Prompting or Fine-tuning? Exploring Large Language Models for Causal Graph Validation

Yuni Susanti[1], Nina Holsmoelle[2,†]

[1] Artificial Intelligence Dept., Fujitsu Ltd., Japan
[2] Osaka University, Japan

Abstract
This study explores the capability of Large Language Models (LLMs) to evaluate causality in causal graphs generated by conventional statistical causal discovery methods—a task traditionally reliant on manual assessment by human subject matter experts. To bridge this gap in causality assessment, LLMs are employed to evaluate the causal relationships by determining whether a causal connection between variable pairs can be inferred from textual context. Our study compares two approaches: (1) prompting-based method for zero and few-shot causal inference (*unsupervised*) and, (2) fine-tuning language models for the causal relation prediction task (*supervised*). While prompt-based LLMs have demonstrated versatility across various NLP tasks, our experiments on biomedical and general-domain datasets show that fine-tuned models consistently outperform them, achieving up to a 20.5-point improvement in F1 score—even when using smaller-parameter language models. These findings provide valuable insights into the strengths and limitations of both approaches for causal graph evaluation.

Keywords
large language models, causal discovery, prompt engineering

1. Introduction

Uncovering underlying causal relationships is a fundamental task across various scientific disciplines, as these relationships form the basis for understanding and decision-making. Statistical causal discovery methods [1, 2] estimate causal structures from observational data, generating causal graphs that visualize these relationships, as illustrated in Figure 1.

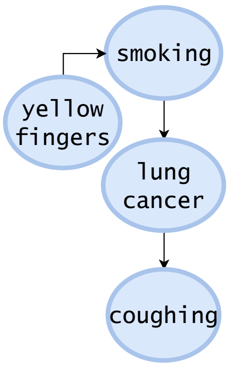

Figure 1: Illustration of a causal graph.

Despite significant advancements in causal discovery, a major challenge remains: verifying the accuracy of causal graphs produced by these predominantly *unsupervised* methods. Typically, this verification relies on domain experts manually validating the graphs, often through controlled experiments. However, such experiments can be prohibitively expensive or, in some fields, entirely unfeasible due to ethical constraints. This highlights the pressing need for alternative, scalable methods to verify causal graphs.

Another approach to verifying causal graphs is using external knowledge from text sources. Causal information is widely distributed across diverse sources, making it an invaluable resource for assisting human experts in validating the accuracy of causal graphs. However, as the number of variables in a causal graph

Workshop on Causal Neuro-symbolic Artificial Intelligence, June 01–5, 2025, Portoroz, Slovenia
† Work performed while at Fujitsu Limited (internship).
susanti.yuni@fujitsu.com (Y. Susanti); n.holsmoelle@gmail.com (N. Holsmoelle)

grows and the volume of textual information expands rapidly, manual verification becomes increasingly impractical. Natural Language Processing (NLP) technologies, including Large Language Models (LLMs) like BERT [3] and ChatGPT, offer a promising solution. These models infer causal relationships between node pairs by leveraging their pre-trained knowledge to analyze the relevant textual context.

In this work, we examine the feasibility of applying NLP technologies to automate causal graph verification. Through quantitative evaluation on causal text datasets, we investigate the performance of two distinct types of NLP models: (1) pre-trained language models fine-tuned for the task of causal relation classification (*supervised*), and (2) prompt-based LLMs (*unsupervised*). To sum up the results, prompt-based LLMs do not necessarily perform better than supervised models on this task, despite their promising performance on diverse clinical NLP tasks [4]. We conduct a detailed analysis to explore the potential factors contributing to this performance gap. Our findings offer valuable insights into the strengths and limitations of these approaches for scalable, automated causal graph validation.

