Biomedical Relation Extraction: What Lies Between the Entities?

Nadiia Shpanko
nadya.shpanko@gmail.com
Jožef Stefan Institute
Ljubljana, Slovenia

Erik Calcina
erik.calcina@ijs.si
Jožef Stefan International
Postgraduate School
Jožef Stefan Institute
Ljubljana, Slovenia

Erik Novak
erik.novak@ijs.si
Jožef Stefan International
Postgraduate School
Jožef Stefan Institute
Ljubljana, Slovenia

Abstract

Clinical trial eligibility criteria are typically written in free text and contain medical concepts, numerical constraints, temporal conditions, and logical relations. Converting these criteria into structured representations requires identifying not only the relevant concepts but also the relations between them. In this work, we study relation classification between annotated entity mentions in the Chia corpus of clinical trial eligibility criteria. We compare three compact GLiNER2 model variants under the same fine-tuning and evaluation protocol. The Large model achieves the best held-out test performance with a micro-F1 score of 0.7510, compared with 0.7260 for Base and 0.7039 for Multilingual. The results show that compact, locally deployable models can perform effective biomedical relation classification while avoiding dependence on external inference services.

Keywords

Relation extraction, Biomedical NLP, Clinical trial eligibility criteria, Information extraction, Local inference

1 Introduction

Clinical trial eligibility criteria determine whether a patient can participate in a study. They combine diagnoses, procedures, medications, measurements, demographic restrictions, temporal conditions, and Boolean logic, yet are usually written as free text. Computational use therefore requires transformation of natural-language criteria into structured representations that support cohort identification, patient-trial matching, and query generation. The Chia corpus [4] was created for this purpose and represents eligibility statements as annotated entities connected by semantic and logical relationships.

Relation Extraction (RE) determines how concepts mentioned in a text are connected. Named Entity Recognition (NER) can identify a measurement such as haemoglobin or a condition such as hypertension, but mentions alone do not capture the relationships expressed by the criterion. A numerical threshold must be linked to its measurement, a temporal expression to the event it constrains, and logical alternatives through `AND` or `OR`. RE therefore requires relation existence and type to be determined for the relevant argument pair while preserving direction where applicable.

Biomedical RE is challenging because realistic schemas mix directed and symmetric relations, heterogeneous labels, and many unrelated candidate pairs. Strong entity recognition does not necessarily translate into strong relation extraction [6], and zero- or few-shot transfer remains unreliable [5]. We therefore focus on relation-specific fine-tuning.

We study these questions on Chia [4], which contains eleven semantic and logical relation types. Its mixture of directed and symmetric relations, multi-argument logical groups, complex mentions, and numerous unrelated candidate pairs makes it a useful evaluation setting for pairwise RE.

A second practical consideration is deployment. Clinical text may contain protected health information, making local processing attractive for privacy-sensitive applications [3]. GLiNER2 provides a compact schema-conditioned framework that can run locally without relying on an external language-model API [8].

This study makes three contributions. First, we formulate RE as reference-anchored pair-conditioned classification, separating relation existence and typing from mention detection. Second, we compare Base, Large, and Multilingual compact schema-driven model variants under the same parameter-efficient fine-tuning and evaluation protocol, including a direct comparison of the Multilingual and English Base checkpoints on the English-only Chia benchmark. Third, we analyse results per relation type and show that frequency alone does not explain difficulty.

The remainder of the paper is structured as follows. Section 2 reviews related work. Section 3 describes the relation extraction methodology. Section 4 presents the dataset, the training configuration, and the evaluation protocol. Section 5 reports the results, Section 6 discusses the findings and limitations, and Section 7 concludes the paper and outlines future work.

2 Related Work

Modern RE benchmarks increasingly contain heterogeneous entity and relation categories rather than a single interaction type. BioRED shows that RE remains difficult even when entity extraction is comparatively strong [6]. A common supervised strategy is to represent candidate spans and classify their pairs. Ji et al. proposed a span-based joint entity and relation model with contextual and span-specific representations [2], while PL-Marker models subject-object span pairs with explicit markers [7]. We take the same pair-conditioned view: the central decision is whether a pair of mentions is related and, if so, under which label.

