Benchmarking zero-shot biomedical relation triplet extraction across language model architectures

Frederik Steensgaard Gade[1,2], Ole Lund[2], Marie Lisandra Zepeda Mendoza[3]
[1]AI & Digital Innovation, Novo Nordisk A/S, Måløv, 2760, Denmark
[2]Section for Bioinformatics, Technical University of Denmark, Kgs. Lyngby, 2800, Denmark
[3]Novo Nordisk Research Centre Oxford Ltd, Oxford, OX3 7FZ, UK
Correspondence: fzsg@novonordisk.com

Abstract

Many language models (LMs) in the literature claim excellent zero-shot and/or few-shot capabilities for named entity recognition (NER) and relation extraction (RE) tasks and assert their ability to generalize beyond their training datasets. However, these claims have yet to be tested across different model architectures.

This paper presents a performance evaluation of zero-shot relation triplet extraction (NER followed by RE of the entities) for both small and large LMs, utilizing 13,867 texts from 61 biomedical corpora and encompassing 151 unique entity types. This comprehensive evaluation offers valuable insights into the practical applicability and performance of LMs within the intricate domain of biomedical relation triplet extraction, highlighting their effectiveness in managing a diverse range of relations and entity types.

Gemini 1.5 Pro, the largest LM included in the study, was the top-performing zero-shot model, achieving an average partial match micro F1 of 0.492 for NER, followed closely by SciLitLLM 1.5 14B with a score of 0.475. Fine-tuned models generally outperformed others on the corpora they were trained on, even in a few-shot setting, but struggled to generalize across all datasets with similar entity types. No models achieved an F1 score above 0.5 for the RTE task on any dataset, and their scores fluctuated based on the specific class of entity and the dataset involved. This observation highlights that there is still large room for improvement on the zero-shot utility of LMs in biomedical RTE applications.

1 Introduction

In the field of biomedical natural language processing (NLP), large efforts are being made to create natural language models (LMs) capable of extracting certain entity types and/or relationships, requiring large sets of manually annotated texts. Recently, large language models (LLMs) have proven useful in extracting information from text in a zero-/few-shot fashion, potentially enabling information extraction (IE) where a smaller user-provided annotation may suffice to accomplish the task at hand (Dagdelen et al., 2024). In this study, we focus on biomedical relation triplet extraction (RTE). RTE consists of identifying entities from a list of allowed entity types (such as genes, diseases, etc.) and the type of relationship that exists between them. Thus, RTE can be broken down into a combined named entity recognition (NER) and relation extraction (RE) task. This extraction is valuable for identifying evidence of specific biological connections in, for example, knowledge base (KB) or knowledge graph construction (KGC). Our goal is to investigate the best architectures for reliable biomedical zero-shot RTE to inform model choice for downstream specific biomedical KB question-answering (QA) tasks.

Multiple papers have benchmarked LLMs for IE tasks on biomedical texts (Dai et al., 2024; Jahan et al., 2024; Chen et al., 2025), and there are multiple established combined benchmark datasets (e.g. BLURB (Gu et al., 2021)) and LLM instruction datasets (e.g. SciRIFF (Wadden et al., 2024)), but two main points remain unaddressed:

1. The generalisability of RTE performance outside of the corpora the models are trained on. Performance reporting for the models usually only includes the validation/test set performance for the datasets they were trained on, thus not truly evaluating their generalisability. Performance reporting for some models on certain datasets may also be sensitive to bias through their inclusion in the LLM pretraining (due to the opaqueness of data being used in training of closed-sourced LMs), necessitating performance benchmarking on less commonly used datasets.

2. A direct comparison of zero-shot capabil-

Proceedings of the 24th Workshop on Biomedical Language Processing (BioNLP 2025), pages 88–100
August 1, 2025 ©2025 Association for Computational Linguistics

ities of generative, decoder-only LLMs to the newest BERT-like (and other) LMs for biomedical NER/RE.

We compared the zero-shot RTE performance across various model architectures using a large combined corpus of gold-standard NER & RE annotation datasets outside of the most commonly used benchmark datasets and across multiple architectures.

2 Datasets

To begin with, we assembled an extensive biomedical gold-standard corpus. For this purpose, we compiled a total of 61 different biomedical corpora suited for public and commercial use, including representative subsets from BigBIO (Fries et al., 2022) featuring NER and/or RE annotations, as well as the ComplexTome (Mehryary et al., 2024) and RegulaTome (Nastou et al., 2024) datasets. Altogether, the combined corpus comprises 13,867 texts, including 9,804 abstracts (70.7%), 1,596 sentences (11.5%), and the remaining 2,467 regarded as miscellaneous (such as case reports, full paper paragraphs, etc.) or undefined. Additionally, 18 of the 61 corpora include annotations for 90 distinct relation types. In total, the entire selected corpus comprises 151 distinct entity types, categorised into 11 groups: `Organism`, `Gene/Protein`, `Chemical`, `Disease`, `Medical`, `Gene-related`, `Protein-related`, `Anatomy`, `Other biological`, `Non-English`, and `Other`. Definitions for these groups can be found in appendix C).

Figure 1 characterises the text length, entity count, relation count, and unique entity/relation types within the test set for each corpus included in our study. Details about the modifications made to the corpora are provided in appendix B.

3 Models and methods

The 12 models included in this study are classified into five categories: BERT/BERT-like (Bidirectional Encoder Representations from Transformers), T5 (Text-to-Text Transfer Transformer), KGC-SFT SLM (Knowledge Graph Construction Supervised Fine-Tuned Small Language Model), biology-SFT SLM, and LLM. A comprehensive list of these models, along with their architecture and maximum context length, can be found in table 1.

The BERT models in this study include GLiNER, NuNER, and ZeroShotBioNER. GLiNER (Generalist and Lightweight Model for Named Entity Recognition) (Zaratiana et al., 2023) is a small, generalist NER model, introduced as an alternative to traditional NER models. Unlike conventional models, GLiNER is not restricted to predefined entities, even though it employs a BERT-like architecture.

Building on GLiNER, GLiNER Multi-task (Stepanov and Shtopko, 2024) extends the capabilities of the model to perform additional information extraction tasks, such as RE and summarisation.

NuNER (Bogdanov et al., 2024) is another generalist alternative to GLiNER, distinguished by its training method, which employs a contrastive learning approach on synthetic data generated by an LLM (GPT-3.5).

ZeroShotBioNER (Košprdić et al., 2024) is a BERT-based model, specifically a fine-tuned version of BioBERT v1.1, trained on 26 biomedical NER classes. It is designed for zero-shot inference across the biomedical domain, particularly targeting chemicals, diseases, and proteins, and is tailored for biological applications.

InstructUIE (Wang et al., 2023) utilizes a T5 architecture and is trained and evaluated on their own curated information extraction benchmark set. This set includes NER datasets from AnatEM, BC5CDR, CHEMDNER, among others, encompassing a wide range of information extraction tasks.

