Enhancing Automatic Term Extraction with Large Language Models via Syntactic Retrieval

Yongchan Chun, Minhyuk Kim, Dongjun Kim, Chanjun Park†, Heuiseok Lim†

Korea University
{cyc9805, mhkim0929, junkim100, bcj1210, limhseok}@korea.ac.kr

Abstract

Automatic Term Extraction (ATE) identifies domain-specific expressions that are crucial for downstream tasks such as machine translation and information retrieval. Although large language models (LLMs) have significantly advanced various NLP tasks, their potential for ATE has scarcely been examined. We propose a retrieval-based prompting strategy that, in the few-shot setting, selects demonstrations according to *syntactic* rather than semantic similarity. This syntactic retrieval method is domain-agnostic and provides more reliable guidance for capturing term boundaries. We evaluate the approach in both in-domain and cross-domain settings, analyzing how lexical overlap between the query sentence and its retrieved examples affects performance. Experiments on three specialized ATE benchmarks show that syntactic retrieval improves F1-score. These findings highlight the importance of syntactic cues when adapting LLMs to terminology-extraction tasks.

1 Introduction

Automatic Term Extraction (ATE) identifies domain-specific terms essential for tasks such as machine translation, information retrieval, and content curation (Tran et al., 2023). Despite its importance, ATE remains underexplored compared to other information extraction (IE) tasks, particularly in low-resource and specialized domains (Rigouts Terryn et al., 2020).

Large Language Models (LLMs) offer new possibilities for IE through in-context learning, yet prior studies (Ma et al., 2023b; Zhang et al., 2023; Wadhwa et al., 2023; Wan et al., 2023; Xu et al., 2024) show they often underperform compared to task-specific pretrained language models (PLMs), struggling with domain precision and boundary detection. While strategies like prompt engineering and retrieval-based demonstrations have improved IE in general, their application to ATE remains largely unexplored.

We address two key challenges in applying LLMs to ATE: (1) **Dataset scarcity and domain diversity**—ATE lacks diverse datasets beyond the biomedical field (Tran et al., 2023; Rigouts Terryn et al., 2020), limiting cross-domain effectiveness. We propose a retrieval method that generalizes across domains. (2) **Boundary identification**—LLMs struggle to extract precise term spans (Ma et al., 2023b; Wang et al., 2023a), a critical issue given the annotation-intensive nature of ATE (qas, 2016).

To address this, we propose a syntactic retrieval method that selects structurally aligned demonstrations. In both in-domain and cross-domain settings, this approach consistently improves ATE performance by enhancing annotation consistency and extraction accuracy.

2 Related Work

2.1 Automatic Term Extraction

Automatic Term Extraction (ATE) is the task of identifying and ranking domain-specific words or multi-word expressions that represent key concepts within a corpus.

Early days of ATE were focused around utilizing statistical methods such as TF-IDF (Salton et al., 1975), termhood (Vintar, 2010), and unithood (Daille et al., 1994; Vu et al., 2008). With the shift towards deep learning, particularly with the emergence of Transformer architectures (Vaswani et al., 2017) like BERT (Devlin et al., 2019), enhanced ATE by enabling automatic feature learning and boosting performance across multilingual and cross-domain tasks (Lang et al., 2021; Tran et al., 2022; Hazem et al., 2022, 2020).

While PLMs have achieved substantial success in ATE tasks, the application of LLMs to this area

† Corresponding Author

Findings of the Association for Computational Linguistics: ACL 2025, pages 9916–9926

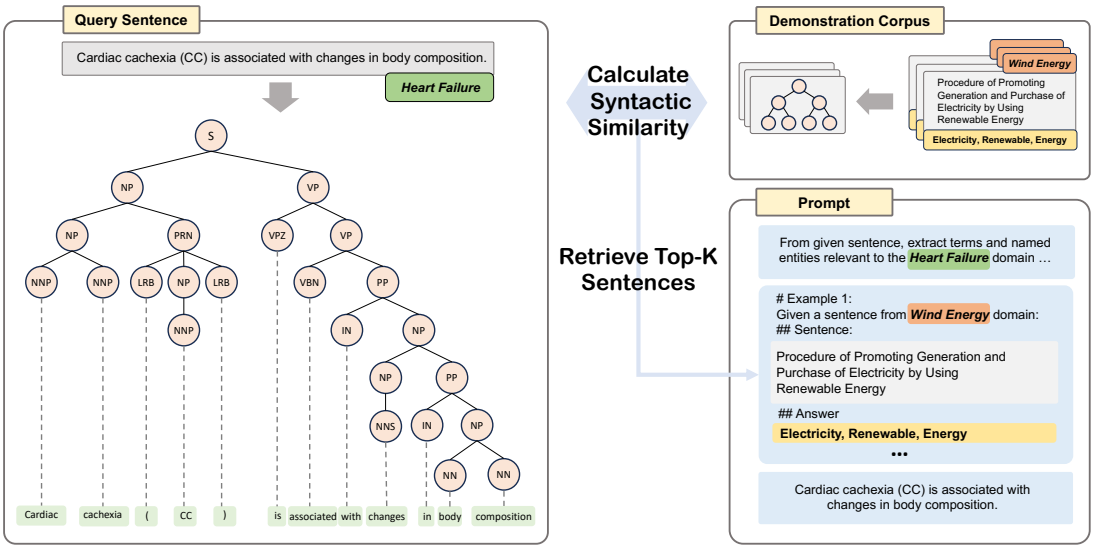

Figure 1: Illustration of the syntactic retrieval process. The example showcases a cross-domain setting in which the domains of the demonstration corpus and the query corpus differ.

has remained relatively underexplored. To address these gaps, our study provides attention to LLMs in the context of ATE.

2.2 Information Extraction with Large Language Models

LLMs have expanded the possibilities of IE through their generative capabilities and vast knowledge base. However, they often underperform compared to task-specific PLMs due to challenges like hallucination and imprecise span boundary identification (Ma et al., 2023b; Wang et al., 2023a; Wadhwa et al., 2023; Sainz et al., 2024).