2. Related Work

The research on causal relation extraction/classification from text sources has been done mostly in supervised setting, especially in biomedical-chemistry domains [5, 6, 7, 8, 9, 10], and open-domain [11, 12, 13, 14]. The pre-training and fine-tuning paradigm in NLP led to state-of-the-art performance in many downstream tasks; likewise, most of the related works listed above fine-tune the pre-trained language models such as BERT [15], or propose some sort of enhancement for BERT such as the work by [16, 10]. Their results on relation extraction on biomedical datasets has been encouraging, motivating us to choose BERT as the model for our fine-tuning experiments. On the other hand, recent works [17, 18, 19, 20, 21, 22] show that Large Language Models (LLMs) effectively provide background knowledge for causal discovery, and their findings suggest that LLM-based prompting methods achieved superior performance than non-LLM approaches. For instance, [4] shows that LLMs perform well at zero and few-shot information extraction from clinical text, despite not being trained specifically for the clinical domain. Similarly, other works [23, 24] suggest that LLMs (i.e., InstructGPT [25], ChatGPT, GPT-3.5, etc.), perform well in various downstream tasks even without tuning the parameters, but only with few examples as instructions/prompts. This inspires us to evaluate such instruction, or prompt-based LLMs, on our causal relation classification task. In this work, we compare the prompt-based LLMs against the *more traditional* supervised model where it is trained/fine-tuned using the training data for causal relation classification task.

3. Approach

Given a pair of entities e_1 and e_2 (i.e., node pairs in causal graph such as *smoking* and *lung cancer*), the LLM is tasked with determining whether a causal relationship exists between them. This formulation frames the problem as a *classification* task, where the relation is categorized as either *causal* or *non-causal*. We explore both prompt-based and fine-tuned LLMs, as below:

3.1. Prompt-based LLMs

In prompt-based learning, a pre-trained language model is adapted to a specific task via priming on natural language prompts—pieces of text that are combined with an input and then fed to the language model to produce an output for that task [4]. Prompt-based learning requires the specification of a prompt template to be applied to the input, thus we designed two settings for the prompt-based LLMs experiments: **Zero-Shot Prompt** and **Few-Shot Prompt** settings.

Zero-Shot Prompt. In the `Zero-Shot Prompt` setting, the prompt directly instructs the LLM to answer a causal question about a given pair of entities, without including any training examples—i.e., following a *zero-shot* approach. Formally, we define the model prediction as

$$\hat{y} = \mathcal{M}(P(e_1, e_2, S)), \quad \hat{y} \in \mathcal{Y}$$

where e_1 and e_2 denote the entity pair, S represents the context sentence, $P(\cdot)$ is the prompt construction function, and $\mathcal{M}$ is the LLM. We hand-crafted the following prompt variations.

A: two-choices, no-context

There is a causal relationship between e_1 and e_2. Answer with 'True' or 'False'

B: two-choices, with-context

Given the following context, classify the relationship between e_1 and e_2 as causal or non-causal. Answer with 'causal' or 'non-causal'. Context: S

C: three-choices, with-context

Given the context below, is there a causal relationship between e_1 and e_2. In case only a correlation, but no strict causation between e_1 and e_2, answer with 'False.' In case of uncertainty, answer with 'Maybe.' In a case where there is clearly a causal relationship, and not just a correlation between e_1 and e_2, answer with 'True.' Context: S

The LLMs are strictly constrained to respond with two choices, as in variations (A) and (B), ensuring a fair comparison with the fine-tuned model. However, in variation (C), we allow the LLMs to return a "Maybe" option when they indicate insufficient evidence to determine causality. Additionally, we varied the prompt by either *including* or *omitting* the textual context sentence S, referred to as **with-context** and **no-context**, respectively.