The two KGC-SFT SLMs, Triplex (SciPhi, 2024) and Phi3 Mini Graph (Emergent Methods, 2024), are fine-tuned versions of Phi3 models specifically designed for generalist RTE.

SciLitLLM 1.5 (Li et al., 2024) is built upon Qwen 2.5 and undergoes continuous pre-training using an internal corpus comprising science textbooks and articles. It is subsequently fine-tuned on SciRIFF (Wadden et al., 2024) as well as a synthetic dataset designed for scientific literature understanding and instructions.

As the representative decoder-only, closed-source LLM, we chose Gemini 1.5 Pro (Gemini Team et al., 2024) due to its computational efficiency. It presents itself as having exceptional ability in long-context needle-in-a-haystack retrieval and demonstrates strong overall performance across a diverse array of tasks.

Models were configured to perform NER and RTE of all applicable types in a single model call, wherever supported. For GLiNER multi-task models and InstructUIE, NER and RE were conducted in two separate model calls. All models were employed at the document level. Details about the

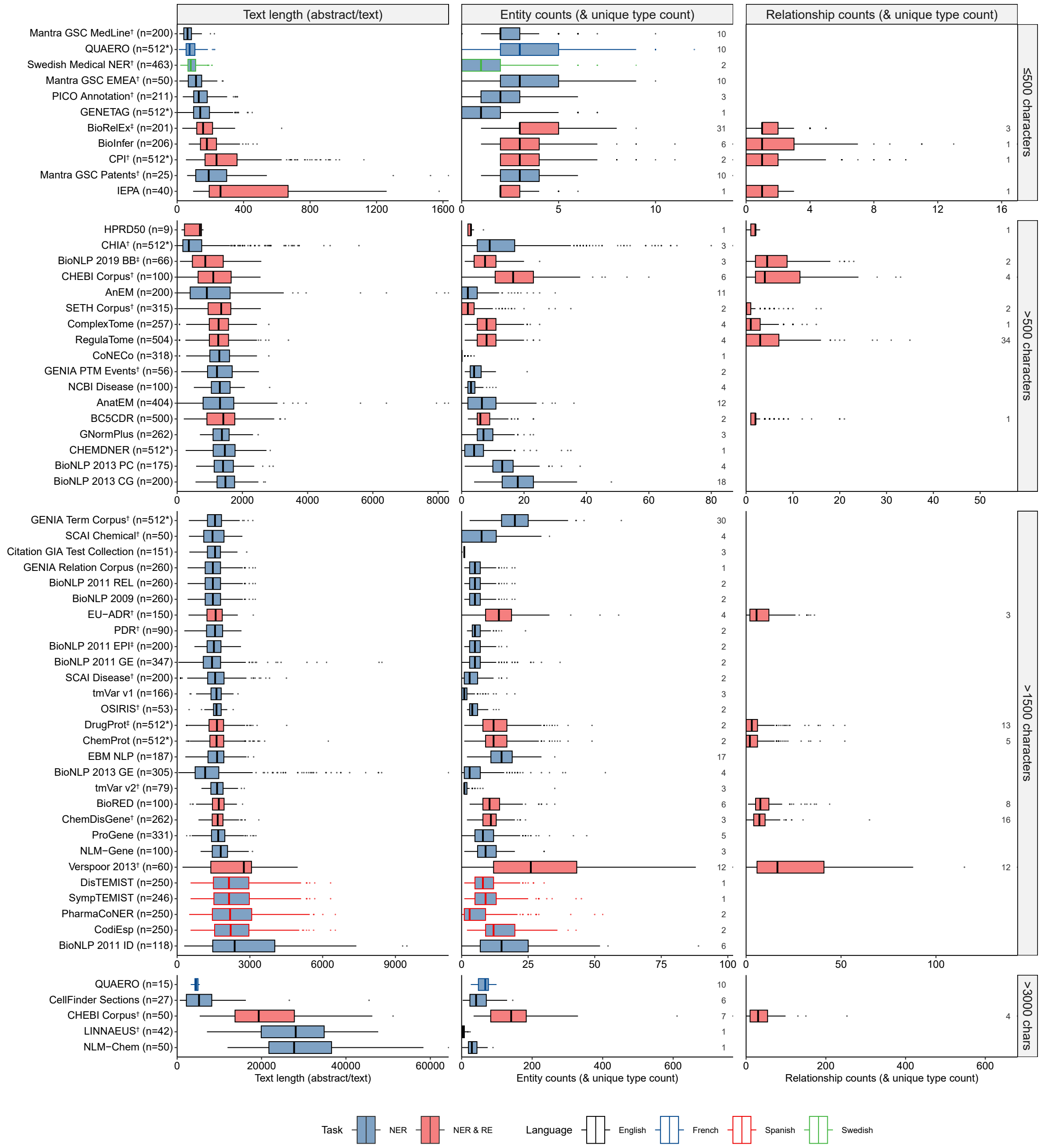

Figure 1: Summary statistics of the text lengths, entity counts (& number of unique entity types), and relation counts (& number of unique relation types) for the test set of each corpus used. The corpora categorised based on their average character count (≤500, >500, >1500, >3000). Details regarding which entity types and relationship types were included, excluded, or merged, can be found in appendix B. n denotes the number of texts in the corpus.

* Datasets were truncated to a maximum of 512 samples to minimise over-representation of certain datasets within the overall corpus.

† For these 20 datasets, no splits were available via BigBIO, therefore, we used the test set from a 50/50 train/test split.

‡ For the BioNLP 2011 EPI, BioNLP 2019 BB, BioRelEx, and DrugProt datasets, annotated test sets were not available, so their development/validation sets were utilised instead.

Architecture	Model	Context length limit	Tasks
LLM	Gemini 1.5 Pro (Feb 2025)	2,097,152	NER & RE
Biology-SFT SLM	SciLitLLM 1.5 (Qwen 2.5 14B)	131,072	NER & RE
KGC-SFT SLM	Triplex (Phi3-3.8B)	131,072	NER & RE
	Phi3 Mini Graph (Phi3-3.8B-128K)	131,072	NER & RE
T5	InstructUIE (Flan-T5 11B)	512	NER & RE
BERT/BERT-like	ZeroShotBioNER (BioBERT V1.1)	512	NER
	NuNER Zero 4K (Longformer Large 4K)	4,096	NER
	GLiNER Medium v2.5 (DeBERTa-V3)	384	NER
	GLiNER Large v2.5 (DeBERTa-V3-Large)	512	NER
	GLiNER Large Bio v0.1 (DeBERTa-V3-Large)*	512	NER
	GLiNER Multi-task v1.0 (DeBERTa-V2-XLarge)	512	NER & RE
	GLiNER Multi-task Large v0.5 (DeBERTa-V3-Large)	512	NER & RE

Table 1: For each model, the table includes its name, model group, token limits for both prompt/input and completion/output, and the tasks each model can perform—specifically NER and RE. Note that the context length limit reflects the maximum number of tokens the architecture can process simultaneously, rather than a verified range for optimal performance.
* This model is not included in the main GLiNER publication by Zaratiana et al. (2023), but is available on HuggingFace (repo_id: `urchade/gliner_large_bio-v0.1`).

prompt/input preparation for each model are provided in appendix D.