The models we compare come from the schema-conditioned line of work. GLiNER introduced label-conditioned span extraction with a bidirectional Transformer encoder, allowing arbitrary labels without autoregressive generation [9]. GLiNER2 extends this approach to several information extraction tasks, including entity recognition, text classification, and hierarchical structured extraction [8]. Both are compact enough for local deployment.

Two further findings shape our setup. Labrak et al. report that instruction-finetuned models handle some clinical natural

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2026, Ljubljana, Slovenia
© 2026 Copyright held by the owner/author(s).
https://doi.org/10.70314/is.2026.sikdd.112

language processing (NLP) tasks well while relation extraction remains weak under zero- and few-shot prompting [5], motivating fine-tuning in our setup. Bassignana and Plank note that reported RE scores depend strongly on protocol, including whether mentions are supplied or predicted, how direction is handled, and whether exact mention boundaries are required [1]. We therefore state the candidate construction explicitly and evaluate relation classification independently of mention detection.

3 Methodology

3.1 Relation extraction formulation

The task is formulated as single-label pair classification. As a concrete example from Chia [4], consider the eligibility criterion "patient has [Condition: hypertension] ([Measurement: systolic blood pressure] > 135 mmHg or [Measurement: diastolic blood pressure] > 85 mmHg)." The criterion contains the clinical concept "hypertension", two blood-pressure measurements, their numerical threshold values, and a logical `OR` connecting the two measurement conditions. This example illustrates the type of structured information represented in Chia and the relations that must be recovered between annotated concepts. In our experiments, the entity mentions and their types are supplied by the reference annotations, and the model is evaluated on the relations between these annotated mentions rather than on entity detection itself. Within each input unit, a single eligibility criterion in Chia, we enumerate all ordered pairs of distinct mentions taken from the reference annotations. No type- or distance-based filtering is applied. Both arguments are marked in the input together with their reference entity types, for example `[ARG1|Measurement]` and `[ARG2|Value]`, and the model assigns each pair a relation type or the `NA` label when no relation applies. We call this construction reference-anchored. Fixing the mentions separates relation existence and typing from mention detection, so the reported score reflects the relation decision and not the entity recogniser.

Direction is retained for semantic relations. For the symmetric `AND` and `OR` relations, the reverse orientation of an annotated edge is excluded from the candidate set rather than labelled `NA`, so these relations are scored only in the annotated orientation.

A pair that carries more than one relation label is expanded into one instance per reference label, and the instances are scored independently. Because the model returns one label per pair, every multi-label pair contributes at least one false negative under this evaluation.

3.2 Distribution-aware training construction

Enumerating all entity pairs produces a pronounced class imbalance, with negative examples substantially outnumbering positive ones. The pairwise expansion of Chia's logical annotations also inflates `OR` relative to the other relation types. Training on this distribution can bias the model toward predicting `NA` and hinder learning of the underrepresented relations.

We therefore construct a tempered training subset that reduces the imbalance without making all relation types equally frequent. Let n_i denote the number of positive training instances of relation class i. Each class is assigned a sampling weight proportional to $n_i^{0.85}$. Because the exponent is smaller than one, differences between frequent and rare classes are compressed: frequent relations remain more common, but their dominance is reduced. The weights are normalised to a fixed budget of positive instances, sampled randomly within each class, and an equal number of `NA` candidates is then drawn, giving a 1:1 ratio of positive to negative training examples. This preserves more `OR` and `AND` examples than uniform class balancing while raising the relative share of the rare relations. The exponent 0.85 was chosen heuristically. It was not tuned on the development set, and no sensitivity analysis or sampler ablation was performed. The budget used here is given in Section 4.2.

3.3 Model variants and fine-tuning

Three pretrained GLiNER2 [8] span models are compared: `gliner2-base-v1`[1], `gliner2-large-v1`[2], and `gliner2-multi-v1`[3]. All models are adapted with Low-Rank Adaptation (LoRA), so only a small fraction of parameters is updated.

The Multilingual variant is included not as an evaluation of cross-lingual RE, since Chia is English-only, but as a comparison with the English Base checkpoint under the same English benchmark. This comparison is motivated by potential deployment in multilingual clinical settings, but the experiment does not isolate multilinguality itself as the cause of any observed performance difference.

For comparability, all three models are fine-tuned under the same training configuration, given in Section 4.3.

