9–13, 2024, Proceedings, Part I*, page 17–29, Berlin, Heidelberg. Springer-Verlag.

Hanh Thi Hong Tran, Matej Martinc, Jaya Caporusso, Antoine Doucet, and Senja Pollak. 2023. The recent advances in automatic term extraction: A survey. *Preprint*, arXiv:2301.06767.

Hanh Thi Hong Tran, Matej Martinc, Antoine Doucet, and Senja Pollak. 2022a. Can cross-domain term extraction benefit from cross-lingual transfer? In *Discovery Science: 25th International Conference, DS 2022, Montpellier, France, October 10–12, 2022, Proceedings*, page 363–378, Berlin, Heidelberg. Springer-Verlag.

Hanh Thi Hong Tran, Matej Martinc, Andraz Pelicon, Antoine Doucet, and Senja Pollak. 2022b. Ensembling transformers for cross-domain automatic term extraction. In *From Born-Physical to Born-Virtual: Augmenting Intelligence in Digital Libraries: 24th International Conference on Asian Digital Libraries, ICADL 2022, Hanoi, Vietnam, November 30 – December 2, 2022, Proceedings*, page 90–100, Berlin, Heidelberg. Springer-Verlag.

Thuy Vu, Ai Ti Aw, and Min Zhang. 2008. Term extraction through unithood and termhood unification. In *Proceedings of the Third International Joint Conference on Natural Language Processing: Volume-II*.

Kang Xu, Yifan Feng, Qiandi Li, Zhenjiang Dong, and Jianxiang Wei. 2025. Survey on terminology extraction from texts. *Journal of Big Data*, 12(1):29.

Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. 2024. A survey on knowledge distillation of large language models. *ArXiv*, abs/2402.13116.

Chuanpeng Yang, Wang Lu, Yao Zhu, Yidong Wang, Qian Chen, Chenlong Gao, Bingjie Yan, and Yiqiang Chen. 2024. Survey on knowledge distillation for large language models: Methods, evaluation, and application. *ArXiv*, abs/2407.01885.

Wenxuan Zhou, Sheng Zhang, Yu Gu, Muhao Chen, and Hoifung Poon. 2024. Universalner: Targeted distillation from large language models for open named entity recognition. *Preprint*, arXiv:2308.03279.

A Use Of AI Assistants

The authors acknowledge the use of ChatGPT and Claude solely for correcting grammatical and spelling errors, and providing assistance with coding.

B Few-shot prompt template

Extending the few-shot prompt templates introduced by (Tran et al., 2024), we design two modifications:

- **Context Enrichment (CE)**: At the beginning of the user prompt, we add the ISO definition of a *term — Terms are "the designation of a defined concept in a special language by a linguistic expression." (ISO 1087). A term is a word or a phrase that has a specific meaning in a particular context/domain, such as a scientific term or a technical concept.*

- **Assistance Response Guidance (ARG)**: Instead of freely letting the language model begin the assistant's response, we prepend to the to the response the sentence: *I have extracted the terms from the text. Here is the list of terms:*, and let the LLM complete it.

While CE aims to provide the LLM with more knowledge about the task at hands, ARG deals with the inherent structure-free and stochastic nature of LLMs' generations, in order to facilitate parsing.

C Prompt for Entity Type Labeling

Prompts 1 and 2 show the system and user prompts used for entity-type-label generation, respectively.

System Prompt
You are a terminology research expert. Your task is to help the user by answering the following question.

Prompt 1: System prompt used for entity-type-label generation.

D Prompts for Data Synthetization

Prompts 3 and 4 show the system and user prompts used for pseudo-label generation, respectively.

E Precision and Recall per Domain and Model

The Table 6 shows the corpus-level precision and recall per dataset and model, while Table 7 shows the document-level scores.