To address these limitations, researchers have explored several strategies: (1) Task reformulation, framing extraction as question-answering or structured prediction (Wei et al., 2023; Zhang et al., 2023; Ma et al., 2023b); (2) Instruction tuning, fine-tuning LLMs with targeted instructions to compensate for limited IE-specific training data (Wang et al., 2023b; Wadhwa et al., 2023); and (3) In-context learning (ICL), optimizing prompt structures and demonstrations to guide extraction (Blevins et al., 2023; Bian et al., 2023; Ma et al., 2023a; Li et al., 2024b).

This work focuses on ICL due to its efficiency, adaptability, and minimal reliance on task-specific annotation. We extend existing ICL approaches for IE and introduce a retrieval-based method tailored to ATE.

3 Motivation and Methodology

In this section, we formulate ATE within the existing in-context learning framework for LLM-based information extraction (Xu et al., 2024), and introduce our proposed methodology.

3.1 Adapting the LLM-Based Information Extraction Framework for ATE

Given a frozen LLM with parameters θ, a fixed instruction I, a query corpus $\mathcal{C}^{\mathrm{q}} = \{q_i\}_{i=1}^{N_{\mathrm{q}}}$, and a demonstration corpus $\mathcal{C}^{\mathrm{d}} = \{(s_j, T_j^{\mathrm{d}})\}_{j=1}^{N_{\mathrm{d}}}$, where each s_j is a sentence and $T_j^{\mathrm{d}} \subset s_j$ is its associated term set, the goal of ATE is to output a term set $T_i \subset q_i$ for each query sentence q_i.

We begin by retrieving the top K most relevant demonstration pairs for q_i using a retrieval function $f : \mathcal{C}^{\mathrm{q}} \times \mathcal{C}^{\mathrm{d}} \to \mathbb{R}$. The retrieved demonstration set $\mathcal{D}_i$ is defined as:

$$\mathcal{I}_i = \underset{j \in [1, N_{\mathrm{d}}]}{\arg\,\mathrm{TopK}}\, f\big(q_i, (s_j, T_j^{\mathrm{d}})\big),$$

$$\mathcal{D}_i = \big\{(s_j, T_j^{\mathrm{d}}) \mid j \in \mathcal{I}_i\big\}.$$

We then construct a prompt as follows:

$$\mathrm{prompt}_i = I \oplus \mathcal{D}_i \oplus q_i, \tag{1}$$

where $\oplus$ denotes string concatenation.

Our objective is to discover a retrieval method $\hat{f}$ that maximizes the probability of generating the correct term set:

$$\hat{f} = \underset{f}{\mathrm{argmax}} \prod_{i=1}^{N_{\mathrm{q}}} \prod_{t \in T_i} p\big(t \,\big|\, \mathrm{prompt}_i; \theta\big)$$

Dataset	Retrieval Method	Llama-3.1-8B-IT			Gemma-2-9B-IT			Mistral-Nemo		
		P	R	F1	P	R	F1	P	R	F1
Cross-domain										
ACTER	BGE-large-en	65.2 ±1.3	51.8 ±1.1	57.7 ±1.0	61.5 ±1.2	56.1 ±1.1	*58.8 ±1.0	64.9 ±1.4	**44.9** ±1.1	52.8 ±1.0
	BGE-en-ICL	66.5 ±1.3	46.4 ±1.1	*54.7 ±1.1	63.2 ±1.2	51.0 ±1.2	*56.5 ±1.2	65.0 ±1.5	40.0 ±1.1	*49.5 ±1.1
	BM25	**66.9** ±1.3	49.4 ±1.1	*56.8 ±1.0	61.4 ±1.1	53.0 ±1.2	*56.8 ±1.1	65.8 ±1.3	43.0 ±1.1	*52.0 ±1.1
	Random	66.4 ±1.5	51.3 ±1.3	57.9 ±1.0	62.9 ±2.1	55.2 ±2.1	*58.8 ±2.1	66.1 ±1.9	44.1 ±1.2	52.6 ±1.1
	FastKASSIM	64.3 ±1.3	**53.0** ±1.0	**58.0** ±1.1	**64.3** ±1.1	**56.6** ±1.1	**60.2** ±1.0	**66.7** ±1.4	44.0 ±1.2	**53.0** ±1.1
In-domain										
ACLR2	BGE-large-en	75.9 ±3.3	78.3 ±2.6	77.1 ±2.6	**77.5** ±3.3	83.5 ±2.3	80.4 ±2.4	72.5 ±3.2	73.0 ±3.0	72.8 ±2.8
	BGE-en-ICL	71.9 ±3.1	74.0 ±2.7	*72.9 ±2.7	74.9 ±3.4	83.2 ±2.4	78.8 ±2.5	71.2 ±3.1	71.7 ±3.1	71.4 ±2.8
	BM25	77.3 ±3.2	78.3 ±2.6	77.7 ±2.6	77.3 ±3.0	**84.4** ±2.2	**80.7** ±2.2	**74.5** ±3.2	**75.2** ±2.9	**74.8** ±2.7
	Random	75.8 ±3.4	75.2 ±3.6	*75.4 ±3.2	77.3 ±3.6	79.4 ±3.5	78.3 ±2.5	72.5 ±3.6	70.0 ±3.5	*71.2 ±2.9
	FastKASSIM	**77.4** ±2.9	**78.7** ±2.4	**78.1** ±2.4	75.9 ±3.2	82.2 ±2.5	78.8 ±2.5	73.1 ±3.1	73.1 ±3.0	73.1 ±2.6
BCGM	BGE-large-en	40.7 ±1.2	55.2 ±1.2	*46.9 ±1.1	33.8 ±1.1	52.6 ±1.2	*41.1 ±1.1	43.4 ±1.2	52.8 ±1.1	*47.6 ±1.1
	BGE-en-ICL	39.0 ±1.2	54.0 ±1.2	*45.3 ±1.2	31.6 ±1.1	52.0 ±1.2	*39.3 ±1.1	37.1 ±1.3	48.1 ±1.2	*41.9 ±1.2
	BM25	38.8 ±1.2	54.6 ±1.1	*45.4 ±1.1	33.5 ±1.1	54.6 ±1.2	*41.5 ±1.2	40.1 ±1.3	50.9 ±1.2	*44.8 ±1.2
	Random	43.1 ±5.5	53.2 ±2.4	*47.1 ±3.3	40.9 ±3.9	56.8 ±1.9	*47.3 ±3.0	48.5 ±6.8	52.0 ±2.7	*50.2 ±4.5
	FastKASSIM	**43.8** ±1.2	**56.6** ±1.2	**49.4** ±1.1	**44.1** ±1.1	**60.8** ±1.2	**51.1** ±1.1	**50.8** ±1.4	**57.2** ±1.2	**53.8** ±1.1

