LLM-Powered Knowledge Graph of Causal Relations in Drug Reviews

Vanni Zavarella[1,*], Lorenzo Bertolini[2], Sergio Consoli[2], Gianni Fenu[1], Diego Reforgiato Recupero[1] and Alessandro Zani[2]

[1]Department of Mathematics and Computer Science, University of Cagliari, Cagliari, Italy
[2]European Commission, Joint Research Centre (JRC), Ispra, Italy

Abstract

This paper presents the employment of JSL-MedLlama, a decoder-only Large Language Model (LLM) trained within the medical domain, to create a knowledge graph of causal relationships from drug reviews. We leverage a dataset of causal narratives from clinical notes, MIMICause, to benchmark JSL-MedLlama for classifying causal narratives using instruction fine-tuning. The results show that it obtains satisfying performance, outperforming other encoder-only baselines. Furthermore, we validate our algorithm robustness and cross-domain generalization by testing it on the Drug Reviews dataset, a collection of patient reviews on specific drugs along with related conditions. We then deploy the model on a subset of around 19,000 Drug Reviews, generating a knowledge graph of 3,050 unique triples connecting 1,149 Drugs and 322 Conditions through the considered causal relations. The results highlight the role of decoder-only LLMs, fine-tuned within the biomedical domain, in advancing causal reasoning and generating valuable resources for real-world biomedical use cases. We make publicly available the drug-condition causal relation knowledge graph to support future research efforts in the field.

Keywords

Causality, Large Language Models, Knowledge Graphs, Clinical NLP, Instruction fine-tuning

1. Introduction

Causal relation extraction (CRE), the task of identifying causal relationships between events or entities in text is critical to advance knowledge discovery in the biomedical domain [1, 2]. Causal reasoning methodologies can be broadly classified into two broad paradigms: qualitative and quantitative. Qualitative approaches predominantly conceptualize causal reasoning as a classification task. In contrast, quantitative methods leverage ad-hoc metrics to quantify causal strength, systematically accounting for the inherent uncertainties that pervade causal inference [3].

Extracting causal relationships from a range of diverse unstructured observational data, including electronic health records (EHRs), clinical notes, and online drug reviews, can serve as valuable sources for causal inference experiments, allowing researchers and healthcare professionals to identify potential risk factors, understand disease progression, and assess treatment effectiveness [2, 4, 5, 6]. However, manually analyzing vast amounts of biomedical literature and clinical texts is infeasible, requiring automated approaches for the extraction of causal relationships [7].

In the biomedical domain, several specialized datasets have been introduced. The ACE corpus [8] consists of MEDLINE case reports annotated with mentions of drugs, adverse effects, dosages, and their interrelations. Similarly, BioCause [9] annotates 851 causal relations extracted from 19 open-access biomedical journal articles. More recently, SemEval-2020 Task 5 introduced a large-scale dataset comprising 25,000 statements annotated for counterfactual reasoning, with explicit tagging of antecedents and consequents [10]. Additionally, Track 4 of BioCreative V released manually curated datasets in the

LLM-TEXT2KG 2025: 4th International Workshop on LLM-Integrated Knowledge Graph Generation from Text (Text2KG) Co-located with the Extended Semantic Web Conference (ESWC 2025), June 1 - June 5, 2025, Portoroz, Slovenia.

*Corresponding author.
vanni.zavarella@unica.it (V. Zavarella); lorenzo.bertolini@ec.europa.eu (L. Bertolini); sergio.consoli@ec.europa.eu (S. Consoli); gianni.fenu@unica.it (G. Fenu); diego.reforgiato@unica.it (D. R. Recupero); alessandro.zani@ec.europa.eu (A. Zani)

form of biological knowledge graphs, capturing causal and correlative relationships between entities using BEL (Biological Expression Language) statements [11].

Despite its significance, achieving robust and generalizable performance in CRE remains challenging due to the complexity, variability, and ambiguity of biomedical texts [12, 13, 14]. In recent years, Large Language Models (LLMs) have emerged as powerful tools for solving various NLP tasks, demonstrating remarkable capabilities in understanding and generating text across multiple domains [15, 16, 17]. LLMs, including transformer-based architectures such as GPT [18] and BERT [19] derivatives, have shown promise in improving CRE by leveraging vast biomedical corpora and pre-trained knowledge to recognize complex causal relationships [20].

