

On the Limitations of Zero-Shot Classification of Causal Relations by LLMs

Vani Kanjirangat*, Alessandro Antonucci and Marco Zaffalon

Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (IDSIA), USI-SUPSI, Lugano (Switzerland)

Abstract

We aim to explore and analyze the capabilities and limitations of the large language models in understanding and distinguishing causal sentences under a zero-shot setting. We experiment on a multi-class dataset of direct causal, conditional causal, and correlational sentences. In the experiments, the GPT and Falcon models are validated against a fine-tuned BERT model under different settings to explore zero-shot capabilities in causality detection. Zero-shot approaches exhibit good performance in other classification tasks, such as sentiment analysis or question answering. Yet, for this task, the fine-tuned approach seems superior, and the situation does not change if language cues are added or a few-shot setting is considered. This is a preliminary analysis of a work in progress. Still, the results suggest that identifying causal relations is a particularly challenging task that is hard to address in a zero-shot setup.

Keywords

Large language models, zero-shot classification, few-shot classification, causal inference.

1. Introduction

The adoption of *large language models* (LLMs) [1, 2] is rapidly growing, primarily because of the zero-shot capabilities exhibited by these tools in a wide range of natural language processing tasks, such as sentiment analysis or recommendations and knowledge-intensive tasks, such as question answering and domain-specific entity recognition [3, 4, 5, 6, 7]. Despite such popularity, it is essential to understand the limitations and address questions such as: where can these models fall back? What are the possibilities of such fallbacks? How can we improve their performance beyond prompting engineering [8, 9, 10]?

This paper is a preliminary report on our work (in progress) on evaluating the potential of state-of-the-art LLMs in the field of causal inference. More specifically, we investigate the performance of LLMs in a classification task with sentences possibly involving causal relations. Our analysis focuses on zero- and few-shot capabilities of LLMs compared against a fine-tuning setting with encoder-based BERT models, which are nowadays the most common choice for classification tasks [11, 12, 13]. Our tests show some limitations of LLM approaches in the causal domain, this being the case either for zero-shot and few-shot setups. Notably, the situation remains the same even if language cues are provided. Such negative results are in line with some recent works presenting LLMs as *causal parrots* [14], not yet capable of genuine causal reasoning [15], beyond just distinguishing between causes and effects [16, 17].

In: R. Campos, A. Jorge, A. Jatowt, S. Bhatia, M. Litvak (eds.): Proceedings of the Text2Story'24 Workshop, Glasgow (Scotland), 24-March-2024

*Corresponding author: Vani Kanjirangat (vani.kanjirangat@idsia.ch).



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

2. Related Work

Recently, a plethora of research has been going on in the direction of exploiting the zero-shot and few-shot capabilities of LLMs. Because of the vast amount of pre-trained data they have been exposed to, *large* (>10B parameters) language models are considered to have an inherent ability to generalise across unseen tasks [18, 19, 20]. For instance, the number of parameters of the recent GPT-3 and GPT-4 models is about, respectively, 175B and 1.76T. Zero-shot and few-shot techniques have been tried with different prompting strategies (e.g., the *chain of thought*) for both classification and generation tasks. In many knowledge-intensive tasks (e.g., question answering), translations, classification tasks (e.g., sentiment analysis) and recommendations, those approaches seem compelling, provided that an adequate, prompt engineering effort is achieved [21, 22, 23, 24]. These techniques may be inaccurate for many other tasks, significantly when the complexity increases, such as multi-task classifications and hard sequence labelling tasks, especially in domain-specific problems [25, 8]. Researchers have come up with soft prompting approaches and *parameter efficient tuning* (PEFT) [26] approaches such as P-tuning [27], prompt-tuning [28, 29] and variations of prompt infusions to overcome these problems, while trying to achieve fine-tuning-based performances. The causal reasoning ability of LLMs has been initially investigated in [30]. The authors observe a good performance with a pairwise causal discovery task, counterfactual reasoning task and actual causality by conducting experiments on datasets of cause-effect pairs. A critical review of the causality inference and reasoning with LLMs on benchmark datasets is reported in [31]. The authors specify the requirements of causal datasets and problems of evaluations with LLMs, such as *memorisation* (the dataset could be part of LLM pre-training). They also indicate that LLMs can answer many datasets by simply computing similarities between options and questions in a vector space. Further, they also indicate that the good performance of LLMs can sometimes be due to spurious language cues in the datasets. In the rest of the paper, we explore the capability of LLMs with simple prompt-based approaches in identifying causality on a multi-class dataset.

