Biomedical Relation Extraction using LLMs and Knowledge Graphs

Pranav Chellagurki[§]
Computer Engineering Department
San José State University
San José, United States
pranav.chellagurki@sjsu.edu

Sai Prasanna Kumar Kumaru[§]
Computer Engineering Department
San José State University
San José, United States
saiprasannakumar.kumaru@sjsu.edu

Rahul Raghava Peela[§]
Computer Engineering Department
San José State University
San José, United States
rahulraghava.peela@sjsu.edu

Neeharika Yeluri[§]
Computer Engineering Department
San José State University
San José, United States
neeharika.yeluri@sjsu.edu

Carlos Rojas
Computer Engineering Department
San José State University
San José, United States
carlos.rojas@sjsu.edu

Jorjeta Jetcheva
Computer Engineering Department
San José State University
San José, United States
jorjeta.jetcheva@sjsu.edu

***Abstract*—Due to the rapid growth of research papers on biomedical topics, it has become increasingly important to make advancements in biomedical Natural Language Processing (NLP). Biomedical NLP enables us to extract important information from text, such as new insights into the role of different genes in disease susceptibility, or the potential for drug therapies that are effective against one disease to work effectively against another. In this paper, we present a comparative evaluation of the binary relation classification capabilities of the current state-of-the-art binary relation classifier, BioBERT [1], against recently released open-source large language models, Gemma-2b, Gemma-7b, and Llama2-7b, which we fine-tune with the benchmark GAD [2] and EU-ADR [3] datasets. In addition, we quantify the potential of discovering new relationships by utilizing knowledge graphs built out of known binary relations.**

***Index Terms*—relation extraction, biomedical natural language processing, biomedical natural language processing, LLMs**

I. Introduction

Over a million new papers focusing on life sciences and biomedical topics are generated each year and made accessible through online sources such as Pubmed[1]. The knowledge encapsulated in each paper could unlock new insights into the role of different genes in susceptibility to disease, or the potential for drug therapies that are effective against one disease to work effectively against another.

The sheer quantity of research papers that have been published makes it difficult for individuals to review directly and requires automated mechanisms to extract key information using Natural Language Processing (NLP) techniques. Fortunately, over the last few years, we have seen a revolution in NLP capabilities, starting with some of the early transformer-based large language models (LLMs) models such as BERT [4], and the recently released GPT-4 [5], Llama2 [6] and Gemma [7] models. However, LLMs typically target general-purpose language capabilities and are thus trained on text sources such as BookCorpus [8] and Wikipedia, leaving them with limited capabilities when tested against the specialized language found in biomedical texts. Fine-tuning these models with biomedical text has been shown to lead to improved performance when performing biomedical NLP tasks in the context of BERT, e.g. [1], [9], [10]. We observe this effect in our own experiments as well (Section IV).

We utilize this fine-tuning approach in our work and focus on exploring the performance, after fine-tuning with biomedical text, of the latest open-source LLMs, including Gemma-2b and Gemma-7b [7], and Llama2-7b [6], where 2b and 7b signify that the model has 2 billion and 7 billion parameters respectively. These are relatively small LLMs and were chosen because they are open-source, well-documented, and have manageable compute requirements, thus enabling us to deploy them on our High-Performance Compute infrastructure.

We focus on the relation extraction (RE) biomedical NLP task, which typically involves classifying relationships between entities such as genes, diseases, drugs ,and proteins [11]. In particular, a relationship can be positive, e.g., the BRCA1 gene is associated with breast cancer, or negative, and thus requires the model to perform a binary classification. Relation extraction is a key task in biomedical NLP where identifying and classifying complex relationships between entities can play a key role in knowledge discovery.

In addition, we quantify the potential of discovering new relationships by building a knowledge graph of currently known binary relationships and then querying the graph for multi-hop relationships that straddle entities connected connected along a path in the graph (rather than directly in a binary relation). Knowledge graphs are also a convenient medium for human scientists to explore biomedical knowledge interactively and derive insights and ideas for future scientific exploration.

We summarize our key contributions below:

- Comparative evaluation of the binary relation classification capabilities of the current state-of-the-art binary

[§]Equal contribution
[1]https://www.nlm.nih.gov/bsd/medline_cit_counts_yr_pub.html

relation classifier, BioBERT [1], against recently released open-source small LLMs, including Gemma-2b, Gemma-7b, and Llama2-7b [6], which we fine-tuned with the benchmark GAD [2] and EU-ADR [12] datasets.
- Quantifying the potential of discovering new relationships by utilizing a knowledge graph built out of known binary relations from the GAD and EU-ADR datasets.

The rest of this paper is organized as follows. We overview related work in Section II. Section III covers our methodology, including model and dataset details, evaluation metrics, and experimental setup. We discuss our results, and conclusion in Section IV and Section V.

II. Related Work

A. Language Models Developments

A variety of NLP models have been developed in recent years, starting with Word2Vec in 2013 [13], and followed by BERT [4] in 2019, and most recently GPT-4 [5] to name just a few prominent examples. These models focus on developing general-purpose language understanding that can be applied to a variety of NLP tasks, e.g., named entity recognition, relation extraction, question-answering, etc. However, %The distinct advantage that modern techniques have is the ability to apply to many problem domains, e.g., named entity recognition, relation extraction, question-answering, etc.

In the area of relation extraction in particular, Papanikolaou et al. introduced a BERT-derived relation extraction approach without direct supervision [14]. Rather than relying on gold data or distant supervision, Papanikolaou et al.'s work capitalizes on syntactic parsing and pre-trained word embeddings to discern precise relations. These relationships subsequently inform the annotation of a broader corpus, which is utilized to fine-tune an established BERT [4] model. Models such as ELMo [15] and BERT [4] learns contextual word embeddings and requires large datasets for pre-training which usually focus on general-purpose language understanding but are not able to capture domain-specific terminology. We discuss models and techniques that augment these models with biomedical language processing capabilities next.

B. Biomedical Language Processing

The development of SciBERT by Beltagy et al. [9], leveraged a large corpus of scientific publications to pre-train a model tailored for scientific texts. This specialized model, distinct from general domain BERT with its SciVocaB vocabulary, demonstrated superior performance in a variety of scientific NLP tasks. Notably, SciBERT achieved state-of-the-art results in many areas such as Named Entity Recognition (NER), and text classification. They significantly outperform BERT-Base with notable F1 score improvements both with and without fine-tuning. SciBERT is tailored more towards broad scientific texts rather than strictly biomedical content.