Table 1: Performance comparison of the evaluated LLMs across different similarity metrics. **P**, **R**, and **F1** refer to precision, recall, and F1-score, respectively. The number of shots used in the experiment is fixed to 10. The highest score along each metric for each dataset is indicated in bold. p-value with less than 0.05 is marked with $*$.

3.2 Limitations of Semantic Retrieval in ATE

Existing LLM-based information extraction approaches typically retrieve semantically similar sentences using cosine similarity between sentence embeddings or entity embeddings (Kim et al., 2024; Wang et al., 2023a; Wan et al., 2023). These methods aim to retrieve examples that contain the correct answer, increasing the likelihood of generating the correct terms. However, this approach has limitations in ATE, especially when the retrieved demonstrations do not overlap with the gold term set, i.e, $|\{T_j^{\mathrm{d}} \mid j \in \mathcal{I}_i\} \bigcap T_i| = 0$, as in low-resource or cross-domain scenarios. In such cases, retrieving semantically similar sentences may not provide useful guidance, as the retrieved examples may come from a different domain and fail to inform the term extraction process. This limitation is particularly problematic for ATE, where datasets covering diverse domains are scarce (Rigouts Terryn et al., 2020; Tran et al., 2023).

3.3 Syntactic Retrieval for ATE

Rather than retrieving examples that directly overlap with the target term set T, we guide the LLM using syntactic patterns to improve term boundary identification. For instance, consider the query sentence q_i: *"The blood pressure measurement is recorded daily."*—a medical domain sentence where possible annotations include *"blood pressure"* or *"blood pressure measurement."* By retrieving a syntactically similar sentence such as *"The rotor speed reading is logged every minute."* from the wind energy domain, with the annotated term *"rotor speed"*, we provide structural guidance for consistent annotation.

To implement syntactic retrieval, we first generate constituency parse trees for the query sentence q_i and each sentence in the demonstration corpus $\{s_1^{\mathrm{d}}, \ldots s_{N_{\mathrm{d}}}^{\mathrm{d}}\}$. We then compute syntactic similarity using FastKASSIM (Chen et al., 2023), an efficient algorithm that leverages a Label-based Tree Kernel to compare parse trees. See Appendix A.5 for further details on FastKASSIM.

After retrieving structurally similar examples, we construct prompts as defined in Equation 1 and pass them to the LLM. The overall process is illustrated in Figure 1.

3.4 Term Overlap Ratio

To analyze the explicit advantage of our retrieval method, we introduce Term Overlap Ratio (TOR), which evaluates the applicability of our method when there is minimal overlap between the terms to be extracted from the query sentence and those in the demonstration set. We define TOR as follows:

$$\mathrm{TOR} = \frac{1}{N_q} \sum_{i=1}^{N_q} \frac{|\{T_j^{\mathrm{d}} \mid j \in \mathcal{I}_i\} \bigcap T_i|}{|T_i|} \tag{2}$$

This metric measures the proportion of T_i that also appear in the retrieved demonstrations. Additionally, we examine the correlation between TOR and micro F1-score to assess how the degree of term overlap impacts overall performance.

Domain	Retrieval Method	Correlation	TOR
Cross-domain	BGE-large-en	-2.48	0.22
	BGE-en-icl	0.88	0.01
	BM25	1.14	0.21
	Random	1.35	0.01
	FastKASSIM	0	0
In-domain	BGE-large-en	15.44	20.51
	BGE-en-icl	9.61	17.38
	BM25	14.62	20.31
	Random	2.47	0.78
	FastKASSIM	5.69	0.98

Table 2: Term Overlap Ratio (TOR) and its correlation with micro F1-score across domains. The number of shots is fixed at 5, and results are averaged across all datasets and models.

4 Experiments

This section presents experimental results for various retrieval methods across multiple datasets. Details on the datasets, models, and baseline retrieval methods are provided in Appendix A. The main results are summarized in Table 1, and qualitative examples of retrieved sentences from the ACTER dataset are shown in Table 6.

4.1 Main Results

Cross-Domain Result FastKASSIM achieves the highest F1-score across all tested models. However, for LLaMA-3.1 and Mistral-Nemo, the differences between FastKASSIM and both BGE-large-en and random retrieval are not statistically significant (p-value > 0.05), indicating comparable performance among these methods. We hypothesize that the competitive performance of BGE-large-en stems from prior findings that transformer-based sentence embeddings capture not only semantic but also partial syntactic information (Chi et al., 2020; Pérez-Mayos et al., 2021; Nikolaev and Padó, 2023). In addition, the random retrieval method ranks second for most models (except Mistral-Nemo), performing better than expected. This suggests that higher diversity among retrieved examples may enhance generalization, making random retrieval a surprisingly effective and efficient alternative in low-resource or time-constrained scenarios.

Nonetheless, FastKASSIM's consistent top performance in general suggests that explicitly disentangling syntactic and semantic features and focusing solely on syntactic structure is more beneficial.

In-Domain Result On the ACLR2 dataset, FastKASSIM outperforms other methods for LLaMA-3.1, while BM25 achieves the best performance on the remaining models. The superior results of semantic- and lexical-based methods over syntactic-based retrieval in this setting are expected, as in-domain sentences with high semantic or lexical similarity are more likely to contain gold terms. As shown in Section 4.1.1, this is supported by higher TOR values and a strong correlation between TOR and micro F1-score.