3. Dataset Settings

For our analysis, we focus on the dataset from [32], developed to automate the identification of causal language use in the scientific literature. The data source was a collection of PubMed¹ abstracts with five main health topics – *nutrition*, *diabetes*, *obesity*, *breast cancer*, and *cholesterol*. Two domain experts were asked to annotate the sentences manually. A good agreement (Cohen’s kappa = 0.98) was reported. The original dataset refers to a multi-class setup with four options: *correlational*, *direct causal*, *conditional causal*, and one without any relations [33]. The entities possibly involved in the causal and correlational relations are not provided. Thus, the identification depends on the specific language patterns used in the input sentences. In the correlational case, the sentence describes some association between variables. With direct causal sentences, the cause and effect are directly mentioned, while in the conditional case, the relation definition carries out an element of doubt. Finally, there are sentences with neither causation nor correlation.

¹<https://pubmed.ncbi.nlm.nih.gov>.

We use the original dataset in the native multi-class setting and a binary classification task. For the binary class, we drop the correlational sentences and combine the direct and conditional causal sentences, thus having only two classes, one with no relations and the other with causal relations. This is intended to allow for a focus on causal relation discrimination. The multi-class dataset includes 1356 no-relation, 494 direct causal, 213 conditional and 998 correlational cases, which makes up 3061 cases. In the binary class setting, we have 1356 no-relation cases and 707 cases of causal relations (which combines direct and conditional cases), with 2063 cases overall.

4. Methodology

To test the capability of LLM models in classifying causal and non-causal sentences under zero-shot settings, we initially design suitable prompts to tackle the task with LLMs. We use both binary and multi-class settings with the prompt including the text input in Fig. 1 and some variations. For the binary settings, we just need to change the classes in the prompt.

```

system_msg = You are a helpful assistant for causal reasoning and cause-and-effect relationship discovery.
Your aim is to identify the entities and to categorize the input sentences into either direct causal relation
or conditional causal relation or correlational relation or no relationship exist

intro_msg = You will be provided with a text. Text: <Text>{text}</Text>

instructions_msg = Please read the provided text carefully to comprehend the context and content.

Examine the roles, interactions, and details surrounding the entities within the text.
Based only on the information in the text, categorize the causal relation as

0. no relation
1. direct causal
2. conditional causal
3. correlational

Your response should analyze the situation in a step-by-step manner, ensuring the correctness of the ultimate conclusion,
which should accurately reflect the likely causal connection based on the information presented in the text.
If no clear causal relationship is apparent,
select the appropriate option accordingly, i.e., 'no relation'.

option_choice_msg = Your response should analyze the situation in a step-by-step manner,
ensuring the correctness of the ultimate conclusion,
which should accurately reflect the likely causal connection between the two entities based on the information presented in the text.

If no clear causal relationship is apparent, select the appropriate option accordingly.

Then provide your final answer within the tags <Answer>[answer]</Answer>, (e.g. <Answer>1</Answer>).

```

Figure 1: A zero-shot prompt for a causal recognition task.

Following the indications from [32] and findings from [31], we create another prompt including *language cues* intended to help the LLM in providing more accurate classifications. In fine-tuning approach, we assume the model automatically captures these patterns given the training data. In a zero-shot setting, with the absence of such training information, we want to see the impact on model performance when some explicit domain knowledge is available. We added the following cues – *association, associated with, predictor* for correlational, *increase, decrease, lead to, effective in, contribute to, reduce* for causal and along with *may, might, appear to, probably* for conditional causal. These cues were then added to the zero-shot prompt (ZS-Cues). Further, we tried them in a few-shot setup (FS-Cues) with some examples from each class (e.g., two samples for each class). Finally, we also consider a 500-shot experiment with labelled samples, used also to train the BERT model under the same settings (500 samples for training).

5. Results and Discussion

For the fine-tuning approach, we use the *best-base-cased model* [34] in both binary and multi-class settings with k-fold cross-validation. We use SimpleTransformers² with four epochs and a learning rate of $2E-5$. We experiment with GPT (3.5 Turbo) and the open-source Falcon model (falcon-7b-instruct³ and falcon-40b-instruct) in zero-shot settings. Falcon-7b-instruct is a 7B parameters causal decoder-only model fine-tuned on a mixture of chats and instructions, while falcon-40b-instruct is a bigger model with 40B parameters. From Tab. 1, it can be observed that Falcon-7b and 40b give inferior performance compared to GPT. Comparing the two Falcon models, the 40b outperformed the 7b model in multi-class and binary settings. This expected result motivates us to focus our experiments on GPT models only.