Formally, we define zero-shot RTE as the process of performing NER followed by RE, given only the allowed entity and relation types. For k-shot RTE, we give k examples from the training set. If no relations were annotated for a given corpus, only the NER task was evaluated.

To mitigate the possibility of hallucinations from the language models, the output was limited to the queried types of entities and relations. Given that all models, except the BERT variants, are causal language models (as opposed to token classifiers), they produce entity name strings rather than token positions. Consequently, to ensure fairness, performance for all models was evaluated using the case-insensitive micro F1 score from MUC-5 (Chinchor and Sundheim, 1993), unless stated otherwise[1], partial boundary, exact-type matching for each unique entity and relationship in the gold-standard data. In this context, a partial match refers to a word match at either boundary. Therefore, the reported performance more closely aligns with the practical application for KGC, where duplicate entities and relationships are consolidated.

4 Results

All models, except the KGC SFT-SLMs, are evaluated across all datasets, with the exception of BC5CDR, BioRED, and ChemDNER; these particular datasets are analyzed separately because some of the models have been fine-tuned specifically using these datasets.

Figure 2 displays the NER rank distribution for each corpus, providing a head-to-head comparison of the models. Additionally, the win rates for NER and RTE are detailed in appendix table A1. Gemini 1.5 Pro and the notably smaller SciLitLLM 14B emerge as the clear frontrunners, whereas InstructUIE and ZeroShotBioNER are the lowest performers overall. However, ZeroShotBioNER excels over all other models in the ChemProt, DrugProt, CHEBI, ChemDisGene, and SETH corpora, which predominantly contain chemical, gene/protein, and disease annotations. Similarly, InstructUIE outperforms all other models in the Citation GIA Test, IEPA, and GENETAG corpora, which exclusively feature gene and protein annotations.

Although ZeroShotBioNER and InstructUIE outperform other models in the specific datasets mentioned, this is not generally the case across the entity types they were fine-tuned on. This is evident in figure 3, which illustrates NER performance by entity type group. Note the two models generally demonstrate lower performance for gene/protein, chemical, and disease entity groups. Moreover, despite being trained on biological entity types, InstructUIE and ZeroShotBioNER do not generalize well to other biological or gene-/protein-related entity types. One might hypothesize that identifying gene-/protein-related entity types parallels the

[1] Due to capitalised words in the beginning of sentences being considered identical to non-capitalised words for the purposes of entity uniqueness.

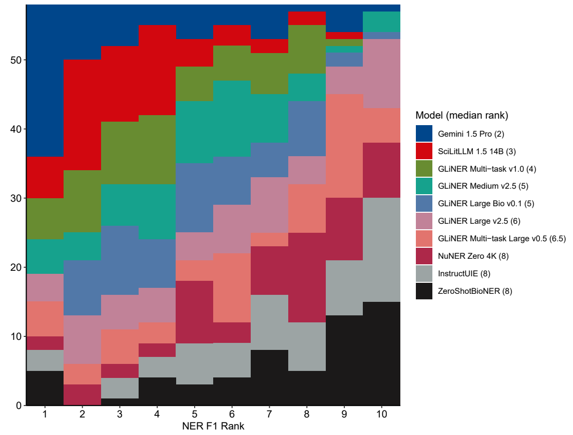

Figure 2: Model ranks of NER micro F1 for all corpora (excluding BC5CDR, BioRED, and ChemDNER).

task of identifying entity relations, which models enabled for RE might excel at.

The NER and RTE performance by corpus mean character count (≤500, >500, >1500, >3000) is shown in figure 4. For models with a short context, the input might be truncated and thus the recall is decreased inherently as a result of the model architecture. However, even for the long-context models, the F1 drops for the longest input texts. In lengthy corpora, RTE performance drops to nearly zero, and across shorter corpora, the general performance for this task remains quite poor across all models.

The performance of ZeroShotBioNER and InstructUIE on BC5CDR, BioRED, and ChemDNER (which were excluded from the previous analyses) is compared with zero-/few-shot prompting of Gemini 1.5 Pro and SciLitLLM 1.5 14B in table 2. Few-shot examples were sourced from the training set.

The KGC SFT-SLMs were evaluated separately on a small subset of datasets, specifically BC5CDR and BioRED, as detailed in table 3. For both datasets, the NER performance of the KGC models is lower than that of all other models, particularly for the more complex dataset, BioRED. Although these models are intended for generalist KGC, their performance falls significantly below that of SciLitLLM 1.5 and Gemini 1.5 Pro (table 2). This discrepancy may be attributed to the lack of biomedical data in their fine-tuning process.

Appendix table A2 compares the partial and strict matching performance of the top three models: Gemini 1.5 Pro, SciLitLLM 1.5 14B, and GLiNER Multi-task v1.0; alongside the two SFT IE models, InstructUIE and ZeroShotBioNER. Gemini 1.5 Pro experiences the largest performance drop when evaluation criteria shift to strict matching. This is due to certain instances, like the one in BC5CDR, where "methamphetamine induces psychosis" is incorrectly labeled as "methamphetamine psychosis" instead of the correct "psychosis." This labeling would be correct under partial matching but incorrect under strict matching. GLiNER Multi-task v1.0 demonstrates the smallest performance loss for NER, achieving the highest F1 score and precision under strict matching conditions. Conversely, SciLitLLM 1.5 14B exhibits the least performance decline when transitioning to strict matching, and even shows an improvement in precision.

5 Discussion

The models explicitly fine-tuned for biology, namely InstructUIE, ZeroShotBioNER, SciLitLLM 1.5, and GLiNER Large Bio v0.1, were generally outperformed by the larger, more generalist models. Exceptions occurred for datasets on which these models were directly fine-tuned or those containing very similar entity types. However, InstructUIE and ZeroShotBioNER did not consistently outperform all other models across datasets featuring entity types similar to those in their fine-tuning datasets. The KGC-specific models demonstrated significantly lower performance compared to other models, possibly due to their lack of biological understanding needed to identify entity and relation types. Overall, Gemini, the largest and most resource-intensive model, achieved the highest scores in the benchmark. Notably, Gemini's performance was only marginally better than the considerably smaller SciLitLLM 1.5, which has 14 billion parameters, in zero-shot biomedical NER, although SciLitLLM had lower RTE performance. We hypothesize that a model fine-tuned on biology and further instruction-tuned specifically for RTE could achieve even better results.

SciLitLLM 1.5 14B was specifically fine-tuned on the literature understanding instruction dataset SciRIFF (Wadden et al., 2024), which includes NER tasks for several of our datasets, such as BioRED and GNormPlus, as well as RE tasks for ChemProt. This may introduce a bias in the performance evaluation.