In contrast, on the BCGM dataset, FastKASSIM consistently outperforms all baselines across models. Together with its strong performance on LLaMA-3.1 for ACLR2, these results suggest that syntactic alignment remains a strong cue even when semantic and lexical overlap is high, reinforcing the utility of syntactic retrieval as a reliable annotation guide.

4.1.1 Analysis Through Term Overlap Ratio

Table 2 presents the TOR results and its correlation with micro F1-score. Since the distributions of TOR and micro F1-score are not normally distributed, we use Spearman's rank correlation for analysis.

In cross-domain settings, FastKASSIM has a TOR of zero, meaning it does not look for sentences that contain gold terms. As a result, micro F1-score shows no correlation with term overlap. In contrast, BGE-large-en exhibits a higher TOR of 0.22, accompanied by a negative correlation. This is possibly due to certain words functioning as domain-specific terms in one field but remain generic in others (e.g., "cough" is a medical term but appears frequently in non-medical contexts). This discrepancy may confuse LLMs, leading to performance degradation as TOR increases.

In in-domain settings, TOR is generally high for BGE-large-en, BGE-en-icl, and BM25, and their performance is positively correlated with TOR. In contrast, FastKASSIM and Random retrieval have lower TOR and weaker correlation. This suggests that while FastKASSIM does not explicitly retrieve documents containing ground-truth terms, it still achieves competitive performance, as shown in Table 1.

5 Ablation Study

In this section, we conduct ablation studies to examine: (1) how the performance of each retrieval method scales with the number of demonstrations, and (2) how our in-context LLM approach compares to strong PLM baselines across selected

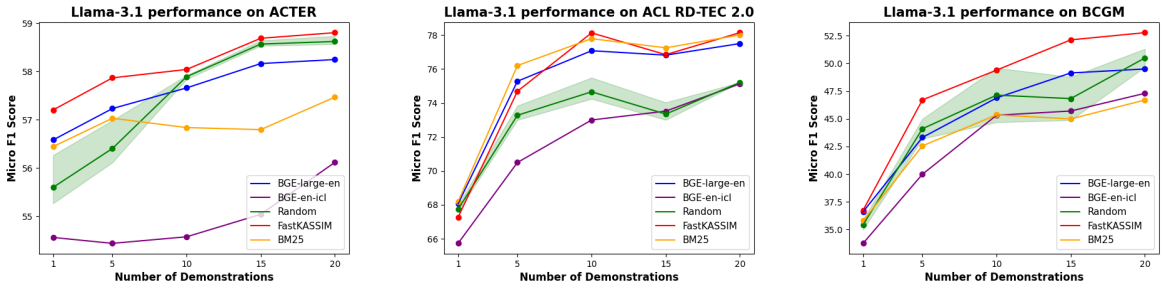

Figure 2: Comparison of retrieval methods for Llama-3.1-8B-IT on ACTER, ACLR2, and BCGM datasets. We report micro F1-score. The 95% confidence interval is reported for random retrieval.

Dataset	Model	P	R	F1
Cross-domain				
ACTER	RoBERTa-large	**69.3** ±1.3	54.7 ±1.2	**61.2** ±1.1
	BART-large	66.1 ±1.2	48.2 ±1.2	55.8 ±1.2
	Gemma-2 FastKASSIM	64.3 ±1.1	**56.6** ±1.1	60.2 ±1.0
In-domain				
ACLR2	RoBERTa-large	**85.8** ±2.5	**88.9** ±2.1	**87.4** ±2.1
	BART-large	81.3 ±2.6	84.8 ±2.3	83.0 ±2.2
	Gemma-2 BM25	77.3 ±3.0	84.4 ±2.2	80.7 ±2.2
BCGM	RoBERTa-large	**88.0** ±0.8	**89.0** ±0.8	**88.5** ±0.8
	BART-large	79.4 ±1.0	77.1 ±1.1	78.2 ±1.0
	Mistral FastKASSIM	50.8 ±1.4	57.2 ±1.2	53.8 ±1.1

Table 3: Performance of pretrained language models (PLMs) on the ACTER, ACLR2, and BCGM datasets. For reference, the table also showcases the best-performing LLM configuration and its retrieval method (see Table 1) based on F1-score. The best score along each metric for each dataset is bolded, and second best score is underlined.

datasets. We also evaluate the impact of different constituency parsers on our syntactic similarity method; detailed results are provided in Appendix C.

5.1 Number of Demonstrations

Figure 2 shows how LLaMA-3.1 scales with the number of demonstrations under the various retrieval strategies. On the ACTER and BCGM datasets, FastKASSIM is consistently the top performer at every shot count. For ACLR2, however, performance oscillates: BGE-large-en, BM25, and FastKASSIM each take the lead at different points, so no single method emerges as uniformly superior. These trends are similar to results discussed in Section 4.1. The corresponding curves for the other models are provided in Figure 3, which shows the similar trend as LLaMA-3.1.

5.2 Comparison with Pretrained Language Models

Earlier PLM works (Lang et al., 2021; Rigouts Terryn et al., 2020) employ *non-sequential* tagging objective, which is out of step with recent advances in ATE (Rigouts Terryn et al., 2021). For a fair comparison, we therefore retrain each PLM using the hyper-parameter settings of (Lang et al., 2021) (batch size, gradient accumulation, learning rate), altering only the tagging objective.

Table 3 presents the results. On ACTER, our best configuration achieved with FastKASSIM on Gemma-2, F1-score of 60.2, is comparable to the PLM baselines. On ACLR2 and BCGM, however, significant performance gap remains, where PLMs outperform LLMs. Additionally, RoBERTa consistently outperforms BART on all datasets, reflecting the difficulties generation-based models face in token-level classification.

In summary, LLMs can match PLM performance in **cross-domain** scenarios, as the broader knowledge base and flexibility of LLMs allow for better adaptation. In **in-domain** settings, PLMs remain superior—likely owing to task-specific fine-tuning.