Few-Shot Prompt. In the `Few-Shot` prompt setting, the prompt includes n-number of labeled training examples to guide the model's prediction, allowing the LLMs to process and learn from these examples before making predictions. This method also often referred as in-context learning [26]. Each training example contains: (a) the entity pair, (b) the relation label (*causal* or *non-causal*), and (c) context sentence. Formally, we define:

$$\hat{y} = \mathcal{M}\left(P_n\left(\{(a^i, b^i, S^i, y^i)\}_{i=1}^n, (a^*, b^*, S^*)\right)\right), \quad \hat{y} \in \mathcal{Y}$$

where $\{(e_1^i, e_2^i, S^i, y^i)\}_{i=1}^n$ is the set of n labeled examples, (e_1^*, e_1^*, S^*) denotes the target test data, $P_n(\cdot)$ is the prompt construction function for few-shot learning, and $\mathcal{M}$ is the LLM. Then, as shown below, the LLM is tasked to classify the test data:

Given the context sentence, classify the relationship between the entities marked with e_1 and e_2 as *causal* or *non-causal*
Context Sentence: Expression of <e1> osteopontin </e1> contributes to the progression of <e2> prostate cancer </e2>. **Result**: e1: *osteopontin*, relation: *causal*, e2: *prostate cancer*'
Context Sentence: Increased expression of <e1> cyclin B1 </e1> sensitizes <e1> prostate cancer </e1> cells to apoptosis induced by chemotherapy. **Result**:

Here, one training example ($n = 1$) is embedded in the prompt, highlighted in blue, while the test example is marked in red. The expected output is shown below:

e1: *cyclin B1*, relation: *causal*, e2: *prostate cancer*

We conducted the Few-Shot Prompt experiment by varying the number of the training data n to be included in the prompt.

3.2. Fine-tuned LLMs

The pre-training of LLMs usually utilizes a great quantity of unlabeled data, and the fine-tuning involves training these pre-trained LLMs on a smaller dataset labeled with examples relevant to the target task. By exposing the model to these new labeled examples, the model adjusts its parameters and internal representations suited for the target task. In this work, we experimented with two models: **(1) BERT** [3] to represent ***small* language model** (under 1b parameters) and **(2) GPT** to represent ***larger*-parameter models**.

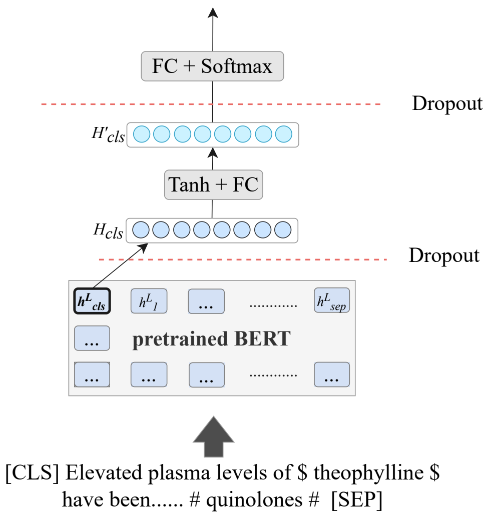

Figure 2: Fine-tuning BERT.

Fine-tuning BERT model BERT is a language model trained using a multi-layer bidirectional transformer encoder and has been demonstrated to enhance various NLP tasks [3]. Figure 2 illustrates our model architecture for fine-tuning the pre-trained BERT model on the causal relation classification task. To ensure a fair comparison with prompt-based LLMs, we adopt a simple fine-tuning approach. Given an input sequence S, we extract its vector representation from the BERT model and utilize the last hidden state of the [CLS] token as the input representation for fine-tuning, following the original paper [3]. We further apply a $Tanh$ activation function and a fully-connected layer (FC) to this representation to obtain the final sequence representation H'_{cls}:

$$H'_{cls} = W_0(tanh(H_{cls})) + b_0 \tag{1}$$

Dropout is applied in the model architecture as a regularization method, as indicated in Figure 2. We used the *binary cross entropy* as the loss function during the training.

Fine-tuning GPT model Fine-tuning the GPT model includes formatting each training example into ***prompt-completion*** pair, where the input example serves as the **prompt**, and the corresponding output serves as the **completion**. The format of these pairs varies depending on the task. While our task is fundamentally a relation *classification* task, it can also be framed as a relation *extraction* task between pairs of entities. We followed GPT fine-tuning instruction and formatted the examples into both task formats:.