The best-performing BERT model was GLiNER Multi-task v1.0, which also achieved the best av-

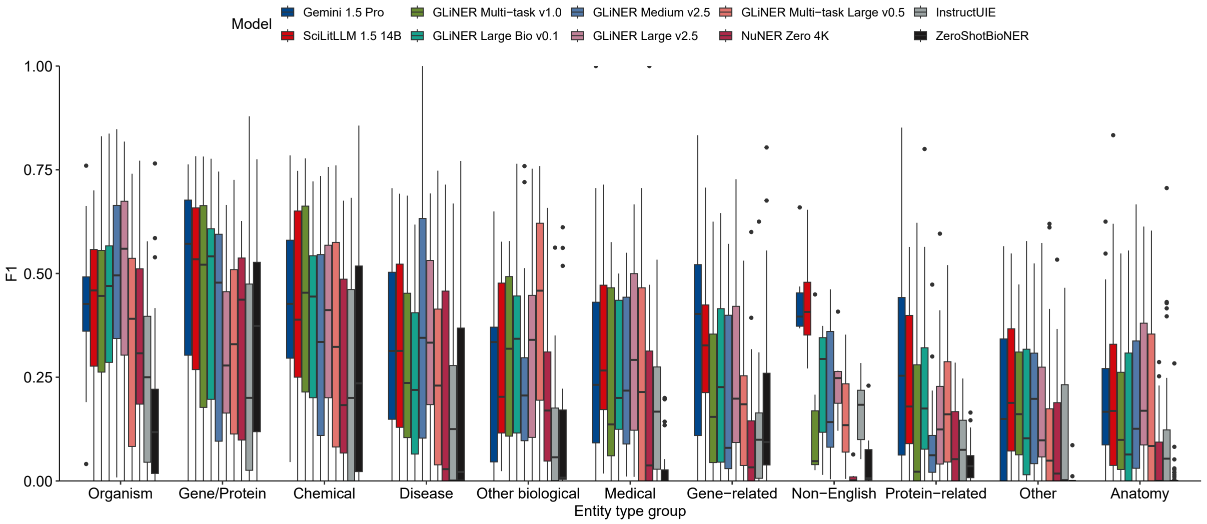

Figure 3: NER F1 scores for each entity type prediction across all datasets (excluding BC5CDR, BioRED, and ChemDNER) stratified by model and entity group. Information on which entity types were grouped is specified in appendix C.

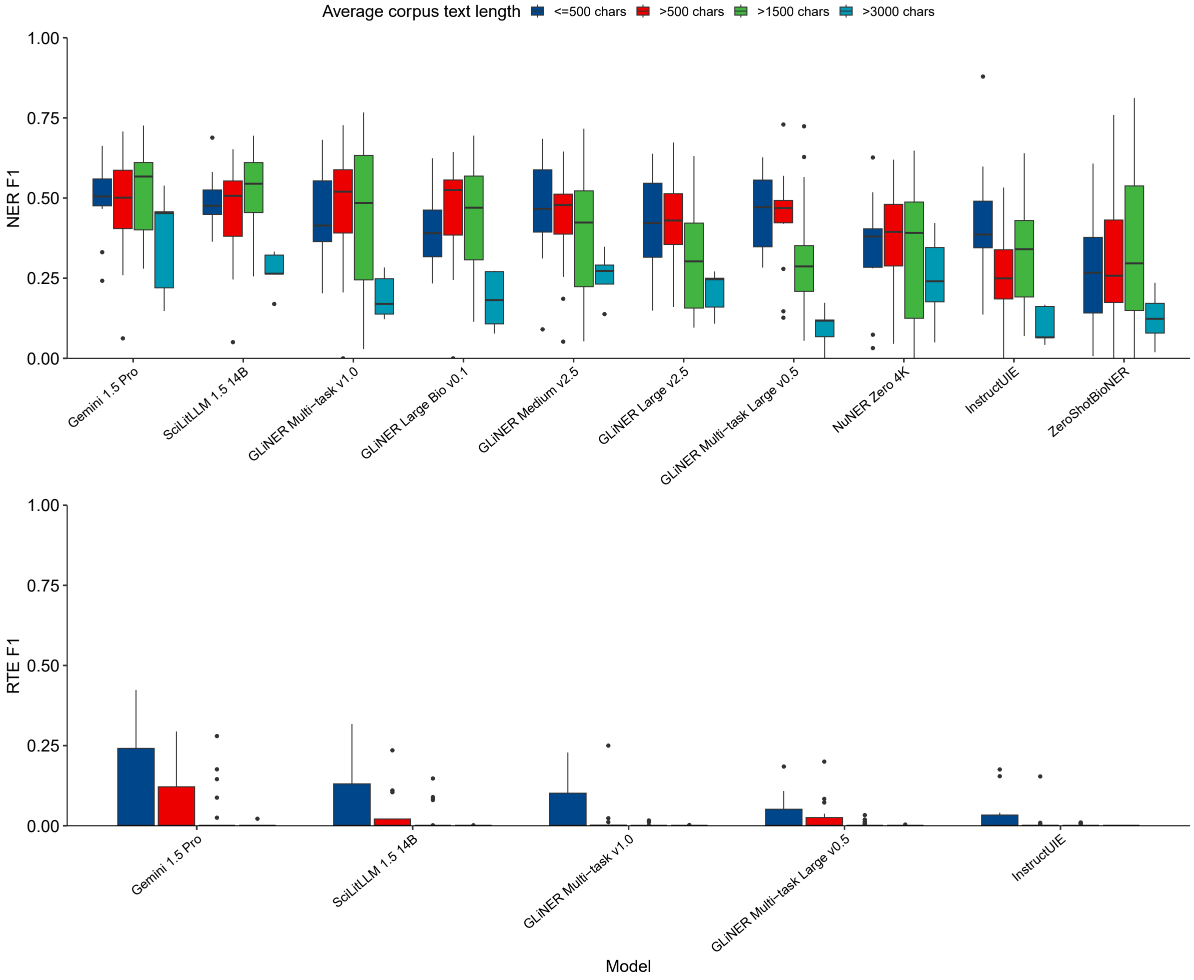

Figure 4: NER and RTE micro F1 score for each corpus across all datasets and (excluding BC5CDR, BioRED, and ChemDNER) stratified by model and average text length.

