Can ChatGPT Understand Causal Language in Science Claims?

Yuheun Kim[1] **Lu Guo**[1] **Bei Yu**[1] **Yingya Li**[2]

[1]School of Information Studies, Syracuse University
{ykim72, lguo15, byu}@syr.edu

[2]Harvard Medical School and Boston Children's Hospital
yingya.li@childrens.harvard.edu

Abstract

This study evaluated ChatGPT's ability to understand causal language in science papers and news by testing its accuracy in a task of labeling the strength of a claim as causal, conditional causal, correlational, or no relationship. The results show that ChatGPT is still behind the existing fine-tuned BERT models by a large margin. ChatGPT also had difficulty understanding conditional causal claims mitigated by hedges. However, its weakness may be utilized to improve the clarity of human annotation guideline. Chain-of-thought prompting was faithful and helpful for improving prompt performance, but finding the optimal prompt is difficult with inconsistent results and the lack of effective method to establish cause-effect between prompts and outcomes, suggesting caution when generalizing prompt engineering results across tasks or models.

1 Introduction

Finding causal relationship is an important goal in scientific research. However, choosing appropriate causal language that accurately reflects the strength of evidence is a non-trivial task when describing research findings. Subjectivity and bias may affect how authors interpret the results. For example, some researchers argued that observational studies can not illuminate causal claims and thus causal language should not be used (e.g., Cofield et al., 2010), while others called for more confidence in causal inference with improved methods and guidelines, (e.g., Pearl and Mackenzie, 2018). On the other hand, average human readers reported difficulty in judging the strength of causal claims mitigated with hedges such as "may" or ambiguous terms like "linked to" (Adams et al., 2017). Manual fact-checking of causal claims in academic publications, news and social media posts also demonstrated evidence of prevalent exaggeration when reporting causal findings (Cofield et al., 2010; Sumner et al., 2014; Haber et al., 2018).

Prior studies have also looked into computational approaches for identifying claim strengths and exaggerated claims. The core component is a text classification task that categorizes research findings by their strengths. The original task definition was based on a manual content analysis (Sumner et al., 2014), which defined seven certainty levels: no statement, explicit statement of no relation, correlational (e.g. "drinking wine is associated with increased cancer rates"), ambiguous (e.g. "drinking wine linked to cancer risk"), conditional causal (e.g. "drinking wine might increase cancer risk"), can cause (e.g, "drinking wine can increase cancer risk"), and unconditionally causal (e.g. "drinking wine increases cancer risk"). However, Adams et al. (2017) found that average human readers can distinguish three categories of relationship only: direct cause statements (e.g. "makes"), can cause statements (e.g. "can make"); and moderate cause statements (e.g. "might cause", "linked", "associated with"), and they encountered difficulty in distinguishing the conditional causal statements and correlational statements in the last group.

In light of these observations, the subsequent computational modeling studies simplified the task to classify four categories: direct causal, conditional causal, correlational, or no relationship (Yu et al., 2019; Tan et al., 2021; Yu et al., 2020; Wright and Augenstein, 2021). These specialized models used various techniques to achieve high accuracy, such as fine-tuning pre-trained BERT models (Yu et al., 2019), or through causal augmentation (Tan et al., 2021). These models also have limitations, such as mistaking a no-relationship sentence as causal or correlational when confounding cues exist. They also rely on thousands of human-annotated training examples.

Recently, the debut of large language models (LLMs) such as ChatGPT shifted the NLP research paradigm toward the direction of "pre-train, prompt, and predict", where downstream tasks are

Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 379–389
July 14, 2023 ©2023 Association for Computational Linguistics

reformulated into textual prompts on zero-shot or few-shot settings (Liu et al., 2023). LLMs trained on sufficiently large and diverse datasets demonstrate promising performance on reasoning tasks without additional task specific training (Radford et al., 2019; Brown et al., 2020). The promising results inspired hope for eliminating the need for specialized models and expensive human annotations (Gilardi et al., 2023). A question rises then - can ChatGPT "understand" causal language in science writing? More specifically, can ChatGPT label the strength of causal claims correctly? Furthermore, since ChatGPT was trained with a variety of textual data, did it inherit the confusion that human readers have regarding conditional causal claims?