A: Fine-tuning GPT, classification format

Prompt: The results provide evidence for altered plasticity of synaptic morphology in memory mutants <e1>dnc</e1> and <e2>rut</e2> and suggest a role...
Completion: non-causal END

B: Fine-tuning GPT, relation extraction format

Prompt: The results provide evidence for altered plasticity of synaptic morphology in memory mutants <e1>dnc</e1> and <e2>rut</e2> and suggest a role...
Completion: dnc rut non-causal END

4. Evaluation

4.1. Datasets and Experiment Settings

We evaluate our approach in (A) **biomedical domain**, focusing on 3 types of causality: gene-gene (GENEC [22]), drug-side effect (DDI [27]), and gene-disease (COMAGC [28]), and (B) **open-domain** causality dataset SEMEVAL [29]. We used GPT model using OpenAI API with `gpt-3.5-turbo` and `text-davinci-003` engines. For experiments with BERT model, we applied BioBERT [30], PubMedBERT [31] for biomedical dataset and BERT (*large, uncased*) for open-domain dataset. Code, dataset and hyper-parameters settings are provided in our Github.

4.2. Results and Discussion

Table 1 summarizes the evaluation results for the biomedical and open-domain datasets. We report the Precision (P), Recall (R), and F1 scores. We apply 5-fold cross-validation and the scores are averaged. We report the standard deviation values of the F1 scores over the 5-folds as shown in parenthesis in Table 1.

In summary, the results indicate that fine-tuned LLM models significantly outperform prompt-based LLMs, achieving a 12.8–20.5 point improvement in F1 score across all datasets. Notably, even with smaller language models like BERT, fine-tuning leads to a substantial performance boost, with F1 scores improving by up to 20.5 points (67.4 vs. 87.9, Zero-Shot Prompt C vs. PubMedBERT on DDI dataset). This contrasted with the previous studies [23, 24] where LLMs perform relatively well, if not better than the fine-tuned models in various tasks including in clinical NLP tasks [4]. One possible reason that the prompt-based model does not perform as well as the fine-tuned model is that causality is rarely written explicitly with causal cues like *"cause," "causing," or "caused"*. Instead, it is often described more implicitly or ambiguously, using keywords such as *"contribute"* or *"play a role"*. Additionally, by fine-tuning the model with training samples, we expose the model to various ways in which causal relationships can be expressed in text. This suggests that identifying causality patterns from training samples is a

Prompt-based	*type*	(Biomed) COMAGC P	R	F1	(Biomed) DDI P	R	F1	(Biomed) GENE P	R	F1	(News) SEMEVAL P	R	F1
Zero-Shot Prompt	A	28.2	61.0	38.1 (.06)	52.2	25.7	34.3 (.02)	23.6	26.6	24.2 (.07)	64.6	66.0	65.3 (.06)
Zero-Shot Prompt	B	28.2	94.2	43.2 (.05)	65.1	69.0	66.7 (.04)	34.3	59.6	**42.3** (0.1)	77.7	84.7	**80.8** (.04)
Zero-Shot Prompt	C	48.8	100	**64.2** (.14)	52.9	93.2	**67.4** (.02)	27.4	71.9	39.5 (.05)	57.4	82.8	67.7 (.03)
	n												
Few-Shot Prompt	5	37.2	83.5	51.0 (.03)	100	37.6	**53.1** (.15)	22.1	25.7	22.7 (.28)	100	46.0	62.7 (.06)
Few-Shot Prompt	15	52.8	41.4	46.1 (.08)	51.4	27.0	35.1 (.05)	26.0	29.1	26.2 (.18)	100	47.9	**64.6** (.04)
Few-Shot Prompt	20	50.2	70.4	**57.0** (.08)	*	*	*	31.7	39.5	**34.3** (.08)	58.9	57.7	58.2 (.02)
Fine-tuning													
BioBERT		77.9	84.4	80.8 (.01)	97.0	76.2	85.2 (.03)	46.1	65.2	53.5 (.07*)	*	*	*
PubmedBERT		80.7	87.4	**83.9** (.03)	93.2	83.3	**87.9** (.01)	50.6	62.1	**55.1** (.03)	*	*	*
BERT-large		*	*	*	*	*	*	*	*	*	93.0	93.0	93.0 (.01)
GPT (*classification*)		80.5	70.1	74.1 (.06)	99.4	78.1	87.4 (.03)	58.6	23.1	31.4 (.08)	99.9	94.8	**96.8** (.03)
GPT (*extraction*)		75.6	58.1	65.5 (.07)	100	62.9	77.1 (.02)	52.4	21.2	30.1 (.06)	100	91.9	95.7 (.03)