Dataset	Model	k-shot	NER			RTE		
			F1	Precision	Recall	F1	Precision	Recall
BC5CDR	ZeroShotBioNER*	SFT	**0.847**	0.777	**0.931**	-	-	-
	InstructUIE	SFT	0.601	0.790	0.485	0.105	0.214	0.070
	Gemini 1.5 Pro	0-shot	0.583	**0.842**	0.446	*0.442*	*0.477*	0.413
		3-shot	0.666	0.826	0.558	0.438	0.458	*0.419*
		10-shot	0.717	*0.836*	*0.627*	**0.497**	**0.488**	**0.507**
	SciLitLLM 1.5 14B	0-shot	0.697	0.796	0.620	0.340	0.446	0.274
		3-shot	0.723	0.811	0.653	0.381	0.471	0.320
		10-shot	*0.738*	0.785	0.696	0.400	0.443	0.364
BioRED	ZeroShotBioNER*	SFT	0.666	0.685	**0.648**	-	-	-
	InstructUIE	No SFT	0.265	0.666	0.165	0.002	0.045	0.001
	Gemini 1.5 Pro	0-shot	0.516	0.725	0.400	0.138	0.232	0.098
		3-shot	*0.669*	*0.755*	0.600	*0.162*	*0.232*	*0.125*
		10-shot	**0.684**	**0.779**	*0.610*	**0.183**	**0.266**	**0.139**
	SciLitLLM 1.5 14B	0-shot	0.607	0.651	0.569	0.021	0.094	0.012
		3-shot	0.600	0.641	0.564	0.057	0.105	0.039
		10-shot	0.622	0.678	0.574	0.085	0.159	0.058
ChemDNER*	ZeroShotBioNER*	SFT	**0.866**	**0.944**	**0.800**	-	-	-
	InstructUIE	SFT	0.658	0.865	0.532	-	-	-
	Gemini 1.5 Pro	0-shot	0.684	0.713	0.657	-	-	-
		3-shot	0.652	0.803	0.549	-	-	-
		10-shot	0.690	0.781	0.619	-	-	-
	SciLitLLM 1.5 14B	0-shot	0.755	0.755	*0.755*	-	-	-
		3-shot	*0.794*	0.878	0.725	-	-	-
		10-shot	0.792	*0.889*	0.714	-	-	-

Table 2: Comparison of model performance of fine-tuned models, ZeroShotBioNER and InstructUIE, with the zero-/few-shot performance of the LLM, Gemini 1.5 Pro, and the biology-SFT SLM, SciLitLLM 1.5 14B. Both ZeroShotBioNER and InstructUIE were fine-tuned on BC5CDR and ChemDNER (denoted with SFT in the table), and ZeroShotBioNER was additionally fine-tuned on BioRED, whilst InstructUIE was not (No SFT). Best performance by dataset is highlighted in bold, and second-best in italics.
* NER-only model/dataset.

Dataset	Matching criteria	Model	NER			RTE		
			F1	Precision	Recall	F1	Precision	Recall
BC5CDR	Partial (strict type)	Triplex	0.458	0.380	**0.576**	0.132	0.242	0.090
		Phi3 Mini Graph*	0.545	0.698	0.448	-	-	-
		GLiNER Multi-task v1.0	**0.611**	**0.771**	0.505	**0.162**	**0.500**	**0.097**
	Relaxed	Triplex	0.486	0.407	0.605	0.121	0.223	0.083
		Phi3 Mini Graph	0.482	0.412	0.581	0.096	0.058	0.290
		GLiNER Multi-task v1.0	0.612	0.770	0.507	0.160	0.493	0.095
BioRED	Partial (strict type)	Triplex	0.015	0.529	0.007	0.002	0.200	0.001
		Phi3 Mini Graph*	0.096	0.295	0.058	-	-	-
		GLiNER Multi-task v1.0	**0.575**	**0.662**	**0.508**	**0.004**	**0.143**	**0.002**
	Relaxed	Triplex	0.014	0.529	0.007	0.002	1.000	0.001
		Phi3 Mini Graph	0.509	0.478	0.544	0.167	0.141	0.205
		GLiNER Multi-task v1.0	0.599	0.691	0.529	0.007	0.219	0.003

Table 3: KGC-SFT SLM performances vs. GLiNER Multi-task performance for NER and RTE with partial, strict-type matching criteria (used through the paper) and relaxed matching (case-insensitive, no schema restriction of output, entity and relation type-agnostic, relation directionality-agnostic). For comparison, Gemini 1.5 Pro 0-shot F1 for RTE in BioRED with relaxed matching criteria is 0.287. Best performance by dataset is highlighted in bold.
* The Phi3 Mini Graph model is unable to follow the instruction to output only specified relation types, and thus restricting the output to the specified schema yields no predictions.

erage performance for strict matching. It is significantly smaller than either SciLitLLM 1.5 14B or Gemini 1.5 Pro, potentially making it the ideal choice when cost and scalability are concerns.

Notably, while SciLitLLM 1.5 Pro was the overall best-performing model among the ones compared, RTE performance was relatively low across the board. No zero-shot model achieved a micro F1 score above 0.5 for any dataset, raising concerns about their effectiveness for RTE tasks. In agreement with Chen et al. (2025), we therefore do not recommend using zero-shot models for biomedical RTE. Although few-shot performance can be comparable to SFT performance for certain models and datasets, fine-tuned models generally outperform non-fine-tuned ones when manually annotated data is available for SFT. In cases where such data is unavailable, few-shot models may be utilized if downstream tasks can accommodate a compromise in performance, possibly due to additional checks at later stages.

While the RTE task yields a simple KG without additional metadata, leveraging information extraction models such as InstructUIE, NuExtract 1.5, and LLMs like Gemini 1.5 Pro could enhance the metadata associated with the triplets. In a biomedical context, this could involve incorporating surrounding biological context such as tissue, organism, intervention, and co-factors. Such contextual enrichment can be done with traditional NLP methods, and could be improved with powerful generalist LLMs (Sosa et al., 2023).

Although some benchmarking datasets are extensive and well-annotated across a wide range of relationships and entities, they present challenges when used to generate KBs or KGs. For instance, RegulaTome includes relationships that are speculative or hypothesized and does not account for the negation of relations. Consequently, using these annotations as the truth set means there is no distinction between verified conclusions and mere speculations—only their mention in the text is captured, while negative results are omitted.

We observe that methods such as GraphRAG (Edge et al., 2024), attempt to leverage the emergent information extraction capabilities of LLMs to enhance knowledge base question answering (KBQA) tasks. However, based on the outcomes of this benchmark, we hypothesize that for results from a GraphRAG-like approach to be valuable in biomedical applications, tailored models are necessary to accurately tag relevant entities and relationships. This is due to the fact that the inherent biological understanding of zero-shot LLMs is typically insufficient for most practical downstream applications.

6 Conclusion

In conclusion, this study benchmarks zero-shot biomedical RTE across a range of LM architectures. Larger models such as Gemini 1.5 Pro and SciLitLLM 1.5 14B excel in NER but face challenges with subsequent RE, with no F1 score surpassing 0.5 in RTE tasks. Notably, GLiNER Multi-task v1.0 stands out as the best-performing BERT-based model, delivering strong performance relative to its smaller size and excelling in strict matching criteria, thus making it a cost-effective option when scalability is a concern.

While fine-tuned models like ZeroShotBioNER perform well on specific datasets, they are generally surpassed by larger, more generalized models even when dealing with slightly out-of-distribution data, underscoring the limitations of current zero-shot models for practical applications in biomedical NLP. Furthermore, although few-shot learning provides some benefits, fine-tuning remains essential for maximizing model performance when annotation is feasible.

Limitations

Conducting a fair evaluation of all available LMs is a challenging task for several reasons. Firstly, accessing and comprehensively testing each model may not be financially viable, necessitating the selection of representative models from various LM categories. Additionally, information regarding the training data is not always publicly available, as seen with Gemini, or models may be trained on known public benchmarks like BLURB, which includes datasets that overlap with our benchmark (EBM PICO, ChemProt, and BC5CDR) or contain shared entity types (JNLPBA) (Gu et al., 2021), thus complicating the fair comparison between models.