BioBERT by Lee et al. [1] is a transformer-based model that is pre-trained on a large biomedical paper abstracts corpus (pulled from PubMed and PMC [16]). BioBERT shares the same architecture as BERT [4], but it outperforms previous state-of-the-art models, including BERT and SciBERT, with significant margins in a variety of biomedical NLP tasks, including named entity recognition, relation extraction, and question answering. This is why we chose to use BioBERT as our state-of-the-art baseline model in this work.

Instead of focusing on domain-specific datasets and pre-training, UmlsBERT by Michalopoulos et al. [17] augmented the model with domain-specific knowledge in the form of contextual embeddings. UmlsBERT outperformed previous BERT-based methods at named entity recognition and medical inference tasks, but does not specifically target relation extraction tasks.

The work by Gu et al. [18] demonstrates that transformer-based models can be trained from scratch in domains such as biomedical texts when sufficiently large datasets are available. However, assembling large annotated datasets is still not feasible for many domain-specific tasks.

C. State-Of-The-Art Large Language Models

Large language models such as Llama2-7b [6], Gemma [7], GPT-3 [19], and GPT-4 [5] are defining the cutting edge in various NLP tasks, including those specific to the biomedical domain. These models are proving indispensable in their ability to process large and complex datasets, deliver insights, and automate tasks that traditionally require significant human labor.

The utilization of GPT-4 in gene-associated disease [20] discovery exemplifies the potential of LLMs in biomedical research. GPT-4 facilitates extensive literature searches, summarizes findings, and identifies diseases linked to specific genes, using literature from databases like PubMed. This ability to analyze and synthesize large volumes of data not only speeds up the research process but also enhances the accuracy and depth of disease-gene correlation studies. Similarly, GPT-3's role in generating high-quality summaries aids researchers in quickly assimilating diverse information, making it a valuable tool for multi-document summarization tasks in biomedical research [20].

These applications highlight a common theme across LLMs transforming the approach to managing biomedical literature and data. For instance, in the context of Named Entity Recognition (NER), models like GPT-NER and PromptNER [21] demonstrate advancements in text generation and few-shot learning, respectively. GPT-NER's approach to converting sequence labeling tasks into text generation challenges, and PromptNER's use of entity type definitions for training on minimal data, show how LLMs can adapt to specific NLP tasks under resource constraints.

Moreover, the research detailed in "Lost in the Middle: How Language Models Use Long Contexts" explores how models from the Llama2-7b series (7b, 13b, and 70b) [22] manage tasks that require processing long texts. This work is critical in understanding how LLMs handle long contexts, such as multi-document question answering or key-value retrieval, which are common challenges in biomedical research. The findings from this study are crucial for optimizing LLMs to handle the

extensive and complex documents typical of medical literature, ensuring that relevant information is effectively utilized irrespective of its position within the text.

Choosing L and Gemma for our research in biomedical relation extraction was motivated by these models' proven capabilities in handling tasks that align closely with the needs of biomedical applications. While LLMs like GPT-3 and GPT-4 have demonstrated their utility in summarization and literature analysis, Llama2-7b and Gemma are open-source and require significantly less compute infrastructure, enabling us to conduct experiments in our High-Performance Cluster setup. Our work aims to extend the model's capabilities into the realm of biomedical relation extraction, an area that, while still developing, promises significant advancements in how biomedical relationships are identified and understood. By leveraging Llama2-7b and Gemma, we are not only harnessing state-of-the-art technology but also pioneering its application in a field ripe for innovation.

D. Knowledge Graphs in NLP Applications

Knowledge graphs (KGs) have become increasingly popular in NLP tasks. For example, several studies focus on providing knowledge graph embeddings [23], [24] as contextual information to the LLM to improve its accuracy and reduce hallucinations. Similarly, work by Si et al. [25] applied a similar technique to clinical embeddings to achieve novel results, and KEPLER by Wang et al. [26] aims to learn via a language model and knowledge embeddings.

The work presented in coLAKE [27] combined training a language model and KG network, and BioByGANS [28] propose a graph neural network for named entity recognition. Finally, LinkBERT by Yasunaga et al. [29] creates a single context from all of the linked content in a single article (Wikipedia and PubMed abstract) to provide greater contextual information.

All of these methods focus on providing context (inputs) to the language model. To the best of our knowledge, ours is the first work to explore the use of a knowledge graph to help derive new relations and to allow users to explore relations and potentially identify new relations to study further.

III. Methodology

In this section, we provide an overview of the machine learning models we use for biomedical relation extraction, the datasets we utilize to fine-tune the models, our evaluation metrics, and experimental setup.

A. Datasets

We use two commonly used benchmark datasets, Gene Association Database (GAD) [2] and European Annotated drugs, diseases, targets, and their Relationships (EU-ADR) [12], to fine-tune our models. These datasets were chosen because they are commonly used to benchmark biomedical NLP tasks. We describe each in more detail next, along with some examples. The properties of both datasets are summarized in Table I.

1) GAD: The Genetic Association Database (GAD) [2] dataset is commonly used for research in the field of genetics and genomics. The dataset was curated to document relationships between genes and genetic mutations with diseases. These relations provide valuable insights for researchers and clinicians as they investigate the involvement of different genes in disease likelihood and progression.

The GAD dataset contains a total of 53,300 relations (Table I), each of which is described as a natural language sentence, and assigned a label of 1 to indicate a positive relationship between the gene and the disease, or 0 to indicate there is no relationship. Below is an example data point from the dataset:

- **Sentence:** Mutations in the BRCA1 gene can cause breast cancer.
  - @GENE$ - BRCA1
  - @DISEASE$ - breast cancer

  Label: 1 (indicates positive relationship)
- **Sentence:** The BRCA1 gene is not associated with lung cancer.

  Label: 0 (indicates negative (no) relationship)

2) EU-ADR: The EU-ADR corpus focuses on extracting relations between drugs, disorders (which can include diseases), and targets (which often include genes, as they are common targets in Pharmacogenomics).