6 Conclusion

We explored the use of LLMs for ATE and proposed a syntactic retrieval method to address two key challenges: dataset scarcity and term boundary identification. Experiments on ACTER, ACLR2, and BCGM showed that syntactic similarity-based retrieval improves ATE performance across both in-domain and cross-domain settings.

We also introduced the Term Overlap Ratio to analyze how different retrieval strategies depend on the presence of gold terms in the demonstration corpus. Our results indicate that syntactic retrieval relies less on such overlap compared to semantic or lexical methods, highlighting its robustness in low-resource scenarios.

Limitations

While our study demonstrates the potential of LLMs for ATE, several limitations remain. **First**, although syntactic retrieval mostly outperforms semantic retrieval, the absolute improvements in F1-score are modest, suggesting inherent limitations in in-context learning for ATE. **Second**, as shown in Section 5.2, LLMs still underperform relative to domain-tuned PLMs. This underscores the need for further adaptation or fine-tuning strategies as done in (Wang et al., 2023b; Wadhwa et al., 2024), which we leave for future work.

Ethics Statement

All experiments in this study were conducted with fairness and transparency in mind. We ensured that our methodologies and evaluation metrics were applied consistently across all models and datasets. Additionally, the datasets used in our experiments are publicly available, and no personally identifiable information (PII) is involved. We adhered to ethical guidelines for data usage and ensured that all results were reported accurately to reflect the true performance of the models.

Acknowledgement

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(NRF-2021R1A6A1A03045425), Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (RS-2024-00398115, Research on the reliability and coherence of outcomes produced by Generative AI), Institute for Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. RS-2022-II220369, (Part 4) Development of AI Technology to support Expert Decision-making that can Explain the Reasons/Grounds for Judgment Results based on Expert Knowledge), and Institute of Information & communications Technology Planning & Evaluation (IITP) under the artificial intelligence star fellowship support program to nurture the best talents (IITP-2025-RS-2025-02304828) grant funded by the Korea government(MSIT).

References

2016. The ACL RD-TEC 2.0: A language resource for evaluating term extraction and entity recognition methods. pages 1862–1868, Portorož, Slovenia. European Language Resources Association (ELRA).

Junyi Bian, Jiaxuan Zheng, Yuyi Zhang, and Shanfeng Zhu. 2023. Inspire the large language model by external knowledge on biomedical named entity recognition. *arXiv preprint arXiv:2309.12278*.

Terra Blevins, Hila Gonen, and Luke Zettlemoyer. 2023. Prompting language models for linguistic structure. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6649–6663.

Reihane Boghrati, Joe Hoover, Kate M Johnson, Justin Garten, and Morteza Dehghani. 2018. Conversation level syntax similarity metric. *Behavior research methods*, 50(3):1055–1073.

Maximillian Chen, Caitlyn Chen, Xiao Yu, and Zhou Yu. 2023. FastKASSIM: A fast tree kernel-based syntactic similarity metric. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 211–231, Dubrovnik, Croatia. Association for Computational Linguistics.

Ethan A. Chi, John Hewitt, and Christopher D. Manning. 2020. Finding universal grammatical relations in multilingual BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577, Online. Association for Computational Linguistics.

Béatrice Daille, Eric Gaussier, and Jean-Marc Langé. 1994. Towards automatic extraction of monolingual and bilingual terminology. pages 515–524.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Amir Hazem, Mérieme Bouhandi, Florian Boudin, and Beatrice Daille. 2020. TermEval 2020: TALN-LS2N system for automatic term extraction. In *Proceedings of the 6th International Workshop on Computational Terminology*, pages 95–100, Marseille, France. European Language Resources Association.

Amir Hazem, Merieme Bouhandi, Florian Boudin, and Beatrice Daille. 2022. Cross-lingual and cross-domain transfer learning for automatic term extraction from low resource data. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 648–662, Marseille, France. European Language Resources Association.

Hongjin Kim, Jai-Eun Kim, and Harksoo Kim. 2024. Exploring nested named entity recognition with large language models: Methods, challenges, and insights. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8653–8670, Miami, Florida, USA. Association for Computational Linguistics.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 423–430, Sapporo, Japan. Association for Computational Linguistics.

Christian Lang, Lennart Wachowiak, Barbara Heinisch, and Dagmar Gromann. 2021. Transforming term extraction: Transformer-based approaches to multilingual term extraction across domains. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3607–3620, Online. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Chaofan Li, MingHao Qin, Shitao Xiao, Jianlyu Chen, Kun Luo, Yingxia Shao, Defu Lian, and Zheng Liu. 2024a. Making text embedders few-shot learners. *Preprint*, arXiv:2409.15700.

Mingchen Li, Huixue Zhou, Han Yang, and Rui Zhang. 2024b. Rt: a retrieving and chain-of-thought framework for few-shot medical named entity recognition. *Journal of the American Medical Informatics Association*, page ocae095.

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. *Preprint*, arXiv:1907.11692.

Xilai Ma, Jing Li, and Min Zhang. 2023a. Chain of thought with explicit evidence reasoning for few-shot relation extraction. *arXiv preprint arXiv:2311.05922*.

Yubo Ma, Yixin Cao, Yong Hong, and Aixin Sun. 2023b. Large language model is not a good few-shot information extractor, but a good reranker for hard samples! In *Findings of the Association for Computational Linguistics: EMNLP 2023*. Association for Computational Linguistics.

Dmitry Nikolaev and Sebastian Padó. 2023. Investigating semantic subspaces of transformer sentence embeddings through linear structural probing. *Preprint*, arXiv:2310.11923.

Laura Pérez-Mayos, Roberto Carlini, Miguel Ballesteros, and Leo Wanner. 2021. On the evolution of syntactic information encoded by BERT's contextualized representations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2243–2258, Online. Association for Computational Linguistics.

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.

Ayla Rigouts Terryn, Véronique Hoste, and Els Lefever. 2021. Tagging terms in text: A supervised sequential labelling approach to automatic term extraction. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 28.

Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. 2024. Gollie: Annotation guidelines improve zero-shot information-extraction. *Preprint*, arXiv:2310.03668.