This paper presents a qualitative approach to causal reasoning by adopting JSL-MedLlama and fine-tuning it using the MIMICause dataset [21], a widely recognized resource for extracting causal relationships in clinical text. We experiment with instruction fine-tuning techniques and compare the resulting model against two strong baselines based on BERT and Clinical-BERT encoders. Our findings show that the tested decoder-only model, fine-tuned on domain-specific biomedical data and further adapted by us to the target task through instruction tuning, achieved satisfying performance and outperformed the considered encoder-only baselines.

Furthermore, we tested our algorithm's robustness and cross-domain generalization on the Drug Reviews dataset, a collection of patient reviews on specific drugs and related conditions. To validate the extracted causal relationships, we annotated a subset of identified instances, achieving high accuracy and strong inter-annotator agreement, confirming the reliability of our approach and its adaptability to real-world biomedical scenarios. Therefore, we deployed the model on a subset of the Drug Reviews dataset, generating a knowledge graph of triples connecting Drug and Condition type entities through four types of causal relations and making the resource publicly available.

The remainder of this paper is structured as follows. Section 2 introduces the task addressed in this work and the dataset used to train our model. Section 3 presents the methodology used to classify causal relations and how it is deployed on the Drug Reviews dataset (Section 3.1). In Section 3.2, we describe the knowledge graph constructed from this dataset and provide relevant analytics. Finally, Section 4 concludes the paper with a summary of findings and directions for future work.

2. Dataset and Task Definition

We train our models to identify causal narratives within clinical notes using the MIMICause dataset [21][1]. The MIMICause dataset is derived from a collection of de-identified discharge summaries sourced from the MIMIC-III (Medical Information Mart for Intensive Care-III) clinical database [22][2], which were annotated for nine types of biomedical entities (*Drug*, *ADE*, *Reason*, *Dosage*, *Strength*, *Form*, *Frequency*, *Route* and *Duration*).

The MIMICause annotation schemas defines that *"a causal relationship/association exists when one or more entities affect another set of entities"* [21]. Eight directed relation types between two entities *e1* and *e2* are defined, where the order of the entity tags determines the direction of causality: *Cause(e1,e2)*, *Cause(e2,e1)*, *Enable(e1,e2)*, *Enable(e2,e1)*, *Prevent(e1,e2)*, *Prevent(e2,e1)*, *Hinder(e1,e2)*, *Hinder(e2,e1).* Additionally, the *Other* class encompasses instances where either a non-causal interaction or no relationship at all exists between a given pair of biomedical entities. For more details on the definitions of the causal relation schema, refer to the original paper [21].

Causal relationships can link entity pairs within the same sentence or, in rare cases, spanning a few sentences in the input text. These relationships may be explicitly signaled by lexical causal connectives, such as "due to", or they may be implicit, requiring inference from the broader context. The MIMICause dataset comprises 2,714 examples, with a train-dev-test split of 1,953 for training, 493 for development, and 268 for testing.

[1] https://huggingface.co/datasets/pensieves/mimicause
[2] Harvard's DBMI Data Portal: https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/

The task of identifying causal relations is formulated as a single-label multi-class relation classification problem:

$$f : (X, e_1, e_2) \rightarrow y$$
$$X = [x_1, x_2, ...x_{n-1}, x_n],$$
$$e_1 = X[i : j] \text{ with } i \leq j \text{ and } i, j \in [1..n],$$
$$e_2 = X[k : l] \text{ with } k \leq l \text{ and } k, l \in [1..n],$$
$$j < k \text{ or } l < i$$

where $y \in [0, ...9]$ is the relation label (4 symmetrical relations plus the Other category), X is an input text sequence, e_1 and e_2 are non-overlapping, continuous token subsequences of X representing the entities between which the causal relation is to be identified (either entity can precede the other).

3. Methods

We perform instruction fine-tuning for classifying causal relations using a SOTA open-source LLM with decoder-only architecture. The reference baselines are the two SOTA encoder-only architectures described in [21], both leveraging BERT-based text encoders combined with fully connected feedforward network (FFN) classifier layers. We will refer to them as BERT+Ent and Clinical-BERT+ENT. Among these baselines, the architecture incorporating the domain-specific Clinical-BERT encoder, denoted as `Clinical-BERT+Ent` in Table 1, yields the best performance.