The EU-ADR dataset contains a total of 3,550 relations (Table I), each of which is described as a natural language sentence, and assigned a label of 1 to indicate a positive relationship between the gene and a disease, or 0 to indicate there is no relationship. Below is an example data point from the dataset:

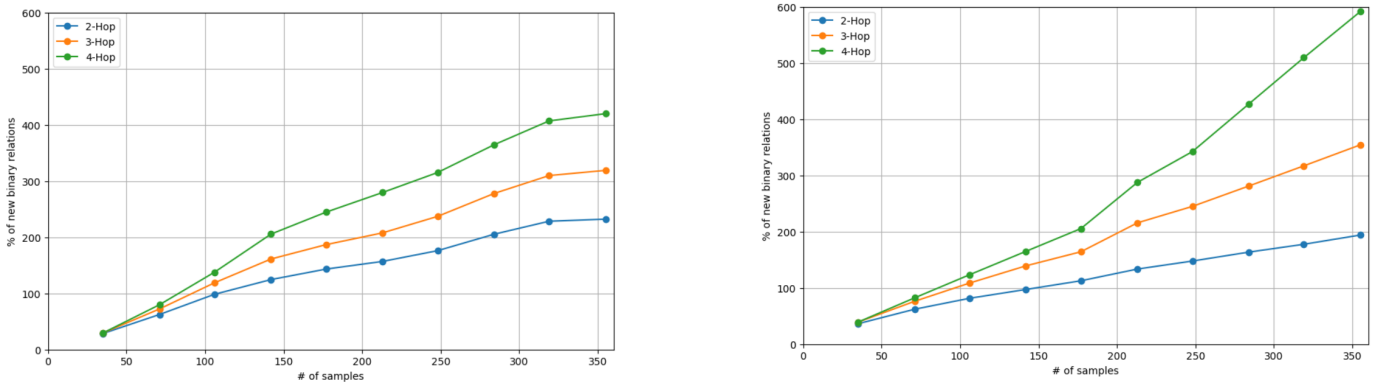
Example sentences and labels:

- **Sentence:** Our work supported @GENE$ genetic variants as possible susceptibility factors for @DISEASE$ and fractures in humans.
- **Labels:**
  - @GENE$ - LRP5
  - @DISEASE$ - osteoporosis
- **Label:** 1 (indicates a positive relationship between the gene and disease)
- **Label:** 0 (indicates negative (no) relationship)

TABLE I: Summary of Dataset Characteristics. "pos." stands for positive class (relation), and "neg." stands for negative class (no relation).

Dataset	**Training Data**			**Test Data**		
	total #	**pos.**	**neg.**	**total #**	**pos.**	**neg.**
GAD	47,970	25,209	22,761	5,330	2,801	2,529
EUADR	3,195	2,358	837	355	262	93

3) Data Preprocessing: In preparation for model training, we pre-processed GAD and EU-ADR to align them to the input requirements of the models. This process involved cleaning and standardizing the textual content, including entity masking which is a substitution of named entity names with generic



(a) % of new relations retrieved from KG for EU-ADR

(b) % of new relations retrieved from KG for GAD

Fig. 1: Illustrates the number of additional relations when we build a knowledge graph from the pair relations. Figure 1a shows for each hop (number of edges from a node) how many more edges we have on the EU-ADR dataset. Similarly, visualized the hops for GAD with Figure 1b.

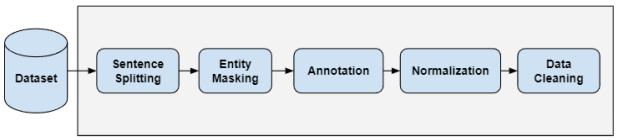


Fig. 2: Data pre-processing steps for BioBERT input data.

placeholders such as "GENE1" or "DRUG2" (Figure 2). This approach was designed to aid the model in learning to recognize and extract significant relationships between entities.

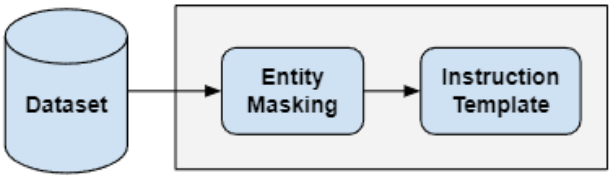


Fig. 3: Data pre-processing steps for Gemma and Llama input data.

For the training of Llama-2-7b and the Gemma models, a standard template is utilized to structure the input data (Figure 3), which ensures consistency and effectiveness in model training:

- Instruction: A clear directive is given to the model specifying the task at hand, e.g., "You will be given a sentence. Your job is to return a boolean variable True/False indicating if there is a relation between the two masked variables."
- Sentence: A sample sentence from the dataset where entities are masked, e.g., "this study proposes that A/A genotype at position -607 in @GENE$ gene can be used as a new genetic marker in Thai population for predicting @DISEASE$ development."
- Response: The expected outcome is based on the presence or absence of a relation, e.g., 'True' or 'False'.

This template ensures that each input instance is formatted consistently, providing clear instructions, and expected outcomes to aid in effective model training. The use of this structured input format helps reduce ambiguity and improves the model's ability to generalize from training data to real-world applications.

4) Training and test set data selection: Due to computational limitations, our study utilized a subset of the GAD dataset comprising 20,000 samples, as opposed to the full dataset of over 40,000 entries. This decision was guided by practical considerations regarding the training time and computational resources available. Training language models (LLMs) on the complete dataset was estimated to require more than 24 hours, while a reduced dataset of 20,000 samples could be processed within approximately 8 hours, allowing for a more efficient use of our computational resources.

Despite these limitations, the subset was carefully selected to maintain a balanced representation of positive and negative relations, ensuring that the integrity and diversity of the dataset were preserved for the training process. Additionally, our approach involved using a substantial validation set to ensure the reliability and generalizability of our results.

The efficiency of this approach is underscored by the performance metrics achieved, which surpassed the baseline model even with the reduced dataset. This outcome suggests that the size of the training set, while important, can be optimized without compromising the model's ability to learn and generalize effectively. This finding aligns with current understandings in machine learning, where the quality of the dataset can be as crucial as its size, and where model architecture and training strategies can mitigate limitations in data volume."

In addition, in our knowledge graph experiments, we utilize a portion of the GAD dataset focusing on gene-disease relations.

B. Machine Learning Models

Pre-training and fine-tuning on domain-specific datasets such as GAD [2] and EU-ADR [12] have been shown to work well [1], [9], [10], [30]. We fine-tuned Llama2-7b [6] with GAD, and EU-ADR with the aim of outperforming BERT-based models such as BioBERT [1]. This process involves adapting LLaMA 2 to the specialized lexicon and contextual nuances of biomedical texts, thereby enhancing its precision and recall in relation extraction tasks.

1) BioBERT: BioBERT [1] is pre-trained on biomedical texts from PubMed abstracts and articles majorly from Wikipedia and Bookscorpus [8].