4 Experiment Setting

4.1 Dataset

Experiments use Chia, a manually annotated corpus of clinical trial eligibility criteria introduced by Kury et al. [4]. The original resource contains 12,409 eligibility criteria from 1,000 interventional Phase IV studies, with 41,487 distinctive entities from 15 entity types and 25,017 relationships from 12 relationship types. Chia aligns its clinical domains with the Observational Medical Outcomes Partnership (OMOP) Common Data Model and represents each criterion as an annotation graph that can be transformed into Boolean logic.

Eleven relation types are targeted. `Has_value` links a clinical concept to its value, `Has_qualifier` attaches a descriptive modifier, `Has_temporal` a temporal restriction, and `Has_index` connects a temporal expression to its reference point. `Has_negation`, `Has_multiplier`, and `Has_mood` connect the corresponding constructs, and `Has_context` links an entity to an observation used as context. `Subsumes` covers dependency, clarification, example, and specification between concepts, while `AND` and `OR` encode Boolean combinations. Chia also defines `Has_scope`, which the post-processed *Without Scope* variant used here excludes.

Chia is structurally difficult for pairwise RE. `OR` annotations can hold more than two arguments: 29.9% of the original `OR` relations connect more than two entities, and the largest reported group contains 25. Converting these into pairwise decisions multiplies the number of `OR` edges (Table 1). The corpus also permits nested, overlapping, and discontinuous annotations, so the relation arguments are more complex than in flat mention-based corpora.

At the criterion-conversion stage, criteria with no retained in-schema entity are excluded because they cannot yield reference-anchored relation candidate pairs. This leaves 10,767 of the 12,409 original criteria, partitioned into 7,714 training, 1,052 development, and 2,001 test criteria. Criteria originating from the same

[1] https://huggingface.co/fastino/gliner2-base-v1
[2] https://huggingface.co/fastino/gliner2-large-v1
[3] https://huggingface.co/fastino/gliner2.5-multi-v1

Table 1: Positive pairwise relation instances by partition. Train, Dev, and Test denote the training, development, and test partitions, respectively. The two logical relations are listed separately because pairwise expansion of their multi-argument groups inflates their counts.

Relation	Train	Dev	Test	Total
Has_value	2,671	384	750	3,805
Has_qualifier	2,549	385	598	3,532
Has_temporal	2,409	321	599	3,329
Subsumes	1,452	202	396	2,050
Has_index	649	79	171	899
Has_negation	616	78	149	843
Has_multiplier	439	61	132	632
Has_mood	347	74	102	523
Has_context	168	16	36	220
AND	2,646	340	688	3,674
OR	10,538	1,685	2,957	15,180
Total	**24,484**	**3,625**	**6,578**	**34,687**

ClinicalTrials.gov study record, identified by a unique ClinicalTrials.gov identifier (National Clinical Trial (NCT) number), are kept in the same partition.

4.2 Data preparation

The original brat annotations are converted into criterion-level JSONL records. Relation direction is preserved, and multi-argument logical annotations are expanded into pairwise relation instances. For example, a three-argument logical group OR(A,B,C) is represented by the pairwise relations OR(A,B), OR(A,C), and OR(B,C). This representation preserves the pairwise connections but no longer represents the original three-argument group as a single structure. Discontinuous mentions are represented by their minimum enclosing span because the evaluated span architecture operates on continuous intervals. In other words, when an annotated mention consists of several separated text segments, we use the shortest continuous text span that covers all of those segments.

After pairwise conversion, positive relation instances are distributed as shown in Table 1. The counts cover the full partitions before tempered sampling and exclude NA candidate pairs. OR alone accounts for 15,180 of the 34,687 positive instances, and the two logical relations together for 54%. This is the skew the sampler of Section 3.2 is built to temper.

Candidate pairs are then constructed from the reference annotations. The development partition yields 21,855 candidate pairs, 3,625 positive and 18,230 NA; the held-out test partition yields 42,091, of which 6,578 are positive and 35,513 NA. A few pairs carry more than one relation label, 14 in training, none in development and 5 in the test partition, and these expand to 28 training and 10 test instances. The tempered sampler of Section 3.2 is run with a budget of 6,000 positive instances and 6,000 NA candidates, giving 12,000 training pairs.