Moreover, performance is sensitive to the matching criteria employed, and the options for this benchmark are restricted due to the nature of the model outputs from causal language models, as they are not token classifiers. More sophisticated matching criteria, such as ontology matching, would be preferable but fall outside the scope of this research.

The models are constrained by their context length, and some might have benefited from re-engineering the task by breaking the texts into sentences—even models with a relatively long context length. Additionally, running the models in multiple rounds, such as one round per entity type, could offer advantages, like increased task specificity. However, this approach also presents drawbacks, including overlap issues and higher costs.

Finally, it is important to recognize that different models may require distinct prompts to achieve optimal performance. Studies have demonstrated that benchmark results are sensitive to prompt engineering (Jahan et al., 2023). Exploring techniques such as chain-of-thought prompting, meta-prompting (Suzgun and Tauman Kalai, 2024), reasoning models (DeepSeek-AI et al., 2025), or other related strategies could potentially enhance performance.

Acknowledgments

The work of FSG is partly funded by the Innovation Fund Denmark (IFD) under File No. 3129-00056 and co-financed through a Novo Nordisk R&ED novoSTAR Industrial PhD fellowship.

Additionally, thanks to Jesper Ferkinghoff-Borg, Julien Fauqueur, and Robert R. Kitchen for their help and support for this project.

References

Sergei Bogdanov, Alexandre Constantin, Timothée Bernard, Benoit Crabbé, and Etienne Bernard. 2024. NuNER: Entity Recognition Encoder Pre-training via LLM-Annotated Data. Computing Research Repository, arXiv:2402.15343. Version 1.

Qingyu Chen, Yan Hu, Xueqing Peng, Qianqian Xie, Qiao Jin, Aidan Gilson, Maxwell B. Singer, Xuguang Ai, Po-Ting Lai, Zhizheng Wang, Vipina K. Keloth, Kalpana Raja, Jimin Huang, Huan He, Fongci Lin, Jingcheng Du, Rui Zhang, W. Jim Zheng, Ron A. Adelman, Zhiyong Lu, and Hua Xu. 2025. Benchmarking large language models for biomedical natural language processing applications and recommendations. *Nature Communications*, 16(1).

Nancy Chinchor and Beth Sundheim. 1993. MUC-5 evaluation metrics. In *Fifth Message Understanding Conference (MUC-5): Proceedings of a Conference Held in Baltimore, Maryland, August 25-27, 1993*.

John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S. Rosen, Gerbrand Ceder, Kristin A. Persson, and Anubhav Jain. 2024. Structured information extraction from scientific text with large language models. *Nature Communications*, 15(1).

Xiang Dai, Sarvnaz Karimi, Abeed Sarker, Ben Hachey, and Cecile Paris. 2024. Multiade: A multi-domain benchmark for adverse drug event extraction. *Journal of Biomedical Informatics*, 160:104744.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiashi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanjia Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. Computing Research Repository, arXiv:2501.12948. Version 1.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. 2024. From Local to Global: A Graph RAG Approach to Query-Focused

Summarization. Computing Research Repository, arXiv:2404.16130. Version 2.

Emergent Methods. 2024. Outperforming Claude 3.5 Sonnet with Phi-3-mini-4k for graph entity relationship extraction tasks — emergentmethods.medium.com. https://emergentmethods.medium.com/outperforming-claude-3-5-sonnet-with-phi-3-mini-4k-for-graph-entity-relationship-extraction-tasks-7c8f6c1ebd79. [Accessed 20-03-2025].

Jason Alan Fries, Leon Weber, Natasha Seelam, Gabriel Altay, Debajyoti Datta, Samuele Garda, Myungsun Kang, Ruisi Su, Wojciech Kusa, Samuel Cahyawijaya, Fabio Barth, Simon Ott, Matthias Samwald, Stephen Bach, Stella Biderman, Mario Sänger, Bo Wang, Alison Callahan, Daniel León Periñán, Théo Gigant, Patrick Haller, Jenny Chim, Jose David Posada, John Michael Giorgi, Karthik Rangasai Sivaraman, Marc Pàmies, Marianna Nezhurina, Robert Martin, Michael Cullan, Moritz Freidank, Nathan Dahlberg, Shubhanshu Mishra, Shamik Bose, Nicholas Michio Broad, Yanis Labrak, Shlok S Deshmukh, Sid Kiblawi, Ayush Singh, Minh Chien Vu, Trishala Neeraj, Jonas Golde, Albert Villanova del Moral, and Benjamin Beilharz. 2022. BigBIO: A Framework for Data-Centric Biomedical Natural Language Processing. Part of Advances in Neural Information Processing Systems 35 (NeurIPS 2022) Datasets and Benchmarks Track.

Google Gemini Team, Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry, Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, James Molloy, Jilin Chen, Michael Isard, Paul Barham, Tom Hennigan, Ross McIlroy, Melvin Johnson, Johan Schalkwyk, Eli Collins, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Clemens Meyer, Gregory Thornton, Zhen Yang, Henryk Michalewski, Zaheer Abbas, Nathan Schucher, Ankesh Anand, Richard Ives, James Keeling, Karel Lenc, Salem Haykal, Siamak Shakeri, Pranav Shyam, Aakanksha Chowdhery, Roman Ring, Stephen Spencer, Eren Sezener, Luke Vilnis, Oscar Chang, Nobuyuki Morioka, George Tucker, Ce Zheng, Oliver Woodman, Nithya Attaluri, Tomas Kocisky, Evgenii Eltyshev, Xi Chen, Timothy Chung, Vittorio Selo, Siddhartha Brahma, Petko Georgiev, Ambrose Slone, Zhenkai Zhu, James Lottes, Siyuan Qiao, Ben Caine, Sebastian Riedel, Alex Tomala, Martin Chadwick, Juliette Love, Peter Choy, Sid Mittal, Neil Houlsby, Yunhao Tang, Matthew Lamm, Libin Bai, Qiao Zhang, Luheng He, Yong Cheng, Peter Humphreys, Yujia Li, Sergey Brin, Albin Cassirer, Yingjie Miao, Lukas Zilka, Taylor Tobin, Kelvin Xu, Lev Proleev, Daniel Sohn, Alberto Magni, Lisa Anne Hendricks, Isabel Gao, Santiago Ontanon, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. Computing Research Repository, arXiv:2403.05530. Version 5.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1).

Israt Jahan, Md Tahmid Rahman Laskar, Chun Peng, and Jimmy Huang. 2023. Evaluation of ChatGPT on biomedical tasks: A zero-shot comparison with fine-tuned generative transformers. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 326–336, Toronto, Canada. Association for Computational Linguistics.

Israt Jahan, Md Tahmid Rahman Laskar, Chun Peng, and Jimmy Xiangji Huang. 2024. A comprehensive evaluation of large language models on benchmark biomedical text processing tasks. *Computers in Biology and Medicine*, 171:108189.