Table 1: Experiment results. Values in **bold** indicates the best F1 score for each method and dataset. $type$ refers to Zero-Shot Prompt variations as explained in 3.1, n refers to the number of training data included in the prompt for Few-Shot setting.

crucial step in accurately recognizing causal relations. Nevertheless, in the Few-Shot Prompt experiments, where n training samples are included in the prompt, the performance does not always improve compared to models without training samples. For example, the scores are 67.4 vs. 53.1 with Zero-Shot Prompt C versus Few-Shot Prompt with n=5 on the DDI dataset. This is illustrated in Table 1, where, surprisingly, the highest F1 score for the prompt-based methods is achieved with Zero-Shot Prompt B and C, both of which do not include any training samples. We hypothesize that the limited size of the training samples may be a factor, and increasing the amount of training data could potentially improve results. In addition, when training examples are too few or chosen poorly, it might confuses the model instead. Further investigation is needed to clarify this point. However, due to the token limitations of the OpenAI API, we were unable to experiment with larger values of n.

Next, we investigated the effect of including the context sentence in the prompt. To do this, we created prompt variations of the Zero-Shot Prompt model by *including* and *not including* the context sentence S in the prompt (**with-context** and **no-context**). The results suggest that, for prompt-based methods, including the context sentence in the prompt can be effective. As shown in Table 1, the Zero-Shot Prompt types B and C (**with-context**) consistently outperform type A (**no-context**). By incorporating context, the model gains additional knowledge to better predict the relationship between the entity pair, rather than relying solely on the information acquired during pre-training. In addition, we observed generally higher scores on the open-domain dataset (96.8 with fine-tuned GPT on SEMEVAL) compared to the biomedical datasets (87.9 with PubMedBERT on DDI). This is expected, as LLMs are predominantly pre-trained on open-domain texts, such as books, articles, and online content. Another contributing factor could be the complexity of biomedical texts, which often include more domain-specific or technical terms compared to open-domain datasets.

5. Conclusion

We present a study exploring the feasibility of applying NLP technologies for causal graph verification. Specifically, we compare prompt-based and fine-tuned LLMs in predicting causality between pairs of entities. Experiments on biomedical and open-domain datasets suggest that fine-tuned models outperform prompt-based LLMs, even with smaller-parameter models like BERT. However, fine-tuned models require sufficient expert-annotated data for training, which can be a significant bottleneck. Constructing training data through expert annotation is often challenging and costly; in this regard, LLMs hold promise as a breakthrough for causal inference research. Due to data limitations, our current evaluation was restricted to pairwise causality validation. Future work will involve extending this approach to multivariate causal graphs, refining prompting strategies, and exploring performance across diverse models.

References

[1] P. Spirtes, C. Glymour, R. Scheines, Causation, Prediction, and Search, 2nd ed., MIT press, mit, 2000.

[2] S. Shimizu, P. O. Hoyer, A. Hyvärinen, A. Kerminen, A linear non-gaussian acyclic model for causal discovery, J. Mach. Learn. Res. 7 (2006) 2003–2030.

[3] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, CoRR abs/1810.04805 (2018) 1161–1186. URL: http://arxiv.org/abs/1810.04805. arXiv:1810.04805.