We fine-tune BioBERT on each of our datasets, GAD and EU-ADR and evaluate how well it performs on binary relation identification. We present the results in Section IV.

2) Llama: Fine-tuning Llama-7b on biomedical datasets, such as GAD and EU-ADR, could significantly enhance its capabilities in biomedical relation extraction. The process involves using datasets that have been structured to align with Llama-2-7b's prompt format, allowing for more efficient learning and adaptation to the biomedical domain. This process typically starts with installing libraries and setting up the model configuration using available resources like Hugging Face, followed by the fine-tuning process which may involve techniques such as 4-bit quantization through QLoRA for efficient training on consumer hardware.

3) Gemma: The Gemma [7] models from Google are a new generation of open source models released this year. They come in two sizes, 2b and 7b, which are tailored for different platforms, from mobile devices to desktop computers. Gemma models can perform text generation and be fine-tuned for specific tasks.

TABLE II: Comparison of Zero-shot model performance (No fine-tuning).

Datasets	Metric	BioBERT	G-2b	G-7b	Llama-7b
GAD	P	52.84	24.03	47.10	27.47
	R	50.11	40.58	49.80	49.37
	F1	33.23	30.00	35.62	34.20
EU-ADR	P	40.73	34.65	36.86	36.56
	R	42.57	40.07	49.80	48.28
	F1	29.22	37.16	42.37	41.61

C. Knowledge Graph

We developed a knowledge graph to synthesize and visualize the predictive analytics from two distinct datasets, GAD (Genetic Association Database) and EU-ADR (Exploring and Understanding Adverse Drug Reactions), leveraging a combination of Python for data manipulation, Flask for web application deployment, and the vis-network JavaScript library for interactive graph visualization. Python, renowned for its robust data manipulation capabilities, is utilized to preprocess and analyze the CSV files containing the datasets. The pandas library, integral to our data processing phase, allows for efficient reading, cleaning, and transformation of these datasets into a structured format suitable for graph-based representation.

Flask, a lightweight and flexible web application framework, serves as the backbone of our application, facilitating the dynamic rendering of the knowledge graph on a web interface. It provides the necessary tools to create a user-friendly environment where users can interact with the data, explore various relationships, and derive insights through an intuitive graphical interface.

The vis-network library is employed to construct the knowledge graph, chosen for its ability to handle complex networks and provide real-time, interactive visualizations. This JavaScript library enables users to not only view but also manipulate the graph directly in the web browser, enhancing the interactivity of the analysis. Nodes in the graph represent entities such as genetic markers or drug reactions, while edges depict the relationships or predictions derived from our analysis of the GAD and EU-ADR datasets.

Together, these technologies form a cohesive system that not only supports the rigorous demands of data processing and analysis but also ensures that the results are accessible and actionable to researchers, clinicians, and policymakers. By integrating these tools, our approach addresses the complex challenge of translating vast amounts of biomedical data into comprehensible and interactive visualizations, thereby contributing to more informed decision-making in genetic research and drug safety surveillance

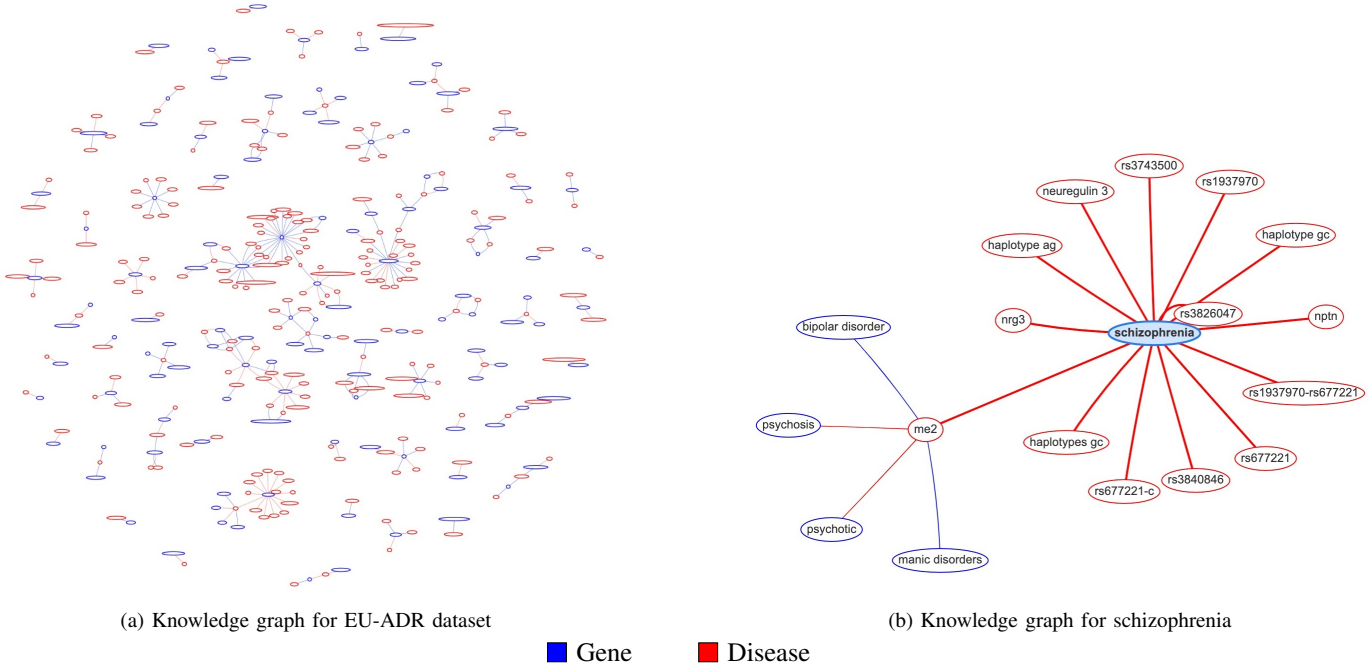
We explore the use of knowledge graphs for discovering new relations in the data.

Following the extraction of entities and subsequent identification of relations among pertinent entities via Large Language Models, a knowledge graph is constructed. This graph visually represents the connections between entities, with lines denoting established relationships. In instances where entities lack a direct relationship, these connections are notably absent from the graph.