For our experiments, we use *johnsnowlabs/JSL-MedLlama-3-8B-v2.0* (shortened as *JSL-MedLlama*)[3], an advanced model developed by John Snow Labs on top of the Llama-3-8B architecture and specifically tailored for medical and healthcare applications, having undergone fine-tuning on extensive medical literature and datasets[4]. The model is accessible through Hugging Face via the Transformer library, thus making our study fully reproducible.

For our instruction fine-tuning implementation, we first transformed the MIMICause training split into instruction prompts, which include for each training instance references to the *e1* and *e2* input entities; then, we fine-tuned our model on the resulting instruction dataset using the `trainer` class[5] from Hugging FaceGiven the computational limitations of fully fine-tuning large generative models, we employed the Low-Rank Adaptation (LoRA) technique for Parameter-Efficient Fine-Tuning [23]. The resulting model is renamed as *CLiMA* (Causal Linking for Medical Annotation).

	Model	Cause	Enable	Prevent	Hinder	Other	Macro F1
BL	BERT+Ent	-	-	-	-	-	0.54
	Clinical-BERT+Ent	-	-	-	-	-	0.56
FT	*CLiMA*	0.85	0.77	0.845	0.8	0.89	**0.829**

Table 1
F1 scores on the test split of the MIMICause dataset of the fine-tuned *JSL-MedLlama* model, *CLiMA*, against the two encoder model baselines (BL). The Macro F1 values are averaged over the nine causal relation categories.

Table 1 reports the evaluation results, grouped by relation (*E1 causal_rel E2* has been grouped with *E2 causal_rel E1*), with the *JSL-MedLlama*-based *CLiMA* model significantly outperforming the baselines.

[3]https://huggingface.co/johnsnowlabs/JSL-MedLlama-3-8B-v2.0
[4]We opt for using a small-range model in order to operate within the constraints of limited compute resources. We train and run model inferences on a single A100 GPU with 40GB SDRAM, applying 4-bit quantization.
[5]https://huggingface.co/docs/transformers/en/main_classes/trainer

Drug	Condition	Review
ciprofloxacin	urinary tract infection	i had a urinary tract infection so bad that when i pee it smells but when i started taking ciprofloxacin it worked it's a good medicine for a urinary tract infections.
nuvaring	birth control	i tried the nuvaring. this was my first form of any birth control. this was very easy to put inside and very easy to take out. i didn't feel the ring ever. i thought it was amazing until i started to get huge deep pimples. they were impossible to get rid of.
ziana	acne	when i first started using ziana, i only had acne in between my eyebrows, chin, and the nose area. my acne worsened while using it and then it got better. but after about 4 months of using it, it became ineffective. so i now have acne between my eyebrows, chin, cheeks, forehead, and the nose area. its great at first but after a while it made my face even worse than before i used the product.

Table 2
Sample reviews with target entity metadata from the *Drug Reviews* dataset.

Across relation classes, the model exhibits a significantly lower performance for *Enable* and *Hinder*, which tend not to be distinguished from *Cause* and *Prevent*, respectively[6].

We make publicly available the model as LORA adapters, with associated training scripts and hyperparameters settings, in the Hugging Face repository: https://huggingface.co/unica/CLiMA.

3.1. Causal Relations from Drug Reviews

We evaluated the cross-domain generalization capabilities of the tested fine-tuned model by deploying it on the open-source dataset of *Drug Reviews (Druglib.com)*[7] available within the UCI Machine Learning Repository[8]. The *Drug Reviews* dataset contains around 215 thousands patient reviews on specific drugs along with related conditions, crawled from online pharmaceutical review sites. This dataset is distributed under a Creative Commons Attribution 4.0 International (CC BY 4.0) license, which allows for the use, sharing and adaptation of the data for research purposes.

