University of Bonn

CAISA lab

Natural Language Processing Lab

(Summer Semester 2025)

Final Report of Team #3

SciREX: Scientific Relation Extraction

Group Members:

Sayantak Karar
Zyad Altahan
Sulaeman Aloradi
Abdelwahab Elshennawy

Instructor: Prof. Dr. Lucie Flek
Advisor: Frederik Labonté

September 23, 2025

1 Introduction

The rapid growth of biomedical literature makes it increasingly difficult to identify and organize meaningful knowledge. This project addresses the problem by focusing on **relation extraction (RE)**, i.e., detecting and classifying semantic relationships between biomedical entities within scientific abstracts.

Our objective is to evaluate and compare multiple paradigms for scientific relation extraction on the BioRED dataset, a manually annotated benchmark of PubMed abstracts with diverse entities and relation types. Specifically, we investigate three complementary approaches: a **classification-based model** using BioBERT, a **question answering formulation** QA4RE, and lastly, **generative models** SciFive and REBEL.

The central research question guiding our study is: *Which modeling paradigm offers the most effective and generalizable solution for biomedical relation extraction under the constraints of the BioRED dataset?*

2 Related Work

Biomedical relation extraction has been studied extensively, with BioRED [1] emerging as a benchmark dataset containing 600 PubMed abstracts annotated with diverse entities and relations. While entity recognition in BioRED reaches F1 $\approx$ 89.3%, relation extraction—especially for novel relations—remains challenging (F1 $\approx$ 47.7%).

Several approaches have been proposed to address these challenges. Shang et al. [2] introduce adaptive document-relation cross-mapping with concept identifiers, achieving up to 72% F1 on BioRED, though limited to predefined relation types. Yamada et al. [3] reframe RE as a question answering task with entity markers, yielding strong results on DrugProt but requiring multiple queries per entity pair. Li and Verspoor [4] propose EMBRE, an entity-masking pretraining strategy that improves novelty detection at the cost of higher computation and sensitivity to NER errors. Ensemble-based methods like SARE [5] show gains by combining multiple pretrained models, but suffer from high complexity and inference cost.

Overall, three themes emerge: (i) explicit modeling of entity types improves performance, (ii) QA and generative formulations add flexibility but increase computational overhead, and (iii) transformer baselines such as BioBERT [6] remain strong for frequent classes but underperform on rare or novel relation types. Dataset imbalance and strict evaluation constraints continue to be open challenges.

3 Methodology

We designed a modular pipeline to compare three paradigms for relation extraction (RE) on the BioRED dataset. Each method follows the same preprocessing, training, and evaluation protocol to ensure comparability.

3.1 BioBERT + Classification Head

BioBERT [6], a domain-specific BERT pretrained on PubMed and PMC, was fine-tuned with a classification head. Each instance consisted of a sentence with two entity mentions, and the model predicted the relation label. Entity markers and the [CLS] token were used to encode entity-level context.

3.2 QA4RE

QA4RE [7] recasts RE as a question-answering task, where entity pairs are turned into natural language queries with multiple-choice answers. This leverages pretrained LLMs for structured relational prediction without explicit relation classification.

3.3 Generative Models

We reformulated RE as a text-to-text generation task using two seq2seq models:

- SciFive [8], a biomedical T5 model, which outputs structured triples (e.g., *Aspirin — Chemical — treats — Inflammation — Disease*).
- REBEL [9], designed for open-domain RE, adapted to biomedical text using delimiter-based serialization for subject–relation–object triples.

3.4 Overall Pipeline

The end-to-end process was:

1. **Preprocessing:** Parse BioRED abstracts into model-specific formats (classification instances, QA prompts, or serialized triples).
2. **Task Formulation:** Train models under three paradigms—classification, generation, and QA.
3. **Training:** Fine-tune each model on the BioRED training set with standard train/dev/test splits.
4. **Evaluation:** Assess performance using macro-F1 and accuracy; qualitative error analysis used confusion matrices and output inspection.