4.3 Models and training configuration

The Base, Large, and Multilingual checkpoints contain about 208.5M, 486.4M, and 307.1M parameters, and their LoRA adapters add about 1.64M, 3.97M, and 1.65M trainable parameters. LoRA targets the encoder query, key, and value projections together with the task-specific projection, classification, and count-prediction layers; the model-specific target-module lists are given with the experiment configuration in the project repository. The shared training configuration is summarised in Table 2. Experiments ran on a single NVIDIA GeForce RTX 4070 Ti SUPER with 16 GB of memory, using PyTorch 2.13.0 with CUDA 12.6, GLiNER2 1.3.2, Transformers 5.14.1, and PEFT 0.20.0.

Table 2: Hyperparameters used for parameter-efficient fine-tuning.

Hyperparameter	Value	Hyperparameter	Value
Training examples	12,000	Warmup steps	150
Fine-tuning method	LoRA	LR scheduler	Linear
Epochs	2	Max gradient norm	1.0
Batch size	4	LoRA rank (r)	8
Gradient accumulation	4	LoRA alpha	16
Effective batch size	16	LoRA dropout	0.0
Learning rate	5×10^{-4}	Training seed	20260810
Weight decay	0.01	Pair-construction seed	20260830

4.4 Evaluation metrics

Relation classification is evaluated with micro-averaged Precision, Recall, and F1, computed over the eleven relation classes. NA marks the absence of a relation rather than a relation type, and it is excluded because negative pairs outnumber positive ones by roughly five to one and would otherwise dominate the metric.

Because candidate spans and types are fixed by the reference annotations, a true positive requires the correct relation label for the correct ordered candidate pair. The confidence threshold is applied to the model-reported confidence of the predicted class: a non-NA prediction below the threshold is mapped to NA, while a predicted NA is retained. Thresholds are selected on the development set by sweeping predefined candidate values, with a finer search around the best-performing values, then fixed and applied unchanged to the held-out test partition. We also report macro-F1, the unweighted mean of the per-relation F1 scores. Unlike micro-F1 it gives equal weight to frequent and rare relations, which is the complementary view needed under this class distribution.

5 Results

Table 3 reports both partitions. Large is the strongest model on the held-out test partition at 0.7510 micro-F1, followed by Base at 0.7260 and Multilingual at 0.7039.

Between the two partitions Base changes little, from 0.7230 to 0.7260 micro-F1, while Large falls from 0.7628 to 0.7510 and Multilingual from 0.7113 to 0.7039. The ranking is preserved on unseen test data.

Table 4 breaks the Large model's test performance down by relation.

The strongest scores are for Has_index, Has_value, and Has_qualifier; Has_context and AND are the weakest. Support alone does not account for this. Has_index reaches 0.934 on 171 positives, while OR, the largest class at 2,957, reaches 0.734 and AND, at 688, reaches 0.538. The low Has_context score should be read cautiously, since only 36 positive examples occur in the test partition.

Inference over the 42,091 test candidate pairs took about 15.8 min for Base, 30.1 min for Large, and 15.7 min for Multilingual,

Table 3: Relation classification results on the development and held-out test partitions, with NA excluded from P/R/F1. Thresholds were selected on the development partition and applied unchanged to the test partition: 0.990 (Base), 0.995 (Large), 0.980 (Multilingual).

Model	Split	P	R	Micro-F1	Macro-F1
Base	dev	0.7043	0.7426	0.7230	0.6778
	test	0.7167	0.7355	0.7260	0.6952
Large	dev	**0.7393**	**0.7879**	**0.7628**	**0.7322**
	test	**0.7351**	**0.7676**	**0.7510**	**0.7494**
Multilingual	dev	0.6837	0.7412	0.7113	0.6461
	test	0.6876	0.7209	0.7039	0.6669

Table 4: Per-relation F1 scores and support of the Large model on the held-out test partition.

Relation	F1	*n*	Relation	F1	*n*
`Has_value`	0.873	750	`Has_negation`	0.827	149
`Has_qualifier`	0.859	598	`Has_multiplier`	0.818	132
`Has_temporal`	0.759	599	`Has_mood`	0.840	102
`Subsumes`	0.644	396	`Has_context`	0.417	36
`Has_index`	0.934	171	`AND`	0.538	688
			`OR`	0.734	2,957

or about 44.4, 23.3, and 44.6 candidate pairs per second on the hardware in Section 4.3.