Model	acter_corp		acter_equi		acter_htfl		acter_wind		coast		genia		acl	
	P	R	P	R	P	R	P	R	P	R	P	R	P	R
IOB sequence labeling *SynTerm*	19.2	67.52	21.83	63.71	30.4	69.86	19.32	80.69	49.04	66.5	30.49	69.17	47.75	70.02
IOB sequence labeling (cross-domain)	43.5	24.5	39.33	44.23	36.27	9.14	33.08	66.85	56.02	8.09	20.12	16.61	68.42	5.72
LLaMA few-shot (cross domain)	10.39	10.53	28.85	74.83	47.52	55.68	21.85	70.62	68.17	31.16	36.49	59.81	68.5	29.24
LLaMA few-shot *SynTerm*	23.09	55.65	35.05	57.44	47.48	37.19	26.64	69.20	72.55	24.54	37.13	63.98	71.20	40.48
Olmo few-shot (cross domain)	28.89	25.5	24.74	74.41	43.44	17.17	23.26	67.14	65.05	39.31	41.51	1.7	67.7	20.29
LLaMA fine-tuned *SynTerm*	26.34	67.74	33.81	69.73	39.89	67.34	29.82	76.55	73.87	61.24	40.86	70.75	71.66	60.1
Olmo fine-tuned *SynTerm*	22.08	58.31	25.81	51.42	35.34	62.31	23.44	71.28	61.71	53.8	34.63	61.4	65.21	51.78

Table 6: Corpus-level evaluation: Precision (P) and Recall (R) scores for each dataset across models.

Model	acter_corp		acter_equi		acter_htfl		acter_wind		coast		genia		acl	
	P	R	P	R	P	R	P	R	P	R	P	R	P	R
IOB sequence labeling *SynTerm*	13.08	62.73	23.15	62.65	28.83	63.38	16.42	68.77	30.81	66.5	18.23	62.88	27.05	72.28
IOB sequence labeling (cross-domain)	45.92	17.8	67.51	35.95	32.67	4.56	46.83	53.98	45.38	6.53	12.37	14.84	57	0.05
LLaMA few-shot (cross domain)	14.62	3.64	57.58	59.44	62.27	33.6	34.15	41.54	61.66	24.84	39.45	46.95	63.62	21.42
LLaMA few-shot *SynTerm*	29.20	29.23	57.91	28.92	56.08	16.27	40.92	41.58	69.78	19.29	39.27	54.85	66.84	30.84
Olmo few-shot (cross domain)	28.93	9.69	49.49	47.38	56.42	40.69	36.88	39.07	61.01	34.58	38.36	52.82	62.35	14.32
LLaMA fine-tuned *SynTerm*	34.77	42.51	60.38	45.05	54.7	44.43	47.9	52.01	74.68	54.76	43.46	62.01	69.75	47.22
Olmo fine-tuned *SynTerm*	28.15	48.54	49.13	39.46	47.4	47.62	32.03	50.25	60.44	51.48	28.54	59.37	60.66	42.63

Table 7: Document-level evaluation: Precision (P) and Recall (R) scores for each dataset across models.

F Qualitative Examples

Table 8 presents a qualitative comparison of different models on the same input sentence. The extracted terms are highlighted in bold. This example underscores the contrast in recall capabilities between various approaches.

In particular, the IOB sequence labeling model trained on *SynTerm* demonstrates high recall, but low precision. In contrast, the few-shot approaches tend to miss key terms, reflecting their comparatively lower recall. Fine-tuned models perform significantly better than their few-shot counterparts, aligning more closely with the gold terms.

G Unique terms among datasets

Figure 3 shows the amount of unique terms per dataset on the diagonal and the amount of unique common terms among the datasets on the lower triangular part. We observe that, as expected, *SynTerm* has the highest overlap with its source datasets, *TermPile* and *TermArXiv*. Furthermore, *SynTerm* shows considerable overlap with all other datasets considered. Contrary to the expected, the overlap comes not only from *TermPile*, but also from the *TerArXiv* dataset, as it shows also common unique term counts in the same order of maginitude as *TermPile*. Interestingly, *SynTerm* has almost all unique terms present in *TermArXiv*, which speaks to the relatively restricted domain and language used in the original dataset.