G. Salton, A. Wong, and C. S. Yang. 1975. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

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, Andraz Pelicon, Antoine Doucet, and Senja Pollak. 2022. *Ensembling Transformers for Cross-domain Automatic Term Extraction*, page 90–100. Springer International Publishing.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Spela Vintar. 2010. Bilingual term recognition revisited: The bag-of-equivalents term alignment approach and its evaluation. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 16(2):141–158.

Thuy Vu, Aiti Aw, and Min ZHANG. 2008. Term extraction through unithood and termhood unification. *Proceedings of the Third International Joint Conference on Natural Language Processing*.

Somin Wadhwa, Silvio Amir, and Byron C Wallace. 2023. Revisiting relation extraction in the era of large language models. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2023, page 15566. NIH Public Access.

Somin Wadhwa, Silvio Amir, and Byron C. Wallace. 2024. Revisiting relation extraction in the era of large language models. *Preprint*, arXiv:2305.05003.

Robert A. Wagner and Michael J. Fischer. 1974. The string-to-string correction problem. *J. ACM*, 21(1):168–173.

Zhen Wan, Fei Cheng, Zhuoyuan Mao, Qianying Liu, Haiyue Song, Jiwei Li, and Sadao Kurohashi. 2023. Gpt-re: In-context learning for relation extraction using large language models. *Preprint*, arXiv:2305.02105.

Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. 2023a. Gpt-ner: Named entity recognition via large language models. *Preprint*, arXiv:2304.10428.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, et al. 2023b. Instructuie: Multi-task instruction tuning for unified information extraction. *arXiv preprint arXiv:2304.08085*.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, et al. 2023. Zero-shot information extraction via chatting with chatgpt. *arXiv preprint arXiv:2302.10205*.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597.

Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, Yang Wang, and Enhong Chen. 2024. Large language models for generative information extraction: A survey. *Preprint*, arXiv:2312.17617.

Alexander Yeh, Alexander Morgan, Marc Colosimo, and Lynette Hirschman. 2005. Biocreative task 1a: gene mention finding evaluation. *BMC bioinformatics*, 6:1–10.

Kai Zhang, Bernal Jiménez Gutiérrez, and Yu Su. 2023. Aligning instruction tasks unlocks large language models as zero-shot relation extractors. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 794–812.

Yu Zhang, Houquan Zhou, and Zhenghua Li. 2020. Fast and accurate neural crf constituency parsing. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI-PRICAI-2020, page 4046–4053. International Joint Conferences on Artificial Intelligence Organization.

A Implementation Details

A.1 Dataset

We conduct our experiments on both cross-domain and in-domain datasets. For the cross-domain setting, we employ the **Annotated Corpora for Term Extraction Research** (Rigouts Terryn et al., 2020), while for in-domain setting, we use two widely studied datasets: **ACL RD-TEC 2.0**(qas, 2016), and **BioCreAtIvE Task 1A: Gene Mention**(Yeh et al., 2005). Table 4 summarizes the dataset statistics.

Dataset Name	Subset	Avg Words	Avg Terms
ACTER	Train	19	2
	Validation	17	3
	Test	19	4
ACLR2	Train	23	3
	Validation	23	4
	Test	19	3
BCGM	Train	22	2
	Validation	23	2
	Test	23	2

Table 4: Average number of words and terms for each dataset.

Annotated Corpora for Term Extraction Research (ACTER) The ACTER dataset spans four distinct domains: Wind Energy, Corruption, Dressage, and Heart Failure. Wind Energy and Corruption consists train dataset, Corruption constructs validation dataset, while Heart Failure constructs test dataset. In addition, terms within ACTER are categorized into four main groups: (1) *Specific Terms*, which are understood primarily by domain experts (e.g., "Cardiac cachexia" in the medical field); (2) *Common Terms*, known to the general public without requiring specialized domain knowledge (e.g., "cough" in the medical domain); and (3) *Out-of-Domain Terms*, which are familiar to experts in other domains (e.g., "*p*-value" in the medical domain). (4) *Named Entities*, name of real-word objects such as person, locations, organizations, etc (e.g "Johns Hopkins Hospital" in the medical domain).

In addition, ACTER is a mulitilingual dataset, spanning English, French and Dutch. Since the focus of our work lies in discovering the optimal retrieval strategy for ATE, we limit our experiments to the English subset of the ACTER.

Also, there have been debates over whether Named Entities should be included as part of term

extraction tasks. We decided to consider Named Entities as terms due to following reasons: (1) LLMs have demonstrated reliable performance in information extraction tasks, including Named Entity Recognition (NER). (2) Recent researches around ATE consider named entities as terms (Lang et al., 2021; Tran et al., 2023).

ACL RD-TEC 2.0 (ACLR2) The ACLR2 dataset was annotated by two experts in the field of computational linguistics, with multiple rounds of inter-annotator agreement. However, due to inherent subjectivity, the annotators were unable to reach full agreement, resulting in two subsets of annotated data. Unlike previous studies that evaluate each subset separately, our work integrates identical terms annotated by both experts to construct a single unified test set. The remaining data is split into training and validation sets, with a ratio of 3:1.

BioCreAtIvE Task 1A: Gene Mention (BCGM) The BCGM dataset, part of the BioCreAtIvE challenge, focuses on gene-related terminology in biomedical texts. It contains sentences from Medline abstracts, with manually annotated terms. These terms primarily include gene and protein names, along with related biological entities such as domains, motifs, and families.

A.2 Models

We evaluate ATE performance mainly on three LLMs: Llama-3.1-8B-IT (Dubey et al., 2024), Gemma-2-9B-IT(Team et al., 2024), and Mistral-Nemo-Instruct-2407. Across all models, we adopt greedy decoding strategy to ensure deterministic output generation.

We also report performance of PLMs as baseline, specifically RoBERTa-large(Liu et al., 2019) and BART-large(Lewis et al., 2020), which have reached SOTA performance in ATE task, as proposed in (Lang et al., 2021; Tran et al., 2023).

Training and evaluation are conducted using the HuggingFace[1] and LlamaIndex[2] libraries.