D. Metrics

We use precision, recall, and f1-score to evaluate the binary classification performance of our models. For each test sentence (sample) each model needs to assign a label of 1 (positive) and 0 (negative) indicating the relationship between the entities referenced in the sentence. The formulas for each of the metrics are provided below. Precision and recall are expressed in terms of true positives (TP), false positives (FP), and false negatives (FN), where TP indicates the number of test samples for which the classifier correctly predicted the positive class, FP indicates the number of test samples for which the classifier predicted the positive class (label 1) but the correct prediction would have been the negative class (label 0), and FN is the number of test samples for which he classifier predicted the negative class but the correct prediction would have been the positive class.

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$



(a) Knowledge graph for EU-ADR dataset

(b) Knowledge graph for schizophrenia

Fig. 4: Demonstrates a visualization of the EU-ADR relations. The colors indicate if the disease (red) or genes (blue). In Figure 4a we see the complete graph. Figure 4b demonstrates the many connections that researchers can glean from visualizing the relations.

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

In biomedical relation extraction, employing metrics like Precision, Recall, and F1 Score is essential for evaluating the effectiveness of models in identifying accurate relationships between biomedical entities. Precision measures the proportion of correctly identified relationships out of all relationships that the model predicted, indicating the model's accuracy and reliability in pinpointing relevant connections. Conversely, Recall assesses the model's ability to capture all relevant relationships within the dataset, reflecting its comprehensiveness in detecting crucial biomedical interactions. The F1 Score, which is the harmonic mean of Precision and Recall, provides a balanced measure of the model's performance, considering both the accuracy and completeness of the extracted relationships. These metrics are invaluable as they collectively offer a nuanced understanding of a model's effectiveness in accurately and comprehensively identifying significant relationships crucial for advancing biomedical research and applications.

E. Experimental Setup

1) Experiments: We outline the experiments we conduct below:

- **Experiment 1:** Binary relation extraction comparative evaluation. In this experiment, we are exploring the question of how the current state-of-the-art model, BioBERT performs relative recently released relatively small (in terms of number of parameters) large language models, Llama-2 (7b parameter version), Gemma (2b and 7b parameter version). A summary of the binary relation experiments is shown in Figure 5.

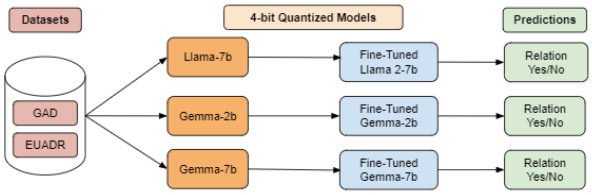


Fig. 5: Summary of Experiments

- **Experiment 2:** Knowledge graph-based relation extraction. We conduct experiments to better understand the potential of knowledge graphs built from binary relations to be used to easily identify additional relations (beyond the ones used to construct them), in addition to being used for knowledge discovery through its visual interactive user interface (Figure 4a). To this end, we quantify the relations that we can discover by querying the knowledge graph and capturing entities that are related to each other across multiple edges (or hops) in the graph. We query the graph in order to discover relationships over 2, 3, and 4 hops.
These newly discovered relations can be used to automatically validate whether these exist in the biomedical

text - searching for a known (or suspected) relation is an easier problem than extracting relations from scratch.

We discuss the results of our experiments in Section IV.

2) Compute Configuration: The high-performance computing (HPC) system utilized for our experiments was equipped with NVIDIA A100-PCIE-40GB graphics processing units (GPUs), featuring a total of four GPUs. The experiments leveraged CUDA Version 12.1. We successfully trained several advanced models including Llama2-7b, Gemma-2b, and Gemma-7b.

In our study, we utilize the BitsAndBytes library to implement a 4-bit quantized format across several models—Gemma 2b, Gemma 7b, and Llama2-7b. This approach is chosen to balance computational efficiency and maintain the precision required for complex natural language processing tasks. Each model is uniformly configured, ensuring that tokenizers align with specific model requirements, and correctly setting padding tokens to support seamless sequence processing. The datasets for training and evaluation are structured with a 70/30 split, and we maintain a separate testing dataset that accounts for at least 10% of the training data.

For optimizing training dynamics, we apply Low-Rank Adaptation (LoRA) which targets essential model components to enhance parameter efficiency without significant computational burden. We carefully set training parameters to optimize resource usage, including a batch size of four, a single gradient accumulation step, and a gradient clipping norm of 0.3. The optimization employs a 32-bit AdamW optimizer, with precise adjustments in learning rate and weight decay to refine model responses. These configurations are supported by a sophisticated training environment equipped with TensorBoard, facilitating detailed monitoring and fine-tuning of the models to ensure peak performance.

We experimented with training the models for 5, 10, and 25 epochs and report our performance results for 25 epochs where we observed the highest performance. This structured approach to training is designed not just to manage computational resources effectively but also to enhance each model's capability to accurately extract and analyze gene-disease relationships, highlighting their potential in advancing biomedical text mining and relation extraction.

IV. Results

A. Experiment 1: Binary relation extraction

We first evaluate our models without fine-tuning them on our biomedical datasets and find that they perform poorly (Figure II) and thus fine-tuning is necessary. This is further evidenced by the much higher performance of the models after fine tuning (Table III).

Overall we find that BioBERT performs substantially better than the small-sized LLM, Gemma-2b in terms of f1 score across both datasets and also substantially worse than the Llama2-7b model (Figure 6).

Interestingly, BioBERT performs better on the EU-ADR dataset relative to the GAD dataset, while all the larger models (7b models) perform worse.

This may be an indication that the models have complementary properties and ensembling them may result in higher performance than using each of them individually.

We look at the model's precision and recall results to get a deeper understanding of where each of them is challenged. We observe that BioBERT has a high recall on EU-ADR, whereas both Gemma models have a very low recall.

While EU-ADR is substantially smaller than GAD, our experiments confirm that it is able to reveal weaknesses in the models' performance that would not have been obvious if we only performed experiments with the GAD dataset.

Upon further analysis of the datasets, we found a discrepancy in vocabulary coverage between the datasets and the pre-trained models - only 58.3% of the EU-ADR vocabulary is represented in the models' pre-trained vocabularies compared to 66% for GAD. This higher number of biomedical terms in EU-ADR is at least in part responsible for the challenges it presents to the models.

B. Experiment 2: Knowledge graph-based relation extraction

For EU-ADR, we find that we are able to identify roughly twice as many relations if we use a 2-hop search, three times as many in a 3-hop search, and 4-times as many in a 4-hop search (Figure 1).