6 Discussion

The three variants reach 0.7039–0.7510 micro-F1 on the held-out test partition, with the same ordering on both partitions. On the English-only Chia benchmark, Multilingual reaches 0.7039 compared with 0.7260 for Base. This observed difference does not establish multilinguality as its cause, and Chia cannot evaluate cross-lingual RE capability.

Three limitations constrain interpretation. Fixed mention spans and types isolate relation classification from upstream detection errors, so the scores are not end-to-end results from raw text. No external RE system is evaluated, so the results support comparison only among the tested variants under this protocol. Finally, the high selected thresholds (0.980–0.995) suggest that confidence values are concentrated near the top of their range, leaving calibration for future work.

Residual difficulty comes from several sources. Pairwise expansion alters the original structure of multi-argument `OR` groups, while the tempered sampler reduces imbalance without reconstructing them. `Subsumes` covers several semantic dependencies under one label, whereas `Has_context` is rare and therefore unstable.

The models run on a single consumer GPU and avoid dependence on third-party inference services, although local deployment still requires appropriate security and validation.

7 Conclusion

We evaluate biomedical relation extraction as reference-anchored relation classification over eleven semantic and logical relation types and compare three compact schema-driven variants under one parameter-efficient fine-tuning protocol.

Large achieves the highest held-out micro-F1 of 0.7510. Performance varies substantially across relation types, reflecting logical structure, semantic ambiguity, and data scarcity rather than frequency alone.

Future work will consider stronger candidate generation, explicit modelling of logical groups, external relation-classification baselines, alternative sampling strategies, relation-specific calibration, additional biomedical RE corpora, and locally deployed clinical information extraction pipelines.

Acknowledgements

This work was supported by the Slovenian Research Agency. Funded by the European Union. UK participants in Horizon Europe Project PREPARE are supported by UKRI grant number 10086219 (Trilateral Research). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Health and Digital Executive Agency (HADEA) or UKRI. Neither the European Union nor the granting authority nor UKRI can be held responsible for them. Grant Agreement 101080288 PREPARE HORIZON-HLTH-2022-TOOL-12-01.

References

[1] Elisa Bassignana and Barbara Plank. 2022. What do you mean by relation extraction? a survey on datasets and study on scientific relation classification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, 67–83.

[2] Bin Ji, Jie Yu, Shasha Li, Jun Ma, Qingbo Wu, Yusong Tan, and Huijun Liu. 2020. Span-based joint entity and relation extraction with attention-based span-specific and contextual semantic representations. In *Proceedings of the 28th International Conference on Computational Linguistics*, 88–99. doi:10.18653/v1/2020.coling-main.8.

[3] Jitendra Jonnagaddala and Zoie Shui-Yee Wong. 2025. Privacy preserving strategies for electronic health records in the era of large language models. *npj Digital Medicine*, 8, 1, 34. doi:10.1038/s41746-025-01429-0.

[4] Fabrício Kury, Alex Butler, Chi Yuan, Li-Heng Fu, Yingcheng Sun, Hao Liu, Ida Sim, Simona Carini, and Chunhua Weng. 2020. Chia, a large annotated corpus of clinical trial eligibility criteria. *Scientific Data*, 7, 1, 281. doi:10.1038/s41597-020-00620-0.

[5] Yanis Labrak, Mickael Rouvier, and Richard Dufour. 2024. A zero-shot and few-shot study of instruction-finetuned large language models applied to clinical and biomedical tasks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2049–2066.

[6] Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N. Arighi, and Zhiyong Lu. 2022. BioRED: a rich biomedical relation extraction dataset. *Briefings in Bioinformatics*, 23, 5, bbac282. doi:10.1093/bib/bbac282.

[7] Deming Ye, Yankai Lin, Peng Li, and Maosong Sun. 2022. Packed levitated marker for entity and relation extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4904–4917. doi:10.18653/v1/2022.acl-long.337.

[8] Urchade Zaratiana, Gil Pasternak, Oliver Boyd, George Hurn-Maloney, and Ash Lewis. 2025. GLiNER2: an efficient multi-task information extraction system with schema-driven interface. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Suzhou, China, 130–140. https://aclanthology.org/2025.emnlp-demos.10/.

[9] Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. GLiNER: generalist model for named entity recognition using bidirectional transformer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 5364–5376. doi:10.18653/v1/2024.naacl-long.300.