Table 1

Zero-shot (F1) performance of GPT and Falcon LLMs.

Model	Approach	Binary class	Multi-class
GPT 3.5 turbo	Zero shot (ZS)	0.59	0.37
Falcon-7b-instruct	Zero shot (ZS)	0.19	0.27
Falcon-40b-instruct	Zero shot (ZS)	0.26	0.38

For further experiments, we use GPT to analyse the performance under different prompt settings and compare them with fine-tuned BERT-based models. Tab. 2 shows that, in both settings, the performance of GPT under zero-shot settings is poor and the fine-tuned BERT model performs better. In the multi-class case, many conditional causal relations are misclassified as direct causal. Yet, the accuracy does not improve significantly in the binary setting. The addition of cues (ZS-Cues) improves the performance, showing the importance of specific patterns that help in classification, especially with multi-class settings. In both cases, ZS-Cues performed better than FS-Cues. This could be because the sentences in this dataset are quite varied (extracted from the scientific literature) and we cannot pre-assume that the selected sentences for few-shot experiments could be the best representative for a class.

Table 2

Comparison of zero-shot, with and without cues, and few-shot LLMs against fine-tuned BERT.

Model	Approach	Binary class	Multi-class
GPT 3.5 turbo	Zero shot (ZS)	0.59	0.37
GPT 3.5 turbo	Zero shot with Cues (ZS-Cues)	0.66	0.51
GPT 3.5 turbo	Few shot with Cues (FS-Cues)	0.62	0.50
BERT-base-cased	Full Fine Tuning (FFT)	0.92	0.87

Tab. 3 reports more details on the zero-shot results. The relatively high recall values for causal sentences denote a good ability of the model in detecting direct causal relation. Yet, the same does not happen with conditional causal relations, typically misclassified as direct ones. This also explains the higher performance in the binary class case.

²<http://simpletransformers.ai>.

³<https://huggingface.co/tiiuae/falcon-7b-instruct>.

Table 3

Results on multi-class and binary settings for zero-shot classifications.

	Multi-class				Binary	
	No Rel.	Causal	Cond. Causal	Corr.	No Rel.	Causal
F1-score	0.45	0.39	0.12	0.54	0.59	0.58
Precision	0.68	0.27	0.10	0.60	0.85	0.45
Recall	0.34	0.70	0.14	0.48	0.45	0.85

For a deeper comparison against the FFT model, we prompt the GPT model with more examples. An option would be fine-tuning the GPT model, but we keep this as a future study, as here the focus is on prompting approaches. Further, there are restrictions on the number of prompt tokens processed by GPT 3.5 model. As a reasonable prompting solution, we use 500 samples (corresponding to a 1:4 train-test split). The same samples are used to train BERT under multi-class settings. This proportion makes the BERT performance comparable with the one with k-fold cross validation (F1=0.81), while a drastic drop is obtained with a 1:9 split (F1=0.36). For GPT, this setup requires a slight change in the prompt (Fig. 1), to include a list of input text and give the corresponding predictions as a list. We then chunked the remaining 2449 test samples, each containing ten samples, to be passed on to the prompt. These steps are intended to optimise prompt efficiency in terms of costs and time. The results are in Tab. 4.

Table 4

BERT-Based FFT vs. GPT models with 500-shot training samples.

Model	No Rel.	Causal	Cond. Causal	Corr.	Avg.
BERT-base-cased	0.86	0.77	0.74	0.86	0.81
GPT 3.5 turbo	0.61	0.42	0.12	0.55	0.43

It can be observed that, with 500 samples, the performance of GPT model was better than its zero-shot counterpart, but is not comparable with BERT model fine-tuned with the same 500 training samples. This seems to confirm, in the causal domain, the general findings discussed in [35]. At the same time, it is also notable that simply adding pattern information, like in ZS-Cues and FS-Cues, makes the LLM performance better than the 500-shot model.

Moreover, in some cases, GPT gives predictions not explicitly mentioned in the prompts. For instance, the prediction was multi-label (neither of the labels included true prediction) or the prediction was categorised into a new class (not described in the prompt). For the evaluations, we had to remove such samples. With zero-shot multi-class, finally, we had 3055 instances and 2016 instances with binary class.