A.3 Baseline Retrieval Methods

We evaluate our syntactic-based retrieval method, FastKASSIM against two semantic similarity models, BGE-large-en-v1.5(Xiao et al., 2023) and BGE-en-icl(Li et al., 2024a)). Additionally, we include a lexical-based approach, BM25 and a random retrieval baseline. For the random baseline, sentences are selected uniformly at random using four different random seeds, and results are averaged across runs. Across all datasets, we use the training set as the demonstration corpus and the test set as the query corpus.

A.4 Evaluation Method

We report precision, recall, and F1-score to evaluate model performance. Specifically, we apply bootstrapping with 10,000 resamples to compute performance statistics and report 95% confidence intervals. To assess statistical significance, we conduct hypothesis testing using p-values, comparing the FastKASSIM-based method against baseline retrieval strategies.

Following recently adopted sequence labeling evaluation approach (Rigouts Terryn et al., 2021), we directly compare the model-generated terms to the gold annotations without additional normalization.

A.5 Syntactic Similarity Metrics

To date, the only metrics explicitly designed for word-, sentence-, and document-level syntactic similarity are the *ConversAtion-level Syntax SImilarity Metric* (CASSIM) (Boghrati et al., 2018) and its successor, the *Fast Tree-Kernel-bAsed Syntactic SIMilarity Metric* (FastKASSIM) (Chen et al., 2023).

CASSIM encodes each sentence as an unlexicalized constituency parse tree and compares document pairs by computing length-normalized Edit Distances (Wagner and Fischer, 1974) between all cross-document sentence pairs. It then applies the Hungarian algorithm to align the most similar sentence pairs and aggregates their distances into a single 0–1 similarity score. In crowdsourced dataset evaluations, CASSIM successfully distinguished syntactically similar from dissimilar sentence pairs, outperforming Linguistic Style Matching and other syntactic baselines.

FastKASSIM is built upon CASSIM, but replaces the Edit Distance with a Label-based Tree Kernel that counts shared subtree fragments. By caching recursive computations, it avoids the tendency of edit distance to overestimate similarity between structurally dissimilar sentences and significantly reduces computational complexity. On the ChangeMyView corpus, FastKASSIM achieves a 2.4–5.3× speedup over CASSIM and demonstrates stronger

[1] https://huggingface.co/
[2] https://github.com/run-llama/llama_index

correlation with human judgments of syntactic similarity. Given its efficiency and improved alignment with human perception, we adopt FastKASSIM as the primary syntactic similarity metric in our work.

B Instruction Templates

This section outlines the instructions used in our experiments. We define [DOMAIN_NAME] as the domain from which terms are to be extracted, and [DEMONSTRATIONS] as the retrieved examples.

B.1 Default Instruction:

> From the given sentence, extract terms and named entities relevant to the [DOMAIN_NAME] domain. If no relevant terms or named entities are found, return "No term".
>
> # Guidelines:
>
> 1. Extract only the terms and named entities present in the sentence.
> 2. Focus solely on English terms.
> 3. Provide only the extracted terms and named entities or "No term," without additional commentary.
> 4. Use commas to separate each term and named entity.
> 5. Maintain the original case (e.g., lowercase, capitalized) of each term.
>
> [DEMONSTRATIONS]
> Given sentence from the [DOMAIN_NAME] domain:

B.2 Instruction Prompt for BGE-en-icl:

> Given a sentence and a specific domain, retrieve sentences from other domains that follow a similar structure while using domain-specific terminology. These examples should help language models identify and extract key terms related to the original domain from the given sentence.
> Domain: [DOMAIN_NAME] Sentence:

C Impact of Parse Tree Construction Methods

This section examines how the choice of constituency parser affects syntactic similarity under the FastKASSIM framework. We compare two major families of parsers: (1) probabilistic models and (2) neural network (NN)-based models. For the probabilistic model, we use the *unlexicalized PCFG* parser (Klein and Manning, 2003) from the Stanford Parser[3]. For the NN-based model, we adopt the *CRF Parser + RoBERTa* (Zhang et al., 2020), implemented in the SuPar library[4].

Dataset	Metric	Llama-3.1-8B-IT		Gemma-2-9B-IT		Mistral-Nemo	
		PCFG	NN	PCFG	NN	PCFG	NN
ACTER	P	64.3 ±1.3	67.9 ±1.3	64.3 ±1.1	64.4 ±1.2	66.7 ±1.4	66.6 ±1.4
	R	53.0 ±1.0	49.5 ±1.0	56.6 ±1.1	54.1 ±1.2	44.0 ±1.2	42.6 ±1.1
	F1	**58.0** ±1.1	57.3 ±1.0	**60.2** ±1.0	58.8 ±1.1	**53.0** ±1.1	52.0 ±1.1
ACLR2	P	77.4 ±2.9	75.0 ±3.3	75.9 ±1.2	76.3 ±3.2	73.1 ±3.1	73.5 ±3.3
	R	78.7 ±2.4	74.4 ±2.7	82.2 ±2.5	81.3 ±2.4	73.1 ±3.0	71.5 ±3.1
	F1	**78.1** ±2.4	74.7 ±2.7	**78.8** ±2.5	78.7 ±2.5	**73.1** ±2.6	72.5 ±2.8
BCGM	P	43.8 ±1.2	42.9 ±8.3	44.1 ±1.1	42.7 ±8.0	50.8 ±1.4	47.0 ±9.2
	R	56.6 ±1.2	59.3 ±7.7	60.8 ±1.2	64.3 ±8.1	57.2 ±1.2	57.8 ±8.1
	F1	49.4 ±1.1	**49.6** ±8.1	**51.1** ±1.1	**51.1** ±8.0	**53.8** ±1.1	51.7 ±8.4

Table 5: Performance of two tree-parsing approaches—an *unlexicalized PCFG* (PCFG) and a neural *CRF + RoBERTa* (NN) model. Columns **P**, **R**, and **F1** denote precision, recall, and F1, respectively. For each setting, the higher F1-score of the two approaches is shown in bold.