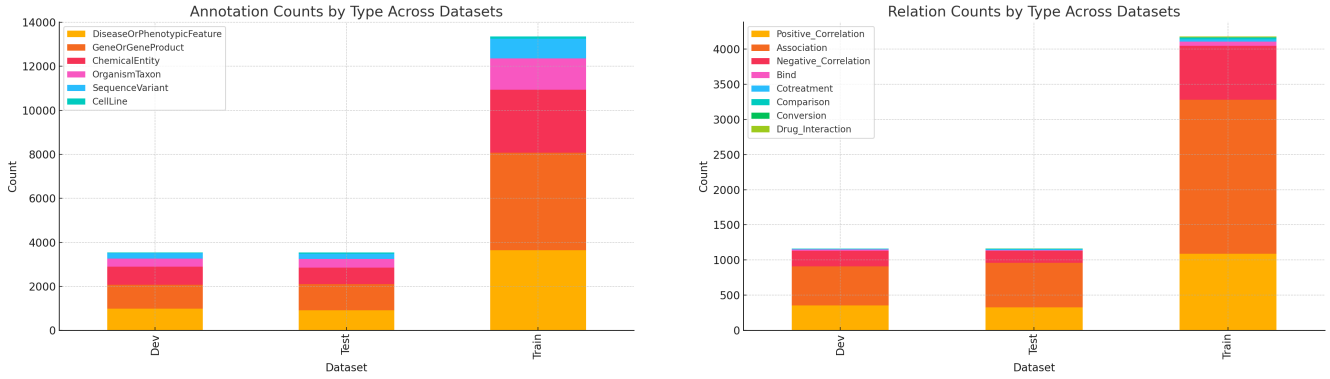
This pipeline enables principled comparison of distinct NLP strategies for biomedical relation extraction.

4 Dataset

We used the BioRED dataset [1], which contains 600 PubMed abstracts annotated with biomedical entities and semantic relations. Entity categories include Gene/Gene Product, Disease/Phenotype, Chemical, Sequence Variant, and Cell Line. Gene and Disease dominate (≈60% of mentions), while Sequence Variant and Cell Line are rare (≈10% combined).

Relation types are similarly imbalanced: *Association* is the most frequent, followed by *Positive-Correlation* and *Negative-Correlation*. Less common but important types include *Bind*, *Cotreatment*, *Drug Interaction*, and *Conversion*. Figure 1 shows the distribution of entities and relations across splits.

This imbalance presents challenges for learning robust models, motivating the use of techniques such as weighted loss or careful evaluation with macro-F1. Nonetheless, the balanced design of train/dev/test splits ensures fair benchmarking.



(a) Annotation Counts

(b) Relation Counts

Figure 1: Entity and relation distributions across datasets.

5 Experimental Setup

We present the experiments on each of the methods (available on Github), in the following sections.

5.1 BioBERT

For the classification-based approach, we fine-tuned `dmis-lab/biobert-base-cased-v1.1` [6] using PyTorch and HuggingFace Transformers. Each instance consisted of a sentence with two marked entities and the corresponding relation label.

We conducted a staged hyperparameter search. First, runs with maximum sequence length 256 explored variations in dropout, weight decay, and learning rate. Top-performing settings were then rerun with length 512. Then,

label smoothing values (0.05–0.2) were tested on the top runs of the previous stage. Initial runs were capped at 10 epochs. Top configurations (Test F1 > 0.5) were extended to 100 epochs with early stopping (patience 5, delta $1e-4$), selecting checkpoints by Dev F1.

The best configuration used a sequence length of 512, weight decay 0.03, and learning rate $2e-5$, achieving the highest F1 on the test set. A complete record of all experiments and hyperparameters is provided in the Appendix for reproducibility.

5.2 QA4RE

In the QA4RE approach [7], relation extraction is reformulated as a multiple-choice question answering task. We evaluated several large language models (LLMs) on BioRED using this format.

The BioRED JSON files (Train, Dev, Test) were converted into QA4RE-style prompts. For each document, entities and relations were extracted, and valid entity-pair relations were transformed into multiple-choice questions with one correct option. The resulting datasets contain 4497 prompts for Train, 1284 for Dev, and 1123 for Test, and a few-shot variant was created by adding examples to each prompt. Each prompt corresponds to one entity pair and relation label (options A–H), with only type-consistent relation options provided.

Models were loaded using HuggingFace (for open-source LLMs) and LiteLLM (for GPT). Both encoder–decoder models (e.g., T5) and causal decoder-only LLMs (e.g., LLaMA, Mistral, GPT) were tested.

The pipeline followed three steps: Load model and tokenizer. Then, for each prompt, generate an answer (e.g., "A", "B") and map to a valid option. Finally, we compare predictions to the gold label and compute accuracy and macro-F1.

5.3 Generative RE Models

In this approach, relation extraction is framed as a text-to-text generation task. Input sentences with entity pairs are serialized into a structured schema of the form:

```
<subj> | <subj type> | <relation> | <obj> | <obj type>
```

This representation was used consistently during training and evaluation.