While similar in topic, the reviews in *Drug Reviews* are different in language style from MIMICause, as they contain slang and are not curated. This allows us to test the robustness of our model on the causal relation extraction task. In the *Drug Reviews* dataset, the target Drug and Condition metadata entities are not always explicitly mentioned in the review text. In order to remain compliant with the instruction prompt settings of our fine-tuned *JSL-MedLlama* model, we first filtered a subset of around 19,200 items from *Drug Reviews* where both Drug and Condition entities are matched within the text. Table 2 lists a few examples of reviews from this subset. Subsequently, for our evaluation we deployed the model on a randomly selected sample of 40 reviews for each possible relation: "Cause", "Prevent", "Hinder", "Enable" and "Other", yielding an overall set of 200 relations to be validated.

We evaluated the correctness and directionality of the extracted causal relations, involving three annotators per relation class. The annotators assessed whether the relation was correct, with options being *True*, if the relation *(E1 causal_rel E2)* was supported by the text, *False* if not, or *Swapped Entities* if the relation was correct but with opposite direction (*E2 causal_rel E1*).

We calculated the average pair-wise Cohen κ inter-rater agreement [24] of all three raters, resulting in a value of 0.739, as well as the Fleiss κ_F agreement [25], resulting in a value of 0.728. These values, ranging in $[-1, +1]$, both indicate a substantial level of agreement among the annotators. We then

[6]In MIMICAUSE, *Enable(e1,e2)* means that the emergence, application or increase of *e1* leads to the emergence or increase of *e2* "jointly to a set of other contributing factors".
[7]*Drug Reviews*: https://archive.ics.uci.edu/dataset/461/drug+review+dataset+druglib+com
[8]https://archive.ics.uci.edu/

Metric	Cause	Enable	Prevent	Hinder	Other	Overall
Cohen κ	0.706	0.591	0.831	0.763	0.770	0.739
Fleiss κ_F	0.707	0.576	0.808	0.746	0.741	0.728
Precision	0.70	0.60	0.78	0.60	0.97	0.73

Table 3
Average pair-wise Cohen κ inter-rater agreement, Fleiss κ_F agreement, and precision score in the human annotation of the *Drug Reviews (Druglib.com)* data sample.

applied a majority vote among the three annotators for the 200 samples of causal relations from our model thus forming a small gold standard. Table 3 summarizes the results categorized by type of relation, presenting also the average pair-wise Cohen's κ inter-rater agreement, Fleiss' κ_F agreement, and the precision score achieved for each of the relations within the gold standard.

The achieved overall precision is 0.73. If we disregard the directionality of the extracted relations, the precision slightly increases to 0.76. In both cases, the level of precision is quite satisfactory, as it closely aligns with the algorithm's overall performance on the original MIMICause test dataset, for which it was specifically trained, proving the robustness and generalization capabilities of our model.

The raters found annotating the `Enable` and `Hinder` relations more challenging, resulting in slightly lower agreement and precision scores (both 0.60). This observation aligns with the performance analysis on MIMICause in Section 3, where these two classes achieved slightly lower F1 scores compared to the others.

3.2. Knowledge Graph

We deploy the fine-tuned JSL-MedLama-3-8B-v2.0 on the 19,200 instances subset of *Drug Reviews* and generate a causal drugs knowledge graph (referred to as *CausalDrugsKG*), comprising 19,200 triples. Out of them, roughly 3,000 are distinct (non-reified) triples, connecting 1,149 unique Drug entities and 322 unique Condition entities via the five considered causal relation categories, i.e. *Cause*, *Enable*, *Prevent*, *Hinder* and *Other*. In the corresponding ontology, designed to describe *CausalDrugsKG* (*causaldrugskg-ont* namespace prefix), each extracted claim is successively reified into instances of the `causaldrugskg-ont:Statement` class, with `causaldrugskg-ont:Statement` representing a specific assertion derived from a collection of drug review items. A sample of generated (un-reified) statements is illustrated in Table 4, together with their support. Here, the support is the number of reviews where full triples were matched).

We made publicly available[9] the automatically generated *CausalDrugsKG* graph in Turtle and RDF serialization format in the European Data portal[10]. The direct link is: https://jeodpp.jrc.ec.europa.eu/ftp/jrc-opendata/ETOHA/ETOHA-OPEN/CausalDrugsKG.ttl.