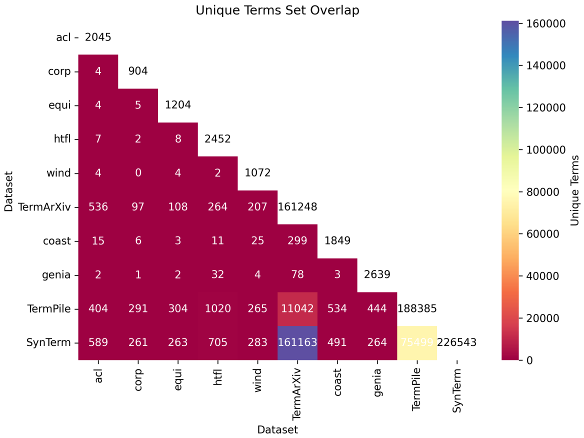

Figure 3: Counts of unique terms among the considered datasets. Off-diagonal counts represent common terms.

H Term Length and Extraction Counts

A possible explanation for the low performance of the few-shot models in some domains like *corp* is given in Table 9. The median number of terms extracted per document for *corp* is zero for both few-shot models (with means of 0.61 and 0.82, respectively). Another contributing factor to the poor performance of the few-shot LLaMA model on *corp* is the extraction of overly long terms. Table 9 reveals that the median length of extracted terms is 27 characters, compared to 11 characters for gold-standard terms.

Although the IOB model trained on *SynTerm* shows a large number of extracted terms per document, many extracted terms are short and imprecise,

Model	**Text**
IOB sequence labeling *SynTerm*	This **is** due **to** the **fact** that **corruption is** often referred **to as** the **crime** without (direct) victim.
IOB sequence labeling (cross-domain)	This is due to the fact that **corruption** is often referred to as the crime without (direct) victim.
`Olmo` few-shot (cross domain)	This is due to the fact that corruption is often referred to as the **crime** without (direct) victim.
`LLaMA` few-shot (cross domain)	This is due to the fact that corruption is often referred to as the crime without (direct) victim.
`LLaMA` fine-tuned *SynTerm*	This is due to the fact that **corruption** is often referred to as the **crime** without (direct) **victim**.
`Olmo` fine-tuned *SynTerm*	This is due to the fact that **corruption** is often referred to as the **crime** without (direct) victim.

Table 8: Models and their extracted terms highlighted in the sentence. The gold-terms are : [’corruption’, ’crime’, ’victim’].

Model	*corp*		*equi*		*htfl*		*wind*		*coast*		*genia*		*acl*	
	Len	Cnt	Len	Cnt	Len	Cnt	Len	Cnt	Len	Cnt	Len	Cnt	Len	Cnt
IOB sequence labeling *SynTerm*	6	11	4	8	7	9	5	8	6	11	7	10	6	45
IOB sequence labeling (cross-domain)	14	806	9	1530	12	556	13	1029	10	258	15	2460	11	181
`Olmo` few-shot (cross domain)	11	0	8	3	13	3	13	2	11	3	13	4	16	3
`LLaMA` few-shot (cross domain)	27	0	8	3	14	2	13	3	11	1	15	4	18	6
`LLaMA` few-shot *SynTerm*	15	2	10	2	17	1	14	2	13	1	15	4	18	7
`LLaMA` fine-tuned *SynTerm*	15	3	9	2	15	3	14	2	14	3	14	4	18	10
`Olmo` fine-tuned *SynTerm*	11	4	8	2	13	4	10	3	13	4	10	5	17	10
Actual	11	2	7	3	9	4	11	1	13	4	11	3	16	14

Table 9: Combined median term lengths (Len) and median term counts per document (Cnt) for each model and dataset.

often including stopwords like “a” or “and” (see Table 9). This results in relatively high F1 scores, but with an imbalance between precision and recall (see Appendix E). This imbalance and performance gap underscores the effectiveness of DiSTER and can be attributed to several key differences in model design and training paradigms. `XLM-R-based` models, fine-tuned for sequence labeling, are optimized for local contextual understanding within narrow task boundaries. In contrast, LLMs are trained on a broader range of tasks and contexts, equipping them with more adaptable reasoning. We posit that this flexibility allows LLMs to better leverage the *SynTerm* dataset, treating it as a text generation task rather than a sequence labeling problem. Qualitative examples for these effects can be found in Appendix F and an analysis of the cross-domain instability of the sequence-labeling models in Appendix K.