Some explicit results from the zero-shot experiments are in Figs. 2, 3 and 4. Tab. 5 reports examples of misclassification from the zero-shot prompting of GPT. Some of these instances are hard to classify even for human experts. GPT almost always classifies a negative causality as a no-relation case (I2, 3, and 4). Some other misclassifications (I8) predicted as causal while conditional in practice are too obvious. Similar patterns are observed in correlational cases (I10). It can be also observed that the addition of cues helped in the correct classification of some instances. The correct predictions are in bold. E.g., I7 and I9 show that the cue *may* could have helped in the correct classification as conditional causal, and for I12 the cue *associated with*.

text: However, obesity seems to be associated with more wound complications.

The text states that "obesity seems to be associated with more wound complications."

This suggests a potential causal relationship between obesity and wound complications. To determine the type of causal relationship, we need to analyze the information provided.

1. Direct Causal Relation: This type of causal relation implies that one event directly causes another event. In this case, if obesity directly causes wound complications, we would expect to see a clear cause-effect relationship between the two.
2. Conditional Causal Relation: This type of causal relation implies that one event causes another event under certain conditions or circumstances. In this case, if obesity causes wound complications only under specific conditions, we would need more information to determine the specific conditions.
3. Correlational Relation: This type of relation implies that two events are statistically related or occur together, but one does not necessarily cause the other. In this case, if obesity and wound complications are correlated but not causally related, we would expect to see a statistical association between the two.

Based on the information provided, it is not clear whether the relationship between obesity and wound complications is a direct causal relation, a conditional causal relation, or a correlational relation.

The text only states that obesity is associated with more wound complications, but it does not provide information on the specific nature of the relationship or any conditions under which the relationship occurs.

Therefore, the appropriate categorization for the causal relationship in this text is: <Answer>3</Answer> (correlational)

Figure 2: An example of correct correlational classification.

text: This study is one of the first proofs of concept showing that a nutrition-sensitive agriculture value chain approach can contribute to improved child nutrition in a remote pastoralist population.

To analyze the causal relationship in the provided text, we need to identify the entities involved and understand their interactions.

Entities:

1. Study
2. Nutrition-sensitive agriculture value chain approach
3. Improved child nutrition
4. Remote pastoralist population

Based on the information in the text, let's analyze the causal relationship:

1. The study is one of the first proofs of concept.
2. The study shows that a nutrition-sensitive agriculture value chain approach can contribute to improved child nutrition.
3. The improved child nutrition is in a remote pastoralist population.

From the above analysis, we can conclude that there is a direct causal relationship between the nutrition-sensitive agriculture value chain approach and improved child nutrition in a remote pastoralist population.

Therefore, the causal relationship in the text can be categorized as a direct causal relation. <Answer>1</Answer>

Figure 3: An example of correct classification of a direct causal relation.

text: There are no indications that endogenous and exogenous gonadal hormones affect the radiation dose-response relationship.

The provided text states that there are no indications that endogenous and exogenous gonadal hormones affect the radiation dose-response relationship.

Based on this information, we can conclude that there is no clear causal relationship between endogenous and exogenous gonadal hormones and the radiation dose-response relationship.

Therefore, the answer is <Answer>0</Answer> (no relation).

Figure 4: An example of a direct causal relation misclassified as no relation.

6. Limitations

Finally, as our paper presents the results of a work in progress, let us discuss the limitations of the present work and the possible enhancements we might consider for the ongoing work.

We have used only one dataset; hence, whether our findings remain valid in the general case might be questionable. The dataset facilitates understanding how well LLMs identify the causal descriptions embedded in scientific literature under a more challenging multi-class setting,

Table 5

Sample sentences predictions from different prompts and ground truth (GT) values.