5.3.1 SciFive

We fine-tuned the `scifive-base-Pubmed` checkpoint [8]. Each input consisted of a sentence-level context with paired entities, serialized into the schema above. The output was the corresponding relation triple. Training ran for up to 44 epochs with AdamW optimization and linear learning rate scheduling. During inference, beam search decoding was applied. The best checkpoint was selected based on validation loss, and evaluation followed the strict BioRED scorer (exact entity spans and relation labels).

5.3.2 REBEL

For REBEL, we used the pretrained `Babelscape/rebel-large` model [9], fine-tuned for 11 epochs on BioRED using HuggingFace's seq2seq trainer with AdamW optimization. The preprocessing and serialization matched SciFive. Inference used beam search decoding, and evaluation again relied on the strict BioRED scorer. No additional post-processing (e.g., entity normalization) was applied, to comply with the dataset's exact-match requirements.

6 Results

The results of our experiments are discussed below.

6.1 Evaluation Metrics

We evaluate model performance using two standard classification metrics:

- **Macro F1 Score**: This is the average F1 score computed independently for each class and then averaged. It gives equal importance to all relation classes, regardless of their frequency, making it especially suitable for class-imbalanced datasets like BioRED.
- **Accuracy**: The overall proportion of correctly predicted relation types among all instances.

6.2 Quantitative Results

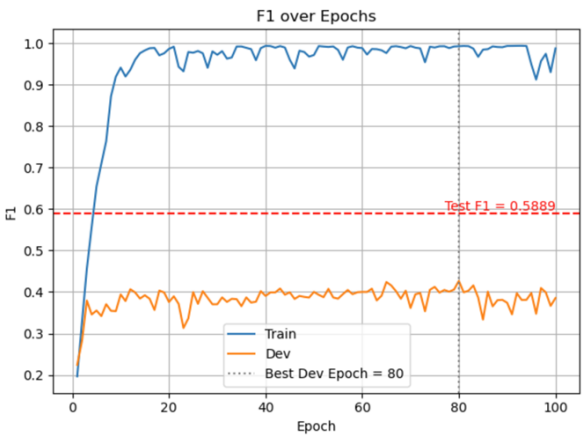
Below, we present the quantitative results for each of our approaches and a corresponding analysis on each of them.

6.2.1 BioBERT

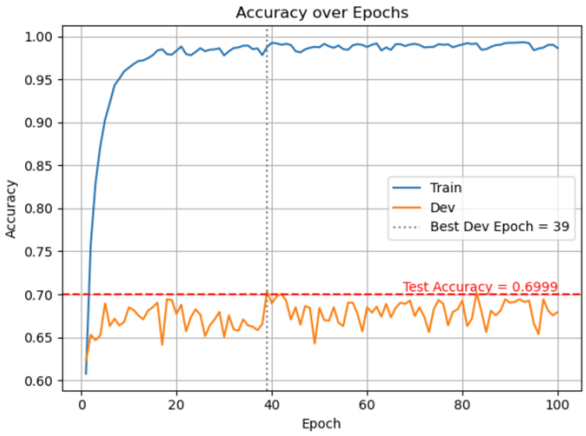
Table 1 presents the top-performing BioBERT configurations on the test set. The best run, with sequence length 512, weight decay 0.03, and learning rate $2e-5$, achieved a test F1 of 0.5889 and Accuracy of 0.6999. Figure 2a and 2b illustrate the F1 and accuracy trends over training. Although some overfitting was observed, selecting checkpoints by validation F1 consistently yielded the best test results, confirming the dev set as a reliable proxy.

Seq Len	Weight Decay	Label Smoothing	Learning Rate	Test F1
512	0.03	No	2e-5	0.5889
512	0.025	No	2e-5	0.5876
512	No	No	2e-5	0.5674
512	0.03	0.05	2e-5	0.5560
512	0.02	No	2e-5	0.5519
512	0.015	No	2e-5	0.5313

Table 1: Top 6 configurations for BioBERT based on Test F1 score (best run highlighted). Full experimental results are included in the appendix.



(a) F1 Score over Epochs



(b) Accuracy over Epochs

Figure 2: Performance metrics of best BioBERT run

Error Analysis The confusion matrix (Figure 3) shows that `Association` dominates both in frequency and misclassification, with frequent confusion against `Positive_Correlation`. Overall, BioBERT benefited from longer sequence lengths and careful regularization, with weight decay emerging as the most impactful hyperparameter.

6.2.2 QA4RE

LLMs' Results The models below in Table 2 are tested on the Test dataset to compare them with the other approaches e.g. BioBERT.