I Directional overlap with k-Nearest Neighbors

In order to observe the total average overlap among two datasets, we extend the analysis introduced by Kambhatla et al. (2023) and include, for that, all data on both sets. Our indicator can be defined as follows. Let

- $\mathcal{D} = \{D_1, \dots, D_N\}$ be a collection of text datasets, each embedded in $\mathbb{R}^d$;
- $k \in \mathbb{N}$ the fixed neighbourhood size;
- $N_k(x) \subset \bigcup_{i=1}^{N} D_i$ the (unordered) set of the k nearest neighbours of sample x measured w.r.t. cosine distance in the embedding space;
- $\mathbf{1}_D(y)$ the indicator that neighbour y belongs to dataset D.

Then the Directional Overlap score $O_{A\to B}$, for two datasets $A, B \in \mathcal{D}$, is

$$O_{A\to B} \;:=\; \frac{1}{|A|\,k} \sum_{x\in A} \sum_{y\in N_k(x)} \mathbf{1}_B(y),$$

i.e. the expected fraction of a point’s k closest neighbours that originate from B. For convenience, one may symmetrize $O_{A\to B}$, such that

$$O(A,B) := \tfrac{1}{2}\big[\,O_{A\to B} + O_{B\to A}\big],$$

and by construction $O(A,B) = O(B,A) \in [0,1]$ and $O(A,A) = 1$. However, in this work, we prefer the Directional Overlap indicator because it reflects the asymmetry of test and train datasets.

J Domain and Term Overlap Analysis

To quantify dataset relationships, we extend the embedding-based Nearest Neighbors analysis from (Kambhatla et al., 2023) to construct a directional overlap indicator (defined in Appendix I). Figure 4 shows the percentage of each origin dataset’s nearest neighbors found in each target dataset. For

User Prompt

Context

- Terms are “the designation of a defined concept in a special language by a linguistic expression.” (ISO 1087).
- A term is a word or a phrase that has a specific meaning in a particular context, such as a scientific term or a technical concept.
- The goal is the identify domain-specific concepts not named entities.
- A numerical value or date is not considered a term.
- A organization, group or person is not considered a term.
- Any non-scientific content such as websites, URLs, email addresses, HTML tags, code snippets, etc, are not considered terms.
- A location (country, state, place, ...) is not considered a term.

Instructions

- You will be given a question about a type of expressions in natural language.
- Your task is to determine whether the given expression is either:

1) a term type, or

0) no term in the sense of the definition above.

- Do not overthink the question. Answer based on your intuition and knowledge.
- If the context is not clear, use the most common interpretation of the expression type.
- After you explain your answer, produce a json object with the following format: `{ "answer": [Option Number] }`, where [Option Number] is 0, 1, or 2.

Question

Given expressions of the type {expression_type}, would you consider it

- a term (1), or
- no term (0)?

"""

Prompt 2: User prompt used for entity-type-label generation.

System Prompt

You are a terminology research expert. Your task is to help the user extracting terms from scientific abstracts.

Prompt 3: System prompt used for pseudo-label generation.

User Prompt

Context

- Terms are “the designation of a defined concept in a special language by a linguistic expression.” (ISO 1087).
- A term is a word or a phrase that has a specific meaning in a particular context/domain, such as a scientific term or a technical concept.
- Only extract terms that are relevant to the domain.
- The goal is to identify domain-specific concepts not named entities.
- A numerical value or date is not considered a term.
- An organization, group or person is not considered a term.
- A location (country, state, place, ...) is not considered a term.

Instructions

- You will be given an excerpt from a scientific abstract.
- Your task is to extract all terms from the text.
- Please return only a comma separated list of correct extractions without any additional information.
- If there are no terms in the text, return an empty string.

Question

The domain: {domain}

The abstract: {abstract}

Return list of extracted terms:

Prompt 4: User prompt used for pseudo-label generation.

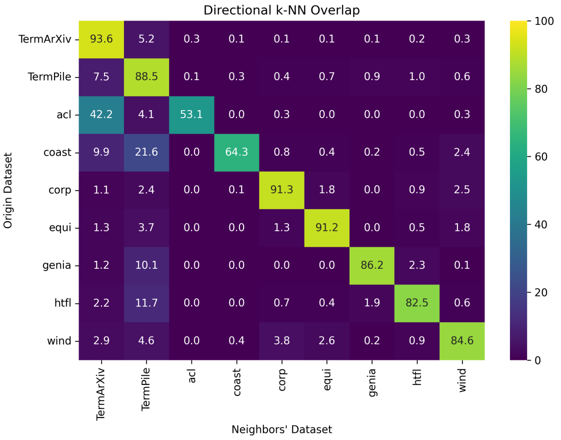

Figure 4: Directional k-NN domain overlap score.