Miloš Košprdić, Nikola Prodanović, Adela Ljajić, Bojana Bašaragin, and Nikola Milošević. 2024. From zero to hero: Harnessing transformers for biomedical named entity recognition in zero- and few-shot contexts. *Artificial Intelligence in Medicine*, 156:102970.

Sihang Li, Jin Huang, Jiaxi Zhuang, Yaorui Shi, Xiaochen Cai, Mingjun Xu, Xiang Wang, Linfeng Zhang, Guolin Ke, and Hengxing Cai. 2024. SciLitLLM: How to Adapt LLMs for Scientific Literature Understanding. Computing Research Repository, arXiv.2408.15545. Version 5.

Farrokh Mehryary, Katerina Nastou, Tomoko Ohta, Lars Juhl Jensen, and Sampo Pyysalo. 2024. STRING-ing together protein complexes: corpus and methods for extracting physical protein interactions from the biomedical literature. *Bioinformatics*, 40(9).

Katerina Nastou, Farrokh Mehryary, Tomoko Ohta, Jouni Luoma, Sampo Pyysalo, and Lars Juhl Jensen. 2024. RegulaTome: a corpus of typed, directed, and signed relations between biomedical entities in the scientific literature. *Database*, 2024.

SciPhi. 2024. Triplex — SOTA LLM for Knowledge Graph Construction - SciPhi AI — sciphi.ai. https://www.sciphi.ai/blog/triplex. [Accessed 20-03-2025].

Daniel N. Sosa, Rogier Hintzen, Betty Xiong, Alex de Giorgio, Julien Fauqueur, Mark Davies, Jake Lever, and Russ B. Altman. 2023. Associating biological context with protein-protein interactions through text mining at pubmed scale. *Journal of Biomedical Informatics*, 145:104474.

Ihor Stepanov and Mykhailo Shtopko. 2024. GLiNER multi-task: Generalist Lightweight Model for Various Information Extraction Tasks. Computing Research Repository, arXiv.2406.12925. Version 2.

Mirac Suzgun and Adam Tauman Kalai. 2024. Meta-Prompting: Enhancing Language Models with Task-Agnostic Scaffolding. Computing Research Repository, arXiv:2401.12954. Version 1.

David Wadden, Kejian Shi, Jacob Morrison, Aakanksha Naik, Shruti Singh, Nitzan Barzilay, Kyle Lo, Tom Hope, Luca Soldaini, Shannon Zejiang Shen, Doug Downey, Hannaneh Hajishirzi, and Arman Cohan. 2024. SciRIFF: A Resource to Enhance Language Model Instruction-Following over Scientific Literature. Computing Research Repository, arXiv:2406.07835. Version 3.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, Jihua Kang, Jingsheng Yang, Siyuan Li, and Chunsai Du. 2023. InstructUIE: Multi-task Instruction Tuning for Unified Information Extraction. Computing Research Repository, arXiv:2304.08085. Version 1.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2023. GLiNER: Generalist Model for Named Entity Recognition using Bidirectional Transformer. Computing Research Repository, arXiv:2311.08526. Version 1.

A Supplementary figures/tables

Model	**NER win-rate**	**RTE win-rate**
Gemini 1.5 Pro	**72.8%**	**89.1%**
SciLitLLM 1.5 14B	**72.8%**	*64.1%*
GLiNER Multi-task v1.0	62.8%	35.9%
GLiNER Large Bio v0.1	55.0%	-
GLiNER Medium v2.5	54.8%	-
GLiNER Large v2.5	46.2%	-
GLiNER Multi-task Large v0.5	40.6%	43.8%
NuNER Zero 4K	34.7%	-
InstructUIE	31.2%	17.2%
ZeroShotBioNER	29.1%	-

Table A1: Model micro F1 win rates in all head-to-head comparisons per dataset (for both NER and RTE, excluding BC5CDR, BioRED, and ChemDNER). Best performance is highlighted in bold,and second-best in italics.

B Dataset modifications

To align the datasets to the same tasks, the relationship type names were renamed to an active form (e.g. `COMPLEX_FORMATION` → `FORMS_COMPLEX_WITH`). Selected entities and relationships were removed, if they were not deemed relevant for the task (such as `part-of` relations). All relation types were capitalised, and all entity types were in PascalCase.

BigBIO dataset import modifications: BioRelEx (to include type of binding: `binds`, `not-binds`, `inconclusively-binds`), ComplexTome (implemented), ProGene (changed splitting to original split), RegulaTome (implemented).

C Entity group definitions

`Organism` `(n=10)` cell, cellline, celltype, livingbeing, microorganism, monocell, organism, organismtaxon, plant, species

`Gene/Protein` `(n=18)` dna, dnafamilyorgroup, gene, geneorgeneproduct, geneormolecularsequence, geneorprotein, geneorproteinfamily, geneorproteinorrna, geneproductormarkergene, geneprotein, peptide, protein, proteinenum, proteinfamiliyorgroup, proteinfamily, proteinfamilyorgroup, proteinisoform, proteinmolecule

`Chemical` `(n=17)` aminoacid, aminoacidmonomer, atom, carbohydrate, chemical, chemicalabbreviation, chemicalentity, chemicalfamily, chemicalordrug, chemicalstructure, compound, drug, metabolite, nucleotide, partchemical, reagent, simplechemical

`Disease` `(n=20)` adverseeffect, compositediseasemention, condition, disease, diseaseclass, diseaseordisorder, diseaseorphenotypicfeature, disorder, disorderfinding, outcome, outcomeadverseeffects, outcomemental, outcomemortality, outcomeother, outcomepain, outcomephysical, participantcondition, phenomena, phenotype, specificdisease

`Medical` `(n=11)` assay, device, diseasemodifier, intervention, interventioneducational, interventionother, interventionpharmacological, interventionphysical, interventionpsychological, interventionsurgical, procedure

`Gene-related` `(n=16)` dnadomainorregion, dnamolecule, dnamutation, dnasubstructure, geneticvariant, mutation, polynucleotide, regulonoperon, rna, rnadomainorregion, rnafamilyorgroup, rnamolecule, sequencevariant, snp, snporsequencevariation, twocomponentsystem

`Protein-related` `(n=13)` complex, fusionprotein, proteincomplex, proteindomain, proteindomainorregion, proteinmotif, proteinmutation, proteinregion, proteinrelatedentity, proteinrnacomplex, proteinsubstructure, proteinsubunit, proteinvariant

`Anatomy` `(n=13)` anatomicalsystem, anatomy, bodypart, bodystructure, developinganatomicalstructure, immaterialanatomicalentity, mul-