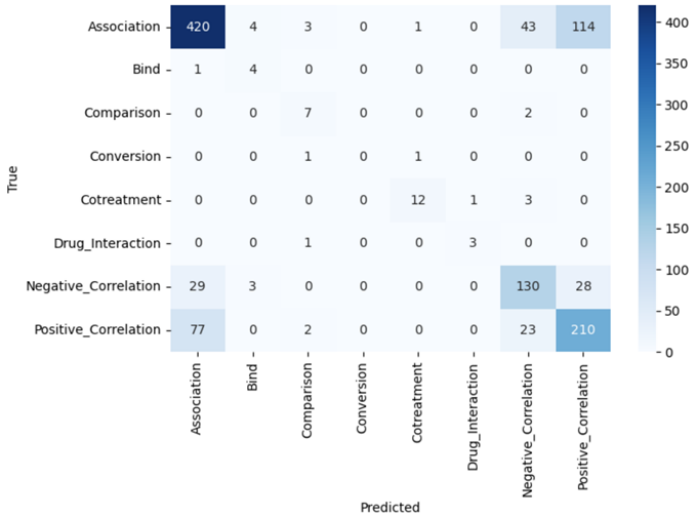


Figure 3: Confusion matrix for the test set (best BioBERT run).

Model	Architecture	Size	Accuracy %	F1-Score %
`flan-t5-small`	Seq2Seq (Enc–Dec)	60M	19	4
`flan-t5-base`	Seq2Seq (Enc–Dec)	220M	20	7
`flan-t5-large`	Seq2Seq (Enc–Dec)	780M	22	5
`flan-t5-xl`	Seq2Seq (Enc–Dec)	3B	24	4
`Mistral-7B-Instruct-v0.3`	Causal LM (Dec-only)	7B	21	11
`Llama-3.2-3B-Instruct`	Causal LM (Dec-only)	3B	35	22
`Llama-3.1-8B-Instruct`	Causal LM (Dec-only)	8B	22	15
`DeepSeek-R1-Distill-Llama-8B`	Causal LM (Dec-only)	8B	23	15
`Gemma-3-4B-IT`	Causal LM (Dec-only)	4B	22	2
`GPT-4o`	Causal LM (Dec-only)	$\sim$ 200B	45*	20*

Table 2: Model Experimental Results on the Test and Dev Datasets. *GPT-4o was tested on Dev dataset (similar to Test). Highlights show the best results.

Zero-shots vs. Few-shots The top-performing models are retested on the train dataset, as it is larger than the Test dataset. In Figure 4, we can see the zero-shot comparison between the models, in which we can see that GPT-4o and Llama-3.2-3B have the best F1-score. On the other hand, when testing using the few-shot models Figure 5, the results were less than expected. This is probably due to the lack of domain knowledge when designing the examples. E.g. Llama-3-3B with few shots was F1 $\sim$13% underperforms zero-shot prompting with F1 $\sim$20%.

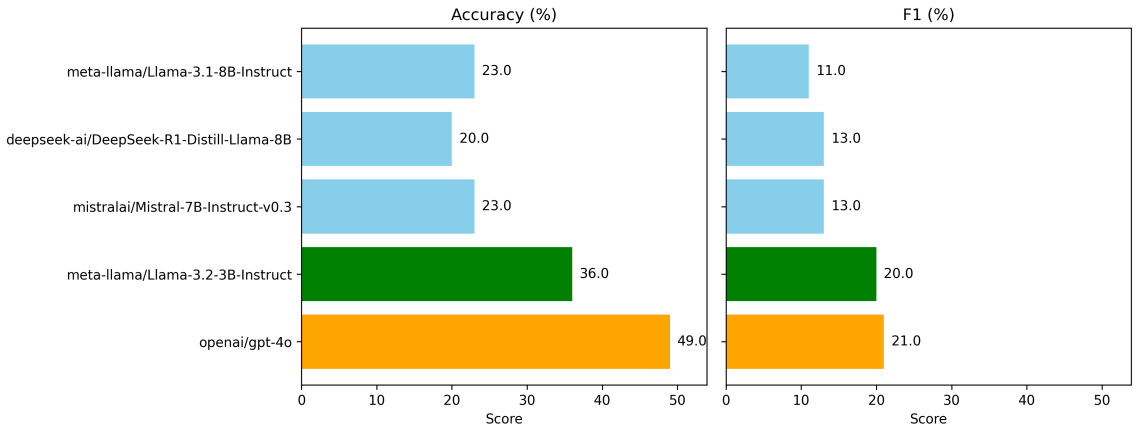


Figure 4: Comparison of top-performing models on the Train dataset (Accuracy and F1-score).