As an illustration of how *CausalDrugsKG* can be queried for retrieving analytical information on target entities, Figure 1 shows a sample SPARQL query that returns all the statements having the target Drug `causaldrugskg:flecainide` as subject, where `causaldrugskg:flecainide` is the knowledge graph entry for the popular antiarrhythmics medication. Figure 2 shows the 10 most frequently occurring Drug and Condition entities in the *CausalDrugsKG* graph, with over 15% of the extracted triples (out of the 19,200) having *birth control* as Condition, followed by *pain*, *depression* and *anxiety*.

4. Conclusions

In this work, we employed *JSL-MedLlama*, a decoder-only LLM for extracting causal relationships from drug reviews, leveraging instruction fine-tuning to enhance its performance. We compared its

[9]Under Creative Commons Attribution 4.0 International (CC BY 4.0).
[10]https://data.jrc.ec.europa.eu/dataset/acebeb4e-9789-4b5c-97ec-292ce14e75d0

Drug	Causal_Relation	Condition	Support
mirena	Cause	birth control	60
accutane	Cause	acne	33
belsomra	Cause	insomnia	4
viibryd	Enable	depression	5
methotrexate	Enable	psoriasis	2
strattera	Enable	adhd	1
lexapro	Prevent	anxiety	269
vyvanse	Prevent	adhd	142
nortriptyline	Prevent	irritable bowel syndrome	4
pristiq	Hinder	depression	17
buspar	Hinder	anxiety	11
amlodipine	Hinder	high blood pressure	4
nexplanon	Other	birth control	406
aviane	Other	birth control	52
belbuca	Other	chronic pain	2

Table 4
Sample statements for the 5 causal relation categories extracted by the fine-tuned *JSL-MedLlama* model, with their support values in the *Drug Reviews* dataset.

```
PREFIX causaldrugskg: <http://causaldrugskg.org/causaldrugskg/resource/>
PREFIX causaldrugskg-ont: <http://causaldrugskg.org/causaldrugskg/ontology#>
SELECT ?statement
FROM <CausalDrugsKG>
WHERE { ?statement a rdf:Statement .
 ?statement rdf:subject causaldrugskg:flecainide . }
```

Figure 1: Sample SPARQL query returning all *CausalDrugsKG* statements with the graph entity `causaldrugskg:flecainide` as `rdf:subject`.

performance against encoder-based baselines using the MIMICause dataset showing how the fine-tuned model achieves superior results in the classification task.

To assess the robustness and cross-domain generalization of our approach, we applied our fine-tuned model to the Drug Reviews dataset, generating *CausalDrugsKG*, a knowledge graph of 3,050 unique triples linking 1,149 drugs to 322 conditions through the five considered causal relation types. The conducted expert annotation on a subset of extracted causal relationships confirmed the accuracy and reliability of the model, reinforcing its applicability to real-world biomedical scenarios.

The results highlight the critical role of LLMs in advancing causal reasoning in the biomedical domain and demonstrate their potential to generate structured knowledge from unstructured patient narratives. To support future research, we publicly release the fine-tuned model as well *CausalDrugsKG*, providing valuable resources for further advancements in biomedical AI.

Acknowledgments

We would like to thank the colleagues of the Digital Health Unit (JRC.F7) at the Joint Research Centre of the European Commission for helpful guidance and support. The views expressed are purely those of the authors and may not in any circumstance be regarded as stating an official position of the European Commission. We acknowledge financial support under the National Recovery and Resilience Plan (NRRP), Mission 4 Component 2 Investment 1.5 - Call for tender No.3277 published on December