Model	Matching criteria	NER			RTE		
		F1	Precision	Recall	F1	Precision	Recall
Gemini 1.5 Pro	Partial (strict type)	0.492	0.611	0.457	0.204	0.236	0.200
	Strict	0.386	0.472	0.365	0.030	0.214	0.016
		-22%	-23%	-20%	-85%	-9%	-92%
SciLitLLM 1.5 14B	Partial (strict type)	0.475	0.541	0.487	0.105	0.232	0.074
	Strict	0.427	0.486	0.440	0.043	0.284	0.024
		-10%	-10%	-10%	**-59%**	**+22%**	**-68%**
GLiNER Multi-task v1.0	Partial (strict type)	0.429	0.581	0.383	0.082	0.437	0.057
	Strict	0.400	0.535	0.359	0.011	0.189	0.006
		-7%	**-8%**	**-6%**	-87%	-57%	-89%
InstructUIE	Partial (strict type)	0.310	0.529	0.264	0.046	0.195	0.030
	Strict	0.257	0.437	0.222	0.013	0.193	0.007
		-17%	-17%	-16%	-72%	-1%	-77%
ZeroShotBioNER	Partial (strict type)	0.301	0.366	0.352	-	-	-
	Strict	0.254	0.301	0.304	-	-	-
		-16%	-18%	-14%	-	-	-

Table A2: Comparison of model performance when transitioning from partial strict-type matching criteria, as used throughout the paper, to strict matching. Strict matching involves case sensitivity, schema restriction of output, and an exact match for entities and relations. The smallest decrease in model performance when switching from partial to strict matching is highlighted in bold for each performance metric.

titissuestructure, organ, organismsubdivision, organismsubstance, pathologicalformation, physiology, tissue

`Other biological` `(n=8)` biologicalactivity, cancer, cellcomponent, cellularcomponent, lipid, multicell, organelle, virus

`Non-English` `(n=6)*` diagnostico, enfermedad, procedimiento, proteina, quimico, sintoma

`Other` `(n=19)` age, characteristic, cohortorpatient, ethnicity, experimentalconstruct, experimenttag, gender, geographicarea, habitat, inorganic, interventioncontrol, object, participant, participantage, participantsamplesize, participantsex, process, size, spectraldata

* The entity names for the French QUAERO and the Swedish Medical NER dataset were in English and thus included in the other groups.

D Model prompting

Inference for GLiNER, GLiNER multi-task, and NuNER were performed using the `gliner` python library, and ZeroShotBioNER using the published implementation. No prompts had to be provided for these `TokenClassifier` models - only entity/relation types were provided. Whenever possible, the default prompt format specified in the model implementation was used. Such prompts are marked with "(default)" - otherwise the prompts were designed.

For zero-shot inference (no examples), only the `<text>`, `<entity_types>`, and `<relation_types>` fields are provided. If no RE annotation exists for a given corpus, this part of the prompt is omitted. For models where we used few-shot prompting (Gemini 1.5 Pro & SciLitLLM 1.5), we show the format of the example given enclosed in parentheses.

InstructUIE (default)

```
NER:

Please list all entity words in the text that fit the category.
Output format is "type1: word1; type2: word2"
Option: <entity_types>
Text: <text>
Answer:

RE:

Given a phrase that describes the relationship between two words,
extract the words and the lexical relationship between them. The
output format should be "relation1: word1, word2; relation2: word3,
word4".
Option: <relation_types>
Text: <text>
Answer:
```

Triplex (default)

```
Perform Named Entity Recognition (NER) and extract knowledge graph
triplets from the text. NER identifies named entities of given
entity types, and triple extraction identifies relationships between
entities using specified predicates.

**Entity Types:**
<entity_types>

**Predicates:**
<relation_types>

**Text:**
<query>
```

Phi3 Mini Graph (default - modified to accept specific types)

```
A chat between a curious user and an artificial intelligence
Assistant. The Assistant is an expert at identifying entities and
relationships in text. The Assistant responds in JSON output only.

The User provides text in the format:

-------Text begin-------
<User provided text>
-------Text end-------

The Assistant follows the following steps before replying to the
User:

1. **identify entities** The Assistant identifies all entities in
the text of the types: <entity_types>. These entities are listed in
the JSON output under the key "nodes", they follow the structure of
a list of dictionaries where each dict is:

"nodes":[{"id": <entity N>, "type": <type>}, ...]

where "type": <type> is the type of the entity.

2. **determine relationships** The Assistant uses the text between
-------Text begin------- and -------Text end------- to determine
the relationships between the entities identified in the "nodes"
list defined above. These relationships are called "edges" and they
follow the structure of:

"edges":[{"from": <entity 1>, "to": <entity 2>, "label":
<relationship>}, ...]

The <entity N> must correspond to the "id" of an entity in the
"nodes" list and relationship must be one of the following types:
<relation_types>.

The Assistant never repeats the same node twice. The Assistant never
repeats the same edge twice.
The Assistant responds to the User in JSON only, according to the
following JSON schema:
{
   "type":"object",
   "properties":{
      "nodes":{
         "type":"array",
         "items":{
            "type":"object",
            "properties":{
               "id":{
                  "type":"string"
               },
               "type":{
                  "type":"string"
               },
               "detailed_type":{
                  "type":"string"
               }
            },
            "required":["id", "type", "detailed_type"],
            "additionalProperties":false
         }
      },
      "edges":{
         "type":"array",
         "items":{
            "type":"object",
            "properties":{
               "from":{
                  "type":"string"
               },
               "to":{
                  "type":"string"
               },
               "label":{
                  "type":"string"
               }
            },
            "required":["from", "to", "label"],
            "additionalProperties":false
         }
      }
   },
   "required":["nodes", "edges"],
   "additionalProperties":false
}

Input:
-------Text begin-------
<text>
-------Text end-------
```

Note: The JSON in the Phi3 Mini Graph prompt is condensed to take up less characters, but formatted here for readability.

Gemini 1.5 Pro

```
Please extract a list of entities, and subsequently a list of
relations between these entities.
The allowed entity types are: <entity_types>.
The allowed relation types are: <relation_types>.
The output should look like:
Entities:
Entity1 (EntityType)
Entity2 (EntityType)

Relationships:
Entity1 (EntityType) --RELATIONSHIP_TYPE-- Entity2 (EntityType)

(Examples:
Example 1:
<example_text>

Entities:
<example_entities>

Relationships:
<example_relationships>)

Do not provide any explanation or deviate from the format. If any
entity does not conform to the entity types stated, they should not
be included. Please now perform the task for the following text:
<text>
```

SciLitLLM 1.5

```
As a biomedical researcher, you are able to extract structured
information from a given piece of text. Please extract a list
of entities, and subsequently a list of relations between these
entities.
The allowed entity types are: <entity_types>.
The allowed relation types are: <relation_types>.
The output should look like:
(entity1_name, entity1_type), (entity2_name, entity2_type),
(entity1_name, RELATION, entity2_name), (entity3_name, RELATION,
entity4_name), ...

(Examples:
Example 1:
<example_text>

Output:
<example_output>)

Do not provide any explanation or deviate from the format. If any
entity does not conform to the entity types stated, they should not
be included. Please now perform the task for the following text:
<text>
```

Additional information

Setup, implementation details, and code can be found at https://github.com/FSGade/BiomedicalZeroShot.