Moreover, as the number of binary relations that are added to the graph grows, the connectivity of the graph becomes richer and the number of new relations that can be discovered via a multi-hop search grows by a factor of 5 (for 2-hop search) to 10 (for 4-hop search).

The effect of discovering relations across more hops is even more pronounced in the knowledge graph built for the GAD dataset (Figure 1). We illustrate here only a subset of GAD as the original number of relations (over 120 thousand) created a substantial computational challenge. Keeping the number of relations the same also allows us to do a better head-to-head comparison.

TABLE III: Comparison of Model Performances

Datasets	Metric	BioBERT	G-2b	G-7b	LlaMA2-7b
GAD	P	78.83	72.02	**97.86**	**99.78**
	R	78.18	70.00	**97.94**	**99.80**
	F1	78.27	69.76	**97.89**	**99.79**
EU-ADR	P	67.87	**93.11**	**98.33**	89.44
	R	70.14	65.35	63.35	**94.88**
	F1	68.73	74.39	74.39	**93.18**

V. Conclusions

In this paper, we present a comparative evaluation of the binary relation classification capabilities of the current state-of-the-art binary relation classifier, BioBERT [1], against LLMs, Gemma-2b, Gemma-7b, and Llama2-7b, which we fine-tuned with the benchmark GAD [2] and EU-ADR [12] datasets.

- **GAD Dataset**: Llama2-7b outperformed all other models with a precision of 99.78%, a recall of 99.80%, and an F1-score of 99.79%. This highlights its superior capacity to leverage context within biomedical literature.

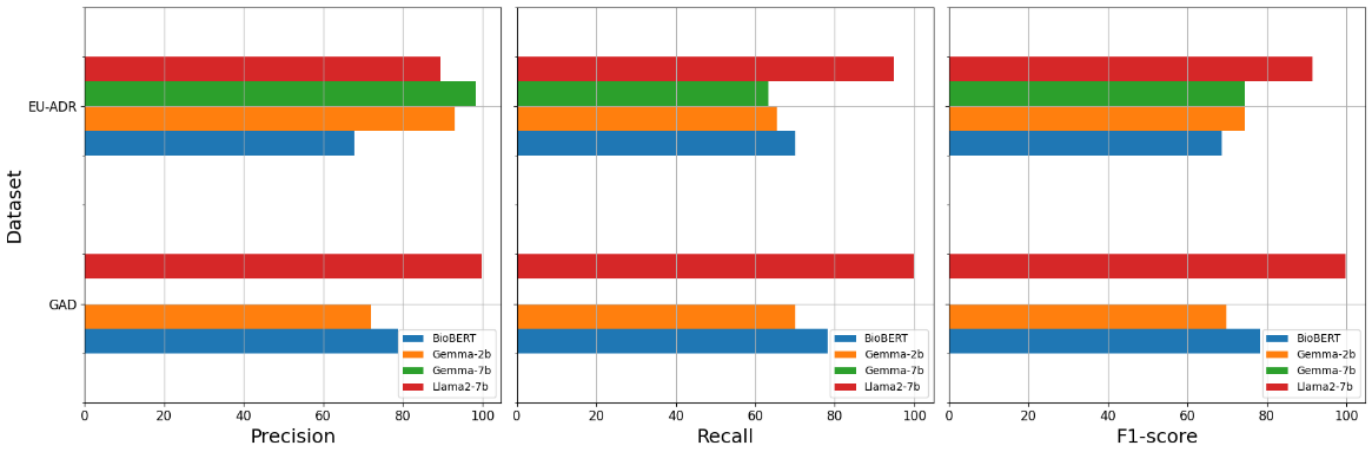


Fig. 6: Performance comparison of fine-tuned models on GAD and EU-ADR

In contrast, BioBERT, while robust, showed lower metrics (Precision: 78.83%, Recall: 78.18%, F1: 78.27%), suggesting limitations in handling more complex entity relations without extensive fine-tuning.

- **EU-ADR Dataset**: Gemma-7b exhibited the highest precision at 98.33% but had lower recall rates (63.35%), indicating a propensity to prioritize accuracy over recall. Llama2-7b maintained a more balanced profile with a precision of 89.44% and recall of 94.88%, leading to an F1-score of 93.18%, underscoring its effectiveness in balancing both metrics.

In our study, we explored the potential of uncovering new relationships by using a knowledge graph constructed from known binary relations. Our results(Figure 1a) show a notable increase in the discovery of additional relationships when extending the search across multiple hops. Specifically, in the EU-ADR dataset, the percentage of new relationships detected grows with the sample size. For instance, at a sample size of 35 (2-hop), we observed an additional 28.57% of relationships, which rose to 29.52% by the 4-hop. This trend intensifies with larger sample sizes, culminating at a sample size of 355, where the additional relationships discovered at the 2-hop increased from 232.21% to 420% by the 4-hop.

The GAD dataset exhibits a similar, yet more pronounced pattern (Figure 1b). These findings highlight the effectiveness of multi-hop searches in knowledge graphs, revealing deeper and sometimes hidden relationships within the data. The consistent growth in the detection of relationships across increasing hops and varying sample sizes demonstrates the models' ability to utilize extensive connective information, paving the way for significant advancements in biomedical research through sophisticated computational methods.

VI. Future work

The trade-off between smaller, fully fine-tuned models and larger models with partial tuning is a central challenge in the above-proposed methodology. Smaller models are more computationally efficient and quicker to train, making them suitable for settings with limited hardware resources. However, they may struggle with the depth of contextual understanding required for complex datasets. Conversely, larger models offer enhanced reasoning capabilities due to their more extensive neural architectures, but often lack full fine-tuning, particularly in their tokenizer components, due to the high computational costs involved. This can lead to inadequate handling of specialized vocabularies, especially in fields like biomedicine where terms frequently evolve.

Future research should thus focus on comprehensive fine-tuning of larger models, including their tokenizers, to fully harness their advanced reasoning abilities. Reducing computational demands through more efficient algorithms or innovative hardware solutions, alongside leveraging distributed computing, could make extensive training more feasible. Additionally, developing adaptive tokenization techniques that adjust to the specialized vocabularies of various datasets would help maintain the effectiveness of large models across different domains, ensuring they stay robust in the face of rapidly changing data landscapes.

REFERENCES

[1] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, p. btz682, Sep. 2019. [Online]. Available: https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btz682/5566506

[2] K. G. Becker, K. C. Barnes, T. J. Bright, and S. A. Wang, "The genetic association database," vol. 36, no. 5, pp. 431–432.