6.2.3 Generative RE Models

On the BioRED test set, both SciFive and REBEL exhibited very low strict scores due to frequent output format errors and under-generation of triples. Table 3 summarizes their performances.

Model / Relation	Support	Precision	Recall	F1
SciFive (strict scorer)				
Association	21	0.0526	0.0476	0.0500
Negative Correlation	12	0.0000	0.0000	0.0000
Positive Correlation	10	0.0000	0.0000	0.0000
Micro Avg.	43	0.0400	0.0233	0.0294
REBEL (strict scorer)				
Micro Avg.	–	0.1526	0.0933	0.1158

Table 3: SciFive and REBEL results on BioRED test set (strict evaluation).

Error Analysis Both models suffered from:

- **Output drift**: extra commentary or malformed delimiters.
- **Entity mismatches**: hallucinated IDs or merged biomedical terms.
- **Low recall**: generating far fewer triples than gold annotations per abstract.

6.3 Comparative Analysis

The relation extraction task was approached using three distinct paradigms: a classification-based model (BioBERT), a question answering formulation (QA4RE), and a generative models (SciFive and REBEL). Each method presents unique strengths and trade-offs in terms of performance, scalability, and generalization.

BioBERT, being a domain-specific encoder-based model, benefited significantly from hyperparameter tuning, particularly with weight decay and increased sequence length. It consistently achieved strong test F1 scores and showed stable generalization trends when early stopping was based on validation performance.

QA4RE's experiments show that seq2seq models (e.g., T5) frequently produce invalid predictions and struggle with biomedical relation extraction, whereas instruction-tuned causal LMs (e.g., Llama, Mistral, DeepSeek) achieve stronger results but still lag behind specialized biomedical models such as BioBERT. Performance is highly sensitive to prompt wording, and few-shot prompting (F1 ∼13%) underperforms zero-shot prompting (F1 ∼20%), suggesting that biomedical RE requires domain-specific calibration rather than general few-shot examples.

Generative RE (SciFive, REBEL). Under BioRED's strict exact-match scoring, generative seq2seq models struggled due to unconstrained decoding and span/ID mismatches. On the test set, *SciFive* (micro) P=0.0400, R=0.0233, F1=0.0294; *REBEL* (micro) P=0.1526, R=0.0933, F1=0.1158. Despite their flexibility and unified text-to-triple interface, performance remains far below BioBERT without grammar-/ID-constrained decoding or copy mechanisms.

Each model brings a complementary perspective to the task. A special highlight goes to the most successful method for the given dataset: the classification-based approach (BioBERT) which offered interpretability and strong performance with focused tuning.

7 Conclusion

This project compared classification, QA-based, and generative approaches for biomedical relation extraction on BioRED. **BioBERT** was the strongest baseline, with its best configuration reaching F1 = 0.5889 and Accuracy = 0.6999. Systematic tuning of weight decay, label smoothing, and sequence length improved results; despite some overfitting, validation F1 consistently guided the best checkpoints.

Challenges like class imbalance (e.g., Association dominating), strict span-matching in evaluation, noisy validation signals, and paradigm differences made fair comparison sensitive to formatting and decoding.

QA4RE offered flexibility, with larger LLMs (e.g., GPT-4o dev, LLaMA-3.2-3B test) reaching F1 ≈ 0.22. Performance, however, depended heavily on prompt quality and few-shot design.

Generative models provided a unified text-to-triple interface but performed poorly under strict scoring due to unconstrained decoding and entity mismatches. On the test set, SciFive achieved P=0.0400, R=0.0233, F1=0.0294; REBEL improved slightly with P=0.1526, R=0.0933, F1=0.1158. Stronger decoding constraints and copy-aware mechanisms are needed to close this gap.

Future work could explore hybrid systems that combine BioBERT's robustness with generative flexibility, or extend QA-based formulations with biomedical-specific prompting and light supervision.

References

[1] Ling Luo, Po-Ting Lai, Chih-Hsuan Wei, Cecilia N Arighi, and Zhiyong Lu. BioRED: a rich biomedical relation extraction dataset. *Briefings in Bioinformatics*, 2022.

[2] Yufei Shang, Yanrong Guo, Shijie Hao, and Richang Hong. Biomedical relation extraction via adaptive document-relation cross-mapping and concept unique identifier, 2025.