I	Samples	ZS	ZS-C	FS-C	GT
1	CUSE was able to distinguish between benign and malignant breast masses with high sensitivity and specificity.	0	0	0	1
2	The presence of an in situ stent did not interfere with surgery.	0	0	0	1
3	There are no indications that endogenous and exogenous gonadal hormones affect the radiation dose-response relationship.	0	0	0	1
4	BRBC does not improve survival of women with MBC in this study, though longer follow up is warranted.	0	0	3	1
5	The decreasing trends in nutritional status and appetite level after SCI require special attention.	1	0	0	0
6	iCBT for depression is an efficacious, accessible treatment option for people with diabetes.	1	1	0	0
7	CETP gene variation may affect coronary risk apart from the level of HDL-C	0	2	2	2
8	A considerable decline in RHR has occurred in Tromsø over the past two decades in men and women of all ages.	0	3	3	3
9	Corticosteroids may delay the time of onset of severe skin reactions and also reduce the incidence of severe radiation dermatitis	1	1	2	2
10	In HIV-infected patients, daptomycin appears to be a useful agent for treating resistant GPs.	1	1	1	2
11	Adverse drug events resulting in intensive care unit admission in oncology patients are common and often associated with significant morbidity, mortality, and cost.	1	1	1	3
12	In patients with obstructive CAD by CCTA, the baseline use of statins was associated with improved clinical outcomes.	1	3	3	3

including correlative and causal relations. Distinguishing between direct and conditional causation is especially difficult. To the best of our knowledge, there are no datasets with analogous characteristics, at least for multi-class settings. Yet, manually annotating scientific abstracts and creating new benchmarks for deeper validation is a realistic and necessary effort.

Moreover, in the current paper, we have focused on the GPT model and compared it with the open-sourced Falcon and BERT-based models. This can be enhanced by comparing with different LLMS. Further, the focus was on using prompt-based techniques, which have a broad scope to be explored. Based on the finding from [31], we investigate techniques such as incorporating language cues while prompting LLMs. One major problem is that LLMs can be sensitive to manually engineered prompt designs; hence, automating prompts and using soft prompt techniques would be the way forward.

7. Conclusions and Outlooks

This work is a preliminary exploration to understand the capabilities and limitations of the GPT model in causality identifications, specifically in multi-class settings. The experiments show that GPT has limited zero-shot and few-shot capabilities in capturing such causal relations, subject to the data in consideration. Focusing on the limitations, in the future, we would like to enhance our experiments on a range of causal data to have conclusive generalisations on the studied facts. Prompt engineering as such has a lot of potential to be explored, while hard-core engineering of prompts may not be always beneficial. Hence, we also plan to explore PEFT techniques such as soft prompting for causal detection and further extractions of causal graphs.

References

- [1] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, Improving Language Understanding by Generative Pre-Training, Technical Report, OpenAI, 2018.
- [2] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, Y. Iwasawa, Large language models are zero-shot reasoners, *Advances in Neural Information Processing Systems* 35 (2022) 22199–22213.
- [3] L. Shu, H. Xu, B. Liu, J. Chen, Zero-shot aspect-based sentiment analysis, arXiv preprint arXiv:2202.01924 (2022).
- [4] A. Kumar, N. Jain, S. Tripathi, C. Singh, From fully supervised to zero shot settings for twitter hashtag recommendation, arXiv preprint arXiv:1906.04914 (2019).
- [5] D. Teney, A. v. d. Hengel, Zero-shot visual question answering, arXiv preprint arXiv:1611.05546 (2016).
- [6] Z.-Y. Dou, N. Peng, Zero-shot commonsense question answering with cloze translation and consistency optimization, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2022, pp. 10572–10580.
- [7] Y. Yang, A. Katiyar, Simple and effective few-shot named entity recognition with structured nearest neighbor learning, arXiv preprint arXiv:2010.02405 (2020).
- [8] L. Floridi, M. Chiriatti, GPT-3: Its nature, scope, limits, and consequences, *Minds and Machines* 30 (2020) 681–694.
- [9] K. Elkins, J. Chun, Can GPT-3 pass a writer’s Turing test?, *Journal of Cultural Analytics* 5 (2020).
- [10] W. Wang, V. W. Zheng, H. Yu, C. Miao, A survey of zero-shot learning: Settings, methods, and applications, *ACM Transactions on Intelligent Systems and Technology (TIST)* 10 (2019) 1–37.
- [11] X. Wang, G. Chen, G. Qian, P. Gao, X.-Y. Wei, Y. Wang, Y. Tian, W. Gao, Large-scale multi-modal pre-trained models: A comprehensive survey, *Machine Intelligence Research* 20 (2023) 447–482.
- [12] V. Khetan, R. Ramnani, M. Anand, S. Sengupta, A. E. Fano, Causal BERT: Language models for causality detection between events expressed in text, in: *Intelligent Computing: Proceedings of the 2021 Computing Conference*, Volume 1, Springer, 2022, pp. 965–980.
- [13] S. Aftan, H. Shah, A survey on BERT and its applications, in: *2023 20th Learning and Technology Conference (L&T)*, IEEE, 2023, pp. 161–166.
- [14] M. Zečević, M. Willig, D. S. Dhimi, K. Kersting, Causal parrots: Large language models may talk causality but are not causal, *Transactions on Machine Learning Research* (2023).
- [15] C. Zhang, D. Janzing, M. van der Schaar, F. Locatello, P. Spirtes, Causality in the time of LLMs: Round table discussion results of CLeaR 2023, *Proceedings of Machine Learning Research* vol TBD 1 (2023) 7.
- [16] L. Zhiheng, Z. Jin, R. Mihalcea, M. Sachan, B. Schölkopf, Can large language models distinguish cause from effect?, in: *UAI 2022 Workshop on Causal Representation Learning*, 2022.
- [17] A. Antonucci, G. Piqué, M. Zaffalon, Zero-shot causal graph extrapolation from text via LLMs, arXiv preprint arXiv:2312.14670 (2023).
- [18] E. Almazrouei, H. Alobeidli, A. Alshamsi, A. Cappelli, R. Cojocaru, M. Debbah, É. Goffinet, D. Hesslow, J. Launay, Q. Malartic, et al., The falcon series of open language models, arXiv