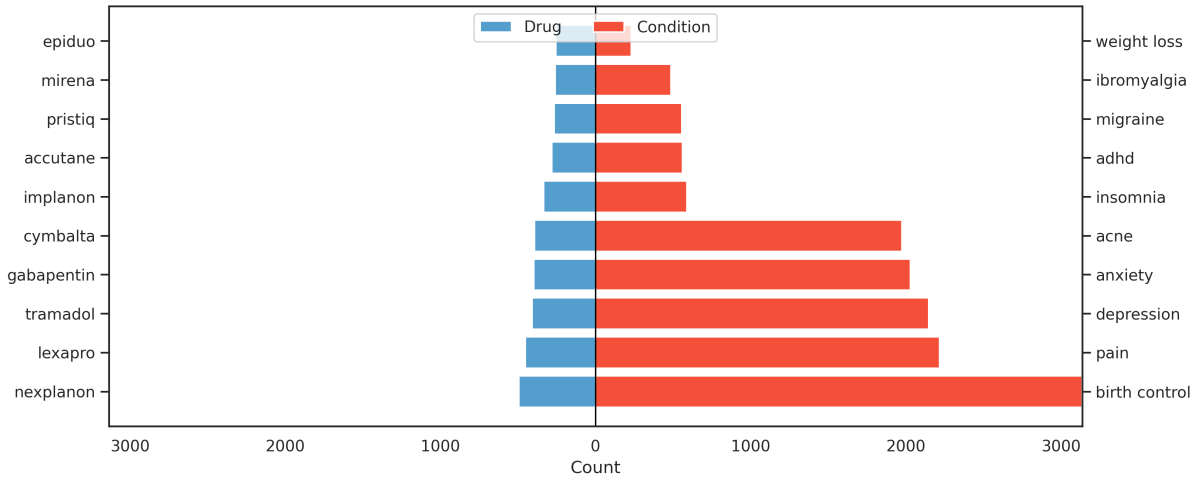

Figure 2: Top 10 Drug and Condition entities in the generated *CausalDrugsKG* graph, with their frequency.

30, 2021 by the Italian Ministry of University and Research (MUR) funded by the European Union – NextGenerationEU. Project Code ECS0000038 – Project Title eINS Ecosystem of Innovation for Next Generation Sardinia – CUP F53C22000430001- Grant Assignment Decree No. 1056 adopted on June 23, 2022 by the Italian Ministry of University and Research (MUR). We also acknowledge the financial support of the project "Data Mesh Platform Builder with AI (DAMPAI)", funded under the "Fondo per la crescita sostenibile" by the "Ministero delle Imprese e del Made in Italy".

Declaration on Generative AI

During the preparation of this work, the author(s) used ChatGPT, Grammarly in order to: Grammar and spelling check. After using this tool/service, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] S. Shimizu, S. Kawano, Special issue: Recent developments in causal inference and machine learning, Behaviormetrika 49 (2022) 275–276. doi:10.1007/s41237-022-00173-z.

[2] A. Akkasi, M.-F. Moens, Causal relationship extraction from biomedical text using deep neural models: A comprehensive survey, Journal of Biomedical Informatics 119 (2021) 103820. doi:https://doi.org/10.1016/j.jbi.2021.103820.

[3] S. Cui, Z. Jin, B. Schölkopf, B. Faltings, The odyssey of commonsense causality: From foundational benchmarks to cutting-edge reasoning, 2024. arXiv:2406.19307.

[4] X. Shen, S. Ma, P. Vemuri, M. R. Castro, P. J. Caraballo, G. J. Simon, A novel method for causal structure discovery from ehr data and its application to type-2 diabetes mellitus, Scientific Reports 11 (2021). doi:10.1038/s41598-021-99990-7.

[5] R. Mozer, A. R. Kaufman, L. A. Celi, L. Miratrix, Leveraging text data for causal inference using electronic health records, 2024. arXiv:2307.03687.

[6] P. Fernainy, A. Cohen, M. E. et al., Rethinking the pros and cons of randomized controlled trials and observational studies in the era of big data and advanced methods: A panel discussion, BMC Proc 18 (Suppl 2) (2024). doi:10.1186/s12919-023-00285-8.

[7] S. Yadav, S. Ramesh, S. Saha, A. Ekbal, Relation extraction from biomedical and clinical text: Unified multitask learning framework, 2020. arXiv:2009.09509.

[8] H. Gurulingappa, A. M. Rajput, A. Roberts, J. Fluck, M. Hofmann-Apitius, L. Toldo, Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from

medical case reports, Journal of Biomedical Informatics 45 (2012) 885–892. doi:https://doi.org/10.1016/j.jbi.2012.04.008.

[9] C. Mihaila, T. Ohta, S. Pyysalo, S. Ananiadou, BioCause: Annotating and analysing causality in the biomedical domain, BMC Bioinformatics 14 (2013) 2 – 2. doi:https://doi.org/10.1186/1471-2105-14-2.

[10] X. Yang, S. Obadinma, H. Zhao, Q. Zhang, S. Matwin, X. Zhu, SemEval-2020 task 5: Counterfactual recognition, in: Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020), Barcelona, Spain, 2020, p. 322–335.