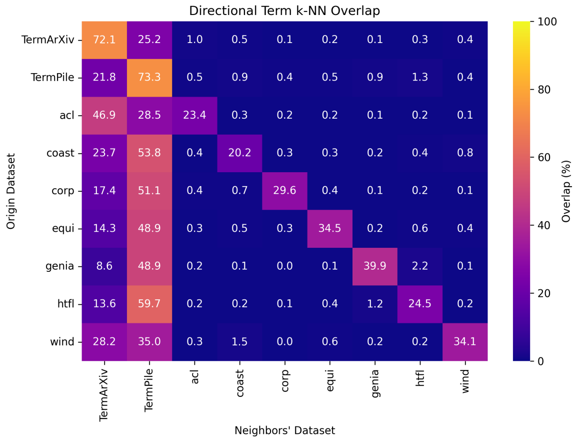

Figure 5: Directional k-NN overlap score for corpus-level terms across domains.

interpretation, read each row as the percentage of an origin dataset's nearest neighbors found in each target dataset.

We observe that *SynTerm* has measurable overlap with all test domains, illustrating the dataset's diversity and broad coverage across different areas. For example for *wind*, we see that 4.6 percent of its nearest neighbors come from the *TermPile* dataset. By adding *TermArXiv* to our *SynTerm* dataset, we add more nearest neighbors for all datasets but especially data points neighboring points from *acl* and *coast*. Analyzing term-level overlap (Figure 5) reveals similar patterns, with most nearest neighbors for *coast* and *acl* terms coming from either *TermArXiv* or *TermPile*. This also holds true for all other test domains to a greater or lesser extent.

These findings explain the strong performance of models fine-tuned on *SynTerm* when evaluated on *coast* and *acl*. The effect is consistent across architectures, with the IOB-based approach gaining 46 and 42 F1 points on *acl* and *coast* respectively when trained on *SynTerm* (Table 2). For *corp* and *equi*, corpus-level term neighbors are prevalent in our synthetic datasets, but domain overlap scores are low. This may explain why the IOB-based approach leverages *SynTerm* to a lesser degree. In order to provide a more precise picture on term overlap, we also provide in Appendix G counts of unique common terms among all datasets.

Notably, LLaMA fine-tuned on *TermArXiv* (containing only scientific terminology with limited overlap to most test domains) demonstrates strong transfer capabilities to unrelated domains like *equi* and *corp*. This suggests LLMs can learn domain-independent representations of terminology, contrasting with sequence tagging models that degrade significantly on out-of-domain data. Such findings highlight LLMs' potential to capture general principles of termhood beyond surface-level domain characteristics.

K Instability using supervised cross-domain Data

While prior work on ACTER reports F1 scores above 50 on *htfl* using the same IOB baseline, these results rely solely on a specific four-domain setting. Expanding to seven domains with varied validation sets reveals a sharp performance drop on *htfl* (Table 2 and Table 4). Surprisingly, increasing domain diversity using leave-one-out training does not guarantee better generalization and sometimes harms performance on individual domains. The directional k-NN overlap heat-map in Appendix J confirms that *htfl* exhibits only maximum 1.9% term-level overlap with the average training mix in this case, explaining its heightened brittleness. This highlights a key trade-off: broader training data may improve average cross-domain results but can reduce domain-specific effectiveness, especially in setups with limited resources.

Replicating the four-domain setup from the ACTER Shared Task confirms that strong results are possible, but only under specific domain splits. As shown in Table 3, the F1 scores of the sequence labeling model vary substantially depending on which domains are used for training and validation. This instability suggests that ATE performance in prior work is highly sensitive to the choice of domains and data splits, raising concerns about the robustness and generalizability of such models. In contrast, using pseudo-labeled data from a LLM yields a more diverse and abundant training corpus, improving cross-domain robustness. Although some domain variance remains, both encoder-based and LLM-based models generalize better, underscoring the importance of data composition and not only model architecture in cross-domain ATE.