- preprint arXiv:2311.16867 (2023).
- [19] S. Zhuang, B. Liu, B. Koopman, G. Zuccon, Open-source large language models are strong zero-shot query likelihood models for document ranking, arXiv preprint arXiv:2310.13243 (2023).
 - [20] Z. Wang, Y. Pang, Y. Lin, Large language models are zero-shot text classifiers, arXiv preprint arXiv:2312.01044 (2023).
 - [21] 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* 33 (2020) 1877–1901.
 - [22] S. Kim, S. J. Joo, D. Kim, J. Jang, S. Ye, J. Shin, M. Seo, The CoT collection: Improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning, arXiv preprint arXiv:2305.14045 (2023).
 - [23] Y. Yu, Y. Zhuang, R. Zhang, Y. Meng, J. Shen, C. Zhang, Regen: Zero-shot text classification via training data generation with progressive dense retrieval, arXiv preprint arXiv:2305.10703 (2023).
 - [24] Z. Yang, Z. Gan, J. Wang, X. Hu, Y. Lu, Z. Liu, L. Wang, An empirical study of GPT-3 for few-shot knowledge-based vqa, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 2022, pp. 3081–3089.
 - [25] M. Moradi, K. Blagec, F. Haberl, M. Samwald, GPT-3 models are poor few-shot learners in the biomedical domain, arXiv preprint arXiv:2109.02555 (2021).
 - [26] Z. Fu, H. Yang, A. M.-C. So, W. Lam, L. Bing, N. Collier, On the effectiveness of parameter-efficient fine-tuning, in: *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 2023, pp. 12799–12807.
 - [27] X. Liu, K. Ji, Y. Fu, W. Tam, Z. Du, Z. Yang, J. Tang, P-tuning: Prompt tuning can be comparable to fine-tuning across scales and tasks, in: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2022, pp. 61–68.
 - [28] X. L. Li, P. Liang, Prefix-tuning: Optimizing continuous prompts for generation, arXiv preprint arXiv:2101.00190 (2021).
 - [29] X. Liu, Y. Zheng, Z. Du, M. Ding, Y. Qian, Z. Yang, J. Tang, GPT understands, too, *AI Open* (2023).
 - [30] E. Kıcıman, R. Ness, A. Sharma, C. Tan, Causal reasoning and large language models: Opening a new frontier for causality, arXiv preprint arXiv:2305.00050 (2023).
 - [31] L. Yang, O. Clivio, V. Shirvaikar, F. Falck, A critical review of causal inference benchmarks for large language models, in: *AAAI 2024 Workshop on “Are Large Language Models Simply Causal Parrots?”*, 2023.
 - [32] B. Yu, Y. Li, J. Wang, Detecting causal language use in science findings, 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)*, 2019, pp. 4664–4674.
 - [33] P. Sumner, S. Vivian-Griffiths, J. Boivin, A. Williams, C. A. Venetis, A. Davies, J. Ogden, L. Whelan, B. Hughes, B. Dalton, et al., The association between exaggeration in health related science news and academic press releases: retrospective observational study, *BMJ* 349 (2014).

- [34] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
- [35] T. Schick, H. Schütze, It's not just size that matters: Small language models are also few-shot learners, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2021, pp. 2339–2352.