[3] Koshi Yamada, Makoto Miwa, and Yutaka Sasaki. Biomedical relation extraction with entity type markers and relation-specific question answering. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 377–384, 2023.

[4] Mingjie Li and Karin Verspoor. Embre: entity-aware masking for biomedical relation extraction. *arXiv preprint arXiv:2401.07877*, 2024.

[5] Yaxun Jia, Haoyang Wang, Zhu Yuan, Lian Zhu, and Zuo-lin Xiang. Biomedical relation extraction method based on ensemble learning and attention mechanism. *BMC bioinformatics*, 25(1):333, 2024.

[6] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, September 2019.

[7] Yu Su Kai Zhang, Bernal Jiménez Gutiérrez. Aligning instruction tasks unlocks large language models as zero-shot relation extractors. In *Findings of ACL*, 2023.

[8] Long N. Phan, James T. Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, and Grégoire Altan-Bonnet. Scifive: a text-to-text transformer model for biomedical literature, 2021.

[9] Pere-Lluís Huguet Cabot and Roberto Navigli. Rebel: Relation extraction by end-to-end language generation. In *Findings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: ACLicoS*, pages 2370–2381. Association for Computational Linguistics, 2021.

Appendix

BioBERT Experiments

This appendix includes the detailed results of all experimental runs for the BioBERT classification model. Each table presents the key hyperparameters used in a run (e.g., learning rate, dropout probability, weight decay, label smoothing) along with the resulting test F1 scores. These results provide transparency, reproducibility, and allow for a comprehensive understanding of how different configurations impacted model performance.

Separate tables are provided for:

- Runs with different dropout and weight decay configurations at sequence length 256

Baseline	Weighted Loss	Dropout	Weight Decay	Learning Rate	Test F1
Yes	No	No	No	2.00E-05	0.3165
Yes	Yes	No	No	2.00E-05	0.3488
Yes	No	Yes	No	2.00E-05	0.3718
Yes	Yes	Yes	No	2.00E-05	0.3209
Yes	No	No	0.005	2.00E-05	0.3588
Yes	No	No	0.01	2.00E-05	0.4832
Yes	No	No	0.015	2.00E-05	0.4164
Yes	No	No	0.02	2.00E-05	0.4171
Yes	No	No	0.025	2.00E-05	0.4458
Yes	No	No	0.03	2.00E-05	0.3981
Yes	No	No	0.035	2.00E-05	0.3428
Yes	No	No	0.05	2.00E-05	0.3457
Yes	No	No	0.1	2.00E-05	0.3654
Yes	No	No	0.5	2.00E-05	0.3272
Yes	No	Yes	0.01	2.00E-05	0.4182
Yes	No	Yes	0.015	2.00E-05	0.3501
Yes	No	Yes	0.02	2.00E-05	0.3936
Yes	No	Yes	0.025	2.00E-05	0.4195
Yes	No	Yes	0.03	2.00E-05	0.4547
Yes	No	Yes	0.035	2.00E-05	0.3372
Yes	Yes	Yes	0.01	2.00E-05	0.3742
Yes	Yes	Yes	0.02	2.00E-05	0.3154
Yes	Yes	Yes	0.025	2.00E-05	0.3849
Yes	Yes	Yes	0.03	2.00E-05	0.3302
Yes	No	No	0.01	1.00E-05	0.3672
Yes	No	No	0.01	5.00E-06	0.393
Yes	No	No	0.015	1.00E-05	0.4213
Yes	No	No	0.015	5.00E-06	0.393
Yes	No	No	0.02	1.00E-05	0.4213
Yes	No	No	0.02	5.00E-06	0.3922
Yes	No	No	0.025	1.00E-05	0.3338
Yes	No	No	0.025	5.00E-06	0.3922
Yes	No	Yes	0.01	1.00E-05	0.3948
Yes	No	Yes	0.01	5.00E-06	0.3303
Yes	No	Yes	0.025	1.00E-05	0.4212
Yes	No	Yes	0.025	5.00E-06	0.3286
Yes	No	Yes	0.03	1.00E-05	0.3897
Yes	No	Yes	0.03	5.00E-06	0.3246