[11] F. Rinaldi, T. R. Ellendorff, S. Madan, S. Clematide, A. van der Lek, T. Mevissen, J. Fluck, BioCreative V track 4: a shared task for the extraction of causal network information using the Biological Expression Language, Database 2016 (2016) baw067. doi:10.1093/database/baw067.

[12] S. R. Sohag, S. M. M. Pasha, Exploring causal relationships in biomedical literature: Methods and challenges, International Journal of Innovative Science and Research Technology (IJISRT) 9 (2025). doi:10.5281/zenodo.14603421.

[13] L. Lishuang, M. Liteng, Z. Beibei, X. Yi, F. Yubo, Q. Xueyang, T. Jingyao, Biomedical event causal relation extraction by reasoning optimal entity relation path, in: M. Sun, J. Liang, X. Han, Z. Liu, Y. He (Eds.), Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 1: Main Conference), Chinese Information Processing Society of China, Taiyuan, China, 2024, pp. 1087–1098. URL: https://aclanthology.org/2024.ccl-1.84/.

[14] J. T. VanSchaik, P. Jain, A. Rajapuri, B. Cheriyan, T. P. Thyvalikakath, S. Chakraborty, Using transfer learning-based causality extraction to mine latent factors for Sjögren's syndrome from biomedical literature, Heliyon 9 (2023) e19265. doi:https://doi.org/10.1016/j.heliyon.2023.e19265.

[15] J. Lehmann, A. Meloni, E. Motta, F. Osborne, D. R. Recupero, A. A. Salatino, S. Vahdati, Large language models for scientific question answering: An extensive analysis of the sciqa benchmark, in: A. Meroño Peñuela, A. Dimou, R. Troncy, O. Hartig, M. Acosta, M. Alam, H. Paulheim, P. Lisena (Eds.), The Semantic Web, Springer Nature Switzerland, Cham, 2024, pp. 199–217.

[16] A. Cadeddu, A. Chessa, V. De Leo, G. Fenu, E. Motta, F. Osborne, D. Reforgiato Recupero, A. Salatino, L. Secchi, A comparative analysis of knowledge injection strategies for large language models in the scholarly domain, Eng. Appl. Artif. Intell. 133 (2024). doi:10.1016/j.engappai.2024.108166.

[17] M. A. K. Raiaan, M. S. H. Mukta, K. Fatema, N. M. Fahad, S. Sakib, M. M. J. e. a. Mim, A review on large language models: Architectures, applications, taxonomies, open issues and challenges, IEEE Access 12 (2024) 26839–26874. doi:10.1109/ACCESS.2024.3365742.

[18] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners, OpenAI Blog (2019) 1–24. URL: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

[19] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 4171–4186. doi:10.18653/v1/n19-1423.

[20] S. Gopalakrishnan, L. Garbayo, W. Zadrozny, Causality extraction from medical text using large language models (llms), Information 16 (2025). doi:10.3390/info16010013.

[21] V. Khetan, M. I. H. Rizvi, J. Huber, P. Bartusiak, B. Sacaleanu, A. Fano, MIMICause: Representation and automatic extraction of causal relation types from clinical notes, in: Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2022, p. 764 – 773. doi:10.18653/v1/2022.findings-acl.63.

[22] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, R. G. Mark, Mimic-iii, a freely accessible critical care database, Scientific Data 3 (2016) 1–9.

[23] Y. Yu, C.-H. H. Yang, J. Kolehmainen, P. G. Shivakumar, Y. Gu, S. R. R. Ren, Q. Luo, A. Gourav, I.-F. Chen, Y.-C. Liu, T. Dinh, A. G. D. Filimonov, S. Ghosh, A. Stolcke, A. Rastow, I. Bulyko, Low-rank

adaptation of large language model rescoring for parameter-efficient speech recognition, in: 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), IEEE, 2023, p. 1–8. doi:10.1109/asru57964.2023.10389632.

[24] M. L. McHugh, Interrater reliability: The kappa statistic, Biochemia Medica 22 (2012) 276 – 282. doi:10.11613/bm.2012.031.

[25] R. Falotico, P. Quatto, Fleiss' kappa statistic without paradoxes, Quality and Quantity 49 (2015) 463 – 470. doi:10.1007/s11135-014-0003-1.