We evaluate both parsers across our implemented models and datasets. Table 5 presents the results. Overall, the PCFG parser outperforms the neural parser. We hypothesize that this is because PCFGs rely solely on syntactic structure, whereas neural parsers incorporate both syntactic and semantic signals, potentially reducing syntactic alignment accuracy. This observation is consistent with our findings in Section 4.1, which show that isolating syntax from semantics improves retrieval quality.

Based on these results, we adopt the *unlexicalized PCFG* parser for all experiments in this work.

[3] https://nlp.stanford.edu/software/lex-parser.shtml

[4] https://github.com/yzhangcs/parser

Query Sentence	Domain	Term
The analysis included a large study sample with more than 60,000 patients across 4372 hospitals.	Heart Failure	patients, hospitals

(a) Example of query sentence and extracted terms.

Similarity Metric	Retrieved Sentence	Domain	Term
BGE-en-large	The author especially thanks his supervisor for his patience and trust during the study.	Corruption	No term
	The contractor will be tasked to set up the network of 27 local research correspondents and cover the coordination/logistic aspects.	Wind Energy	contractor
	Seven participants came to the public meeting.	Corruption	No term
	The studies of this thesis can be surely used for further works.	Wind Energy	No term
	This survey is conducted every two years.	Corruption	No term
BGE-en-icl	7,355 5,241 2,750 1,815	Wind Energy	No term
	52,534 21,238 55,501 86,160	Wind Energy	No term
	2 20 52 88 152 239 318 418 490 556 590 605 610 605 600 590 580 570	Wind Energy	No term
	82 4.8 Results.	Wind Energy	No term
	Hence, by simply including all these power plants operating on the grid (excl.	Corruption	No term
BM25	Technical Wind Energy Potential (MW) 83.000 14.000 12.000 57.000 22.000 42.000 35.000 43.000 20.000	Wind Energy	Technical Wind Energy Potential, MW
	Any regular income Members receive in respect of each item declared in accordance with the first subparagraph shall be placed in one of the following categories: EUR 500 to EUR 1 000 a month; EUR 1 001 to EUR 5 000 a month; EUR 5 001 to EUR 10 000 a month; more than EUR 10 000 a month.	Corruption	income
	Lastly sample blade design studies are given by specifying a set of input values.	Wind Energy	blade design
	The Convention enjoys broad support: more than 100 member-countries have ratified it, including Belgium.	Corruption	Belgium
	Studies on this concept concluded that it was more cost-effective to use multiple turbines or larger turbines than to pay for the complex structure needed to support.	Wind Energy	turbines, turbines
FastKASSIM	The Commission conducted public consultations in 2010 on the audit policy lessons from the financial crisis.	Corruption	Commission, public, audit, policy, financial crisis
	Belgium ratified this Convention in 2007.	Corruption	Belgium
	Lessons learned from similar experiences in the past	Corruption	No term
	I know that the leaders of a certain country cream something off payments for the supply of commodities.	Corruption	cream something off payments
	For example, a two-bladed rotor with a tail vane would yaw in a series of jerking motions because at the instant the rotor was vertical it offered no centrifugal force resistance to the horizontal movement of the tail vane in following changes in wind direction.	Wind Energy	two-bladed rotor, tail vane, yaw, rotor, centrifugal force, tail vane, wind direction

(b) Retrieved sentences using different similarity metrics.

Table 6: Comparison of query sentence and terms with sentences retrieved in ACTER dataset using different similarity metrics, including BGE-en-large, BGE-en-icl, BM25 and FastKASSIM. The terms extracted from each retrieved sentence are listed alongside their respective domain.

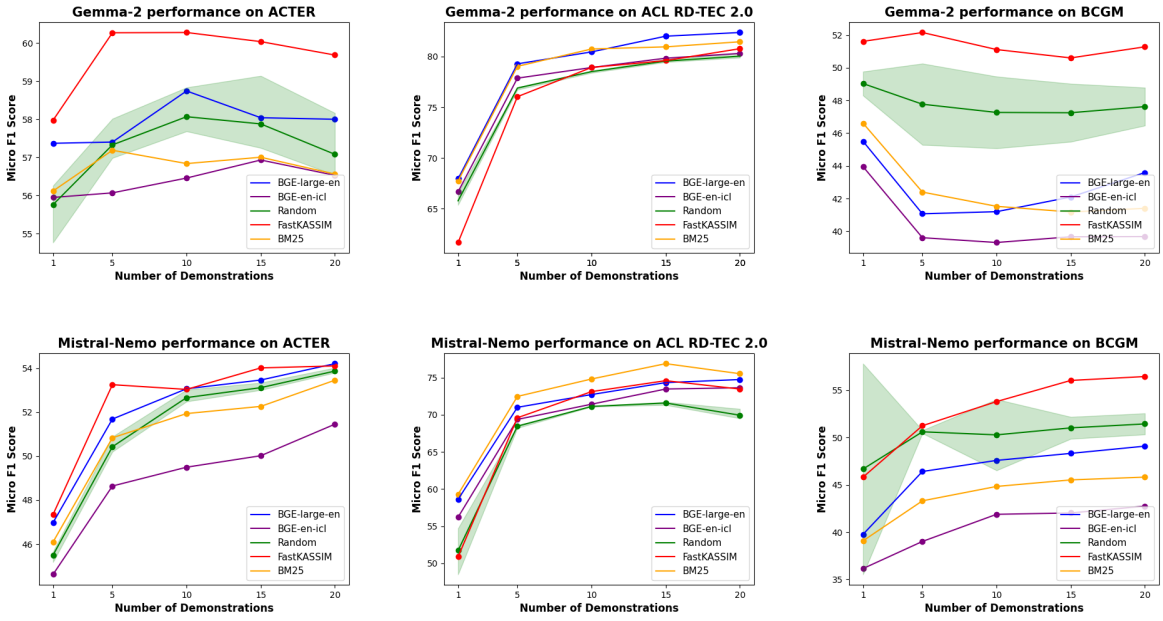

Figure 3: Comparison of retrieval methods for Gemma-2 and Mistral-Nemo on ACTER, ACLR2, and BCGM datasets. The 95% confidence interval is reported for random retrieval.