[4] M. Agrawal, S. Hegselmann, H. Lang, Y. Kim, D. Sontag, Large language models are few-shot clinical information extractors, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 1998–2022. URL: https://aclanthology.org/2022.emnlp-main.130.

[5] C. S. G. Khoo, S. Chan, Y. Niu, Extracting causal knowledge from a medical database using graphical patterns, in: Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, ACL '00, Association for Computational Linguistics, USA, 2000, p. 336–343. URL: https://doi.org/10.3115/1075218.1075261. doi:10.3115/1075218.1075261.

[6] C. Mihăilă, S. Ananiadou, Semi-supervised learning of causal relations in biomedical scientific discourse, BioMedical Engineering OnLine 13 (2014) S1. URL: https://doi.org/10.1186/1475-925X-13-S2-S1. doi:10.1186/1475-925X-13-S2-S1.

[7] J. Gu, L. Qian, G. Zhou, Chemical-induced disease relation extraction with various linguistic features, Database 2016 (2016) 1161–1186. URL: https://doi.org/10.1093/database/baw042, baw042.

[8] I. Reklos, M.-P. Albert, Medicause: Causal relation modelling and extraction from medical publications, in: Proceedings of the Text2KG 2022: International Workshop on Knowledge Graph Generation from Text, Text2KG, Crete,Hersonissos, Greece, 2022, pp. 1161–1186.

[9] V. Khetan, M. I. Rizvi, J. Huber, P. Bartusiak, B. Sacaleanu, A. Fano, MIMICause: Representation and automatic extraction of causal relation types from clinical notes, in: Findings of the Association for Computational Linguistics: ACL 2022, Association for Computa-

tional Linguistics, Dublin, Ireland, 2022, pp. 764–773. URL: https://aclanthology.org/2022.findings-acl.63. doi:10.18653/v1/2022.findings-acl.63.

[10] Y. Susanti, K. Uchino, Causal-evidence graph for causal relation classification, in: Proceedings of the 39th ACM/SIGAPP Symposium on Applied Computing, SAC '24, Association for Computing Machinery, New York, NY, USA, 2024, p. 714–722. URL: https://doi.org/10.1145/3605098.3635894. doi:10.1145/3605098.3635894.

[11] C. S. G. Khoo, J. KORNFILT, R. N. ODDY, S. H. MYAENG, Automatic Extraction of Cause-Effect Information from Newspaper Text Without Knowledge-based Inferencing, Literary and Linguistic Computing 13 (1998) 177–186. URL: https://doi.org/10.1093/llc/13.4.177. doi:10.1093/llc/13.4.177. arXiv:https://academic.oup.com/dsh/article-pdf/13/4/177/10888761/177.pdf.

[12] D.-S. Chang, K.-S. Choi, Incremental cue phrase learning and bootstrapping method for causality extraction using cue phrase and word pair probabilities, Information Processing & Management 42 (2006) 662–678. URL: https://www.sciencedirect.com/science/article/pii/S0306457305000580. doi:https://doi.org/10.1016/j.ipm.2005.04.004.

[13] E. Blanco, N. Castell, D. Moldovan, Causal relation extraction, in: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), European Language Resources Association (ELRA), Marrakech, Morocco, 2008, pp. 1161–1186. URL: http://www.lrec-conf.org/proceedings/lrec2008/pdf/87_paper.pdf.

[14] A. Balashankar, S. Chakraborty, S. Fraiberger, L. Subramanian, Identifying predictive causal factors from news streams, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 2338–2348. URL: https://aclanthology.org/D19-1238. doi:10.18653/v1/D19-1238.

[15] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[16] P. Su, K. Vijay-Shanker, Investigation of improving the pre-training and fine-tuning of bert model for biomedical relation extraction, BMC Bioinformatics 23 (2022) 120. URL: https://doi.org/10.1186/s12859-022-04642-w. doi:10.1186/s12859-022-04642-w.