- Top-performing configurations rerun with sequence length 512

Baseline	Weighted Loss	Dropout	Weight Decay	Learning Rate	Test F1
Yes	No	No	No	2.00E-05	0.5532
Yes	No	No	0.01	2.00E-05	0.49
Yes	No	No	0.015	2.00E-05	0.5161
Yes	No	No	0.02	2.00E-05	0.5537
Yes	No	No	0.025	2.00E-05	0.5299
Yes	No	No	0.03	2.00E-05	0.5374
Yes	No	Yes	0.01	2.00E-05	0.3749
Yes	No	Yes	0.02	2.00E-05	0.3567
Yes	No	Yes	0.025	2.00E-05	0.365
Yes	No	Yes	0.03	2.00E-05	0.3127
Yes	No	No	0.01	5.00E-06	0.365
Yes	No	No	0.015	1.00E-05	0.4155
Yes	No	No	0.015	5.00E-06	0.365
Yes	No	No	0.02	1.00E-05	0.4155
Yes	No	No	0.02	5.00E-06	0.324
Yes	No	No	0.025	5.00E-06	0.324
Yes	No	Yes	0.01	1.00E-05	0.3942
Yes	No	Yes	0.025	1.00E-05	0.4057

- Label smoothing experiments with varying smoothing factors for top run configurations
  - Learning Rate 2.00E-5

Baseline	Seq Len	Weighted Loss	Label Smooth	Dropout	Weight Decay	Test F1
Yes	512	No	0.1	No	No	0.3559
Yes	512	No	0.05	No	No	0.4345
Yes	512	No	0.2	No	No	0.5055
Yes	512	No	0.1	No	0.02	0.4406
Yes	512	No	0.05	No	0.02	0.4201
Yes	512	No	0.2	No	0.02	0.4292
Yes	512	No	0.1	Yes	0.02	0.4874
Yes	512	No	0.05	Yes	0.02	0.3881
Yes	512	No	0.2	Yes	0.02	0.4578
Yes	512	No	0.1	No	0.03	0.4742
Yes	512	No	0.05	No	0.03	0.5401
Yes	512	No	0.2	No	0.03	0.4658
Yes	512	No	0.1	No	0.025	0.4837
Yes	512	No	0.05	No	0.025	0.3099
Yes	512	No	0.2	No	0.025	0.4061
Yes	512	No	0.1	No	0.015	0.5365
Yes	512	No	0.05	No	0.015	0.4552
Yes	512	No	0.2	No	0.015	0.3178
Yes	256	No	0.1	No	0.01	0.4521
Yes	256	No	0.05	No	0.01	0.3541
Yes	256	No	0.2	No	0.01	0.4872
Yes	256	No	0.1	No	0.025	0.3548
Yes	256	No	0.05	No	0.025	0.3885
Yes	256	No	0.2	No	0.025	0.5256
Yes	256	No	0.1	No	0.03	0.4016
Yes	256	No	0.05	No	0.03	0.3653
Yes	256	No	0.2	No	0.03	0.3696

– Learning Rate 1.00E-5

Baseline	Seq Len	Weighted Loss	Label Smooth	Dropout	Weight Decay	Test F1
Yes	512	No	0.1	No	No	0.3902
Yes	512	No	0.05	No	No	0.3985
Yes	512	No	0.2	No	No	0.3331
Yes	512	No	0.1	No	0.02	0.3742
Yes	512	No	0.05	No	0.02	0.448
Yes	512	No	0.2	No	0.02	0.3934
Yes	512	No	0.1	Yes	0.02	0.3672
Yes	512	No	0.05	Yes	0.02	0.3385
Yes	512	No	0.2	Yes	0.02	0.3789
Yes	512	No	0.1	No	0.03	0.493
Yes	512	No	0.05	No	0.03	0.382
Yes	512	No	0.2	No	0.03	0.3432
Yes	512	No	0.1	No	0.025	0.371
Yes	512	No	0.05	No	0.025	0.4849
Yes	512	No	0.2	No	0.025	0.3992
Yes	512	No	0.1	No	0.015	0.3742
Yes	512	No	0.05	No	0.015	0.448
Yes	512	No	0.2	No	0.015	0.3934
Yes	256	No	0.1	No	0.01	0.416
Yes	256	No	0.05	No	0.01	0.3707
Yes	256	No	0.2	No	0.01	0.3994
Yes	256	No	0.1	No	0.025	0.4052
Yes	256	No	0.05	No	0.025	0.3672
Yes	256	No	0.2	No	0.025	0.3886
Yes	256	No	0.1	No	0.03	0.4115
Yes	256	No	0.05	No	0.03	0.3491
Yes	256	No	0.2	No	0.03	0.4094