[3] E. M. van Mulligen, A. Fourrier-Reglat, D. Gurwitz, M. Molokhia, A. Nieto, G. Trifiro, J. A. Kors, and L. I. Furlong, "The eu-adr corpus: Annotated drugs, diseases, targets, and their relationships," *Journal of Biomedical Informatics*, vol. 45, no. 5, pp. 879–884, 2012, text Mining and Natural Language Processing in Pharmacogenomics. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1532046412000573

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. [Online]. Available: http://arxiv.org/abs/1810.04805

[5] OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, R. Avila, I. Babuschkin, S. Balaji, V. Balcom, P. Baltescu, H. Bao, M. Bavarian, J. Belgum, I. Bello, J. Berdine, G. Bernadett-Shapiro, C. Berner, L. Bogdonoff, O. Boiko, M. Boyd, A.-L. Brakman, G. Brockman, T. Brooks, M. Brundage, K. Button, T. Cai, R. Campbell, A. Cann, B. Carey, C. Carlson, R. Carmichael, B. Chan, C. Chang, F. Chantzis, D. Chen, S. Chen, R. Chen, J. Chen, M. Chen, B. Chess, C. Cho, C. Chu, H. W. Chung, D. Cummings, J. Currier, Y. Dai, C. Decareaux, T. Degry, N. Deutsch, D. Deville, A. Dhar, D. Dohan, S. Dowling, S. Dunning, A. Ecoffet, A. Eleti, T. Eloundou, D. Farhi, L. Fedus, N. Felix, S. P. Fishman, J. Forte, I. Fulford, L. Gao, E. Georges, C. Gibson, V. Goel, T. Gogineni, G. Goh, R. Gontijo-Lopes, J. Gordon, M. Grafstein, S. Gray, R. Greene, J. Gross, S. S. Gu, Y. Guo, C. Hallacy, J. Han, J. Harris, Y. He, M. Heaton, J. Heidecke, C. Hesse, A. Hickey, W. Hickey, P. Hoeschele, B. Houghton, K. Hsu, S. Hu, X. Hu, J. Huizinga, S. Jain, S. Jain, J. Jang, A. Jiang, R. Jiang, H. Jin, D. Jin, S. Jomoto, B. Jonn, H. Jun, T. Kaftan, Kaiser, A. Kamali, I. Kanitscheider, N. S. Keskar, T. Khan, L. Kilpatrick, J. W. Kim, C. Kim, Y. Kim, J. H. Kirchner, J. Kiros, M. Knight, D. Kokotajlo, Kondraciuk, A. Kondrich, A. Konstantinidis, K. Kosic, G. Krueger, V. Kuo, M. Lampe, I. Lan, T. Lee, J. Leike, J. Leung, D. Levy, C. M. Li, R. Lim, M. Lin, S. Lin, M. Litwin, T. Lopez, R. Lowe, P. Lue, A. Makanju, K. Malfacini, S. Manning, T. Markov, Y. Markovski, B. Martin, K. Mayer, A. Mayne, B. McGrew, S. M. McKinney, C. McLeavey, P. McMillan, J. McNeil, D. Medina, A. Mehta, J. Menick, L. Metz, A. Mishchenko, P. Mishkin, V. Monaco, E. Morikawa, D. Mossing, T. Mu, M. Murati, O. Murk, D. Mély, A. Nair, R. Nakano, R. Nayak, A. Neelakantan, R. Ngo, H. Noh, L. Ouyang, C. O'Keefe, J. Pachocki, A. Paino, J. Palermo, A. Pantuliano, G. Parascandolo, J. Parish, E. Parparita, A. Passos, M. Pavlov, A. Peng, A. Perelman, F. d. A. B. Peres, M. Petrov, H. P. d. O. Pinto, Michael, Pokorny, M. Pokrass, V. H. Pong, T. Powell, A. Power, B. Power, E. Proehl, R. Puri, A. Radford, J. Rae, A. Ramesh, C. Raymond, F. Real, K. Rimbach, C. Ross, B. Rotsted, H. Roussez, N. Ryder, M. Saltarelli, T. Sanders, S. Santurkar, G. Sastry, H. Schmidt, D. Schnurr, J. Schulman, D. Selsam, K. Sheppard, T. Sherbakov, J. Shieh, S. Shoker, P. Shyam, S. Sidor, E. Sigler, M. Simens, J. Sitkin, K. Slama, I. Sohl, B. Sokolowsky, Y. Song, N. Staudacher, F. P. Such, N. Summers, I. Sutskever, J. Tang, N. Tezak, M. B. Thompson, P. Tillet, A. Tootoonchian, E. Tseng, P. Tuggle, N. Turley, J. Tworek, J. F. C. Uribe, A. Vallone, A. Vijayvergiya, C. Voss, C. Wainwright, J. J. Wang, A. Wang, B. Wang, J. Ward, J. Wei, C. J. Weinmann, A. Welihinda, P. Welinder, J. Weng, L. Weng, M. Wiethoff, D. Willner, C. Winter, S. Wolrich, H. Wong, L. Workman, S. Wu, J. Wu, M. Wu, K. Xiao, T. Xu, S. Yoo, K. Yu, Q. Yuan, W. Zaremba, R. Zellers, C. Zhang, M. Zhang, S. Zhao, T. Zheng, J. Zhuang, W. Zhuk, and B. Zoph. GPT-4 Technical Report. [Online]. Available: http://arxiv.org/abs/2303.08774

[6] Llama 2: Open Foundation and Fine-Tuned Chat Models. [Online]. Available: http://arxiv.org/abs/2307.09288