[17] E. Kıcıman, R. Ness, A. Sharma, C. Tan, Causal reasoning and large language models: Opening a new frontier for causality, 2023.

[18] R. Tu, C. Ma, C. Zhang, Causal-discovery performance of chatgpt in the context of neuropathic pain diagnosis, 2023. arXiv:2301.13819.

[19] M. Willig, M. Zečević, D. S. Dhami, K. Kersting, Can foundation models talk causality?, 2022. arXiv:2206.10591.

[20] C. Zhang, S. Bauer, P. Bennett, J. Gao, W. Gong, A. Hilmkil, J. Jennings, C. Ma, T. Minka, N. Pawlowski, J. Vaughan, Understanding causality with large language models: Feasibility and opportunities, 2023. arXiv:2304.05524.

[21] J. Gao, X. Ding, B. Qin, T. Liu, Is ChatGPT a good causal reasoner? a comprehen-

sive evaluation, in: H. Bouamor, J. Pino, K. Bali (Eds.), Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 11111–11126. URL: https://aclanthology.org/2023.findings-emnlp.743. doi:10.18653/v1/2023.findings-emnlp.743.

[22] Y. Susanti, M. Färber, Knowledge graph structure as prompt: Improving small language models capabilities for knowledge-based causal discovery, in: G. Demartini, K. Hose, M. Acosta, M. Palmonari, G. Cheng, H. Skaf-Molli, N. Ferranti, D. Hernández, A. Hogan (Eds.), The Semantic Web – ISWC 2024, Springer Nature Switzerland, Cham, 2025, pp. 87–106.

[23] X. Wei, X. Cui, N. Cheng, X. Wang, X. Zhang, S. Huang, P. Xie, J. Xu, Y. Chen, M. Zhang, Y. Jiang, W. Han, Zero-shot information extraction via chatting with chatgpt, 2023. arXiv:2302.10205.

[24] K. Jeblick, B. Schachtner, J. Dexl, A. Mittermeier, A. T. Stüber, J. Topalis, T. Weber, P. Wesp, B. Sabel, J. Ricke, M. Ingrisch, Chatgpt makes medicine easy to swallow: An exploratory case study on simplified radiology reports, 2022. arXiv:2212.14882.

[25] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, R. Lowe, Training language models to follow instructions with human feedback, 2022. arXiv:2203.02155.

[26] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: Proceedings of the 34th International Conference on Neural Information Processing Systems, NIPS '20, Curran Associates Inc., Red Hook, NY, USA, 2020, pp. 1877 – 1901.

[27] M. Herrero-Zazo, I. Segura-Bedmar, P. Martínez, T. Declerck, The ddi corpus: An annotated corpus with pharmacological substances and drug–drug interactions, Journal of Biomedical Informatics 46 (2013) 914–920. URL: https://www.sciencedirect.com/science/article/pii/S1532046413001123. doi:https://doi.org/10.1016/j.jbi.2013.07.011.

[28] H.-J. Lee, S.-H. Shim, M.-R. Song, H. Lee, J. C. Park, Comagc: a corpus with multi-faceted annotations of gene-cancer relations, BMC Bioinformatics 14 (2013) 323. URL: https://doi.org/10.1186/1471-2105-14-323. doi:10.1186/1471-2105-14-323.

[29] I. Hendrickx, S. N. Kim, Z. Kozareva, P. Nakov, D. Ó Séaghdha, S. Padó, M. Pennacchiotti, L. Romano, S. Szpakowicz, SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals, in: Proceedings of the 5th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Uppsala, Sweden, 2010, pp. 33–38. URL: https://aclanthology.org/S10-1006.

[30] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, BioBERT: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (2019) 1234–1240. URL: https://doi.org/10.1093/bioinformatics/btz682. doi:10.1093/bioinformatics/btz682.

[31] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-specific language model pretraining for biomedical natural language processing,

ACM Trans. Comput. Healthcare 3 (2021). URL: https://doi.org/10.1145/3458754. doi:10.1145/3458754.