– Learning Rate 5.00E-6

Baseline	Seq Len	Weighted Loss	Label Smooth	Dropout	Weight Decay	Test F1
Yes	512	No	0.1	No	No	0.3959
Yes	512	No	0.05	No	No	0.3743
Yes	512	No	0.2	No	No	0.3310
Yes	512	No	0.1	No	0.02	0.3391
Yes	512	No	0.05	No	0.02	0.3831
Yes	512	No	0.2	No	0.02	0.3312
Yes	512	No	0.1	Yes	0.02	0.3141
Yes	512	No	0.05	Yes	0.02	0.3287
Yes	512	No	0.2	Yes	0.02	0.3306
Yes	512	No	0.1	No	0.03	0.3453
Yes	512	No	0.05	No	0.03	0.3758
Yes	512	No	0.2	No	0.03	0.3309
Yes	512	No	0.1	No	0.025	0.3391
Yes	512	No	0.05	No	0.025	0.3831
Yes	512	No	0.2	No	0.025	0.3312
Yes	512	No	0.1	No	0.015	0.3286
Yes	512	No	0.05	No	0.015	0.3777
Yes	512	No	0.2	No	0.015	0.3324
Yes	256	No	0.1	No	0.01	0.3413
Yes	256	No	0.05	No	0.01	0.3948
Yes	256	No	0.2	No	0.01	0.3240

Continued on next page

Baseline	Seq Len	Weighted Loss	Label Smooth	Dropout	Weight Decay	Test F1
Yes	256	No	0.1	No	0.025	0.4013
Yes	256	No	0.05	No	0.025	0.3946
Yes	256	No	0.2	No	0.025	0.3215
Yes	256	No	0.1	No	0.03	0.4010
Yes	256	No	0.05	No	0.03	0.3956
Yes	256	No	0.2	No	0.03	0.3118

- Extended runs (up to 100 epochs) using early stopping for top configurations. This table gives us our top 6 runs that we report in Table 1 (the F1 scores highlighted)

Baseline	Stop Epoch	Seq Len	Weighted Loss	Label Smoothing	Dropout	Weight Decay	Learning Rate	Test F1
Yes	100	512	No	No	No	0.02	2.00E-05	0.5519
Yes	100	512	No	No	No	No	2.00E-05	0.5674
Yes	25	512	No	0.05	No	0.03	2.00E-05	0.5560
Yes	100	512	No	No	No	0.03	2.00E-05	0.5889
Yes	26	512	No	0.1	No	0.015	2.00E-05	0.5177
Yes	100	512	No	No	No	0.025	2.00E-05	0.5876
Yes	63	256	No	0.2	No	0.025	2.00E-05	0.4577
Yes	78	512	No	No	No	0.015	2.00E-05	0.5313
Yes	100	512	No	0.2	No	No	2.00E-05	0.4856

All results reported in the main paper related to BioBERT are drawn from these tables.

QA4RE Few-shot Experiment

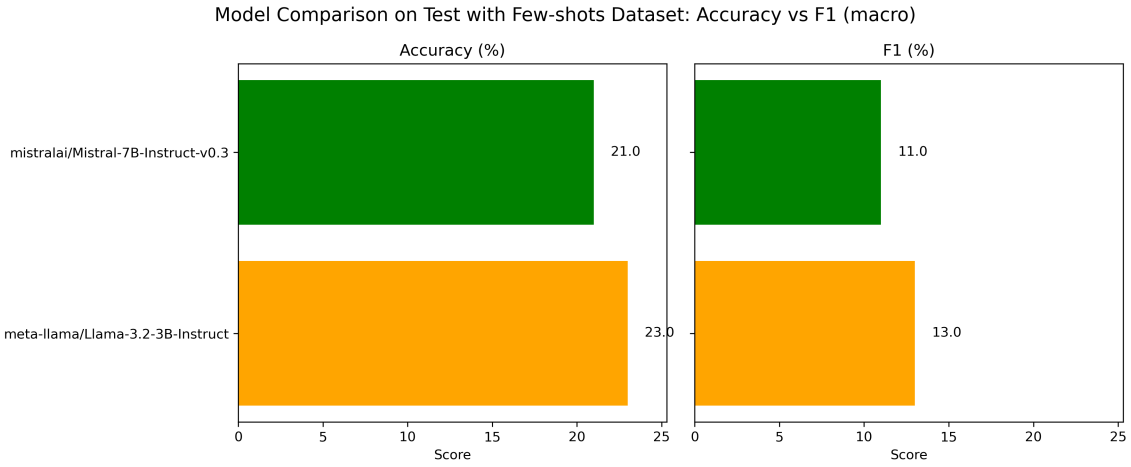


Figure 5: Comparison of models on the Test dataset with few shots (Accuracy and F1-score).