[7] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love, P. Tafti, L. Hussenot, P. G. Sessa, A. Chowdhery, A. Roberts, A. Barua, A. Botev, A. Castro-Ros, A. Slone, A. Héliou, A. Tacchetti, A. Bulanova, A. Paterson, B. Tsai, B. Shahriari, C. L. Lan, C. A. Choquette-Choo, C. Crepy, D. Cer, D. Ippolito, D. Reid, E. Buchatskaya, E. Ni, E. Noland, G. Yan, G. Tucker, G.-C. Muraru, G. Rozhdestvenskiy, H. Michalewski, I. Tenney, I. Grishchenko, J. Austin, J. Keeling, J. Labanowski, J.-B. Lespiau, J. Stanway, J. Brennan, J. Chen, J. Ferret, J. Chiu, J. Mao-Jones, K. Lee, K. Yu, K. Millican, L. L. Sjoesund, L. Lee, L. Dixon, M. Reid, M. Mikuła, M. Wirth, M. Sharman, N. Chinaev, N. Thain, O. Bachem, O. Chang, O. Wahltinez, P. Bailey, P. Michel, P. Yotov, R. Chaabouni, R. Comanescu, R. Jana, R. Anil, R. McIlroy, R. Liu, R. Mullins, S. L. Smith, S. Borgeaud, S. Girgin, S. Douglas, S. Pandya, S. Shakeri, S. De, T. Klimenko, T. Hennigan, V. Feinberg, W. Stokowiec, Y.-h. Chen, Z. Ahmed, Z. Gong, T. Warkentin, L. Peran, M. Giang, C. Farabet, O. Vinyals, J. Dean, K. Kavukcuoglu, D. Hassabis, Z. Ghahramani, D. Eck, J. Barral, F. Pereira, E. Collins, A. Joulin, N. Fiedel, E. Senter, A. Andreev, and K. Kenealy. Gemma: Open Models Based on Gemini Research and Technology. [Online]. Available: http://arxiv.org/abs/2403.08295

[8] Y. Zhu, R. Kiros, R. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning books and movies: Towards story-like visual explanations by watching movies and reading books," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 19–27.

[9] I. Beltagy, K. Lo, and A. Cohan. (2019, Sep.) SciBERT: A Pretrained Language Model for Scientific Text. ArXiv:1903.10676 [cs]. [Online]. Available: http://arxiv.org/abs/1903.10676

[10] X. Zheng, H. Du, X. Luo, F. Tong, W. Song, and D. Zhao, "BioByGANS: biomedical named entity recognition by fusing contextual and syntactic features through graph attention network in node classification framework," *BMC Bioinformatics*, vol. 23, no. 1, p. 501, Nov. 2022. [Online]. Available: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-022-05051-9

[11] R. Xing, J. Luo, and T. Song, "BioRel: Towards large-scale biomedical relation extraction," vol. 21, no. 16, p. 543. [Online]. Available: https://doi.org/10.1186/s12859-020-03889-5

[12] E. M. van Mulligen, A. Fourrier-Reglat, D. Gurwitz, M. Molokhia, A. Nieto, G. Trifiro, J. A. Kors, and L. I. Furlong, "The EU-ADR corpus: Annotated drugs, diseases, targets, and their relationships," vol. 45, no. 5, pp. 879–884. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1532046412000573

[13] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality."

[14] Y. Papanikolaou, I. Roberts, and A. Pierleoni. Deep Bidirectional Transformers for Relation Extraction without Supervision. [Online]. Available: http://arxiv.org/abs/1911.00313

[15] M. E. Peters, M. Neumann, L. Zettlemoyer, and W.-t. Yih. Dissecting Contextual Word Embeddings: Architecture and Representation. [Online]. Available: http://arxiv.org/abs/1808.08949

[16] A. E. Johnson, T. J. Pollard, and L. Shen, "Li wei h. lehman, mengling feng, mohammad ghassemi," *Benjamin Moody, Peter Szolovits, Leo A. Celi, and Roger G. Mark,'MIMIC-III, a freely accessible critical care database', Scientific Data*, vol. 3, p. 160035, 2016.

[17] G. Michalopoulos, Y. Wang, H. Kaka, H. Chen, and A. Wong, "UmlsBERT: Clinical Domain Knowledge Augmentation of Contextual Embeddings Using the Unified Medical Language System Metathesaurus," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, pp. 1744–1753. [Online]. Available: https://aclanthology.org/2021.naacl-main.139

[18] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing," vol. 3, no. 1, pp. 1–23. [Online]. Available: https://dl.acm.org/doi/10.1145/3458754

[19] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[20] J. Chang, S. Wang, C. Ling, Z. Qin, and L. Zhao. (2024) Gene-associated disease discovery powered by large language models. [Online]. Available: https://doi.org/10.48550/arXiv.2401.09490

[21] I. Keraghel, S. Morbieu, and M. Nadif. (2024) A survey on recent advances in named entity recognition. [Online]. Available: https://doi.org/10.48550/arXiv.2401.10825

[22] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni,

and P. Liang. (2024) Lost in the middle: How language models use long contexts. [Online]. Available: https://doi.org/10.1162/tacl_a_00638
[23] Z. Zhang, X. Han, Z. Liu, X. Jiang, M. Sun, and Q. Liu. ERNIE: Enhanced Language Representation with Informative Entities. [Online]. Available: http://arxiv.org/abs/1905.07129
[24] B. He, D. Zhou, J. Xiao, X. jiang, Q. Liu, N. J. Yuan, and T. Xu. Integrating Graph Contextualized Knowledge into Pre-trained Language Models. [Online]. Available: http://arxiv.org/abs/1912.00147
[25] Y. Si, J. Wang, H. Xu, and K. Roberts, "Enhancing clinical concept extraction with contextual embeddings," vol. 26, no. 11, pp. 1297–1304. [Online]. Available: https://academic.oup.com/jamia/article/26/11/1297/5527248
[26] X. Wang, T. Gao, Z. Zhu, Z. Zhang, Z. Liu, J. Li, and J. Tang, "KEPLER: A Unified Model for Knowledge Embedding and Pre-trained Language Representation," vol. 9, pp. 176–194. [Online]. Available: https://direct.mit.edu/tacl/article/doi/10.1162/tacl_a_00360/98089/KEPLER-A-Unified-Model-for-Knowledge-Embedding-and
[27] T. Sun, Y. Shao, X. Qiu, Q. Guo, Y. Hu, X. Huang, and Z. Zhang. CoLAKE: Contextualized Language and Knowledge Embedding. [Online]. Available: http://arxiv.org/abs/2010.00309
[28] X. Zheng, H. Du, X. Luo, F. Tong, W. Song, and D. Zhao, "BioByGANS: Biomedical named entity recognition by fusing contextual and syntactic features through graph attention network in node classification framework," vol. 23, no. 1, p. 501. [Online]. Available: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-022-05051-9
[29] M. Yasunaga, J. Leskovec, and P. Liang. LinkBERT: Pretraining Language Models with Document Links. [Online]. Available: http://arxiv.org/abs/2203.15827
[30] R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, and T.-Y. Liu, "BioGPT: Generative pre-trained transformer for biomedical text generation and mining," vol. 23, no. 6, p. bbac409. [Online]. Available: https://academic.oup.com/bib/article/doi/10.1093/bib/bbac409/6713511