In this study, we evaluate ChatGPT's ability to understand causal claims in science papers and news. We tested GPT3.5 (text-davinci-003) and ChatGPT (gpt-3.5-turbo) performance in classifying causal claim strength using the annotated corpora provided by Yu et al. (2019) and Yu et al. (2020). Specifically, we ask the following research questions:

- RQ1: Does ChatGPT outperform fine-tuned BERT models for classifying causal claim strength?
- RQ2: How does ChatGPT interpret conditional causal claims?
- RQ3: Do GPT3.5 and ChatGPT agree on their predictions? Does ChatGPT give similar answers to semantically-similar prompts?
- RQ4: How do instructional elements in prompts, such as Chain-of-Thought, context, and system messages, affect ChatGPT performance?

Our paper is organized as follows: Section 2 summarizes studies on prompt engineering and various classification tasks evaluated on ChatGPT. In the Methods section, we introduce the different prompt designs we experimented (section 3.1), explain how we evaluated the model's performance (section 3.2), provide a summary of the data we used (section 3.3) and present the API details for the experiment (section 3.4). We then report experiment results that address RQ1 and RQ2 (section 4.1) as well as RQ3 and RQ4 (section 4.2). Based on the experiment we test the entire dataset (section 4.3) and also evaluate the result of applying ensemble models (section 4.4). Finally, we discuss our limitations and present our conclusion with discussions in section 5.

2 Related Work

Since prompts provide crucial information for LLMs such as ChatGPT, a number of studies have explored prompt engineering strategies (Liu et al., 2023). Here we summarize several common prompt design approaches with a focus on text classification tasks, which are most relevant to our study.

Zero-shot, one-shot, and few-shot learning are three types of prompting designs used to guide LLMs. Zero-shot prompting provides task descriptions or instructions without explicit examples. One-shot prompting uses a single example for the desired task. Few-shot prompting is similar to the one-shot design, but it involves providing the model with a small number of examples instead of just one for the model to learn from and generate task-aligned responses (Brown et al., 2020).

Prompts can be generated either manually or automatically (Brown et al., 2020; Radford et al., 2019; Petroni et al., 2019). While manual, intuitive approach is straightforward, it can be time-consuming to identify the most effective prompt and there is no guarantee to find one (Jiang et al., 2020). Researchers then sought automatic approaches (Liu et al., 2023; Gao et al., 2021; Raffel et al., 2020), or even asking ChatGPT itself to generate prompts (Zhong et al., 2023). However, since LLMs sometimes do not follow instructions, their answers may be ill-formatted or even invalid. When that happens, human intervention is needed, which increases the time cost for post-processing LLM results (Kocoń et al., 2023). Therefore, in this study, we focused on manually-generated prompts.

Text classification tasks often use instruction prompts to explicitly tell LLMs what to do. For instance, Qin et al. (2023)'s prompt starts with an instruction of task description: *"For each snippet of text, label the sentiment of the text as positive or negative. The answer should be exact 'positive' or 'negative' "*, followed with the text to be labeled. Ye et al. (2023) formulate their prompt as *"Definition: ... Input: ... Answer: ..."* where an example of definition can be *determine the speaker of the dialogue, agent or customer*. More context information about the task may be added to the instruction prompt, such as providing the definition of genre for genre classification (Kocoń et al., 2023).

Another commonly-used instructional element is the Chain-of-Thought (CoT), which has been found to improve LLMs' performance on certain arithmetic, commonsense, and symbolic reasoning tasks (Wei et al., 2022). While CoT was initially an instance of few-shot prompts, a decent zero-shot performance in reasoning tasks and classification tasks was demonstrated by adding a simple CoT prompt such as *"Let's think step by step"* at the end of a question (Kojima et al., 2022; Zhong et al., 2023).

ChatGPT also provides a unique feature, system messages, which can be used as part of the prompt to guide the model's behavior (Shen et al., 2023), such as *"You are a helpful assistant that can classify sentences as either causal or correlational research findings"* which specifies the model to behave as a professional for our task.

Previous studies have evaluated ChatGPT on various classification tasks (Qin et al., 2023; Bang et al., 2023; Huang et al., 2023; Kocoń et al., 2023), using various prompt designs. The results indicate promise and limitations. In the case of sentiment classification, ChatGPT was found to have difficulty in understanding neutral sentiment, or give unbalanced predictions on negative vs. positive sentiment, raising questions on the extent to which ChatGPT really "understands" sentiment as a linguistic concept (Wang et al., 2023). It is also difficult to directly compare the results due to different sample sets and prompts. The sample sizes were usually small since most studies were conducted before OpenAI made the API available.

3 Methods

3.1 Prompt Design

We experimented with intuitive trial-and-error approaches as well as consulting prior studies on the prompt designs that have demonstrated good performance in other text classification tasks. In this study we focused on zero-shot prompting design for two reasons. First, it is the most common strategy that end users choose to interact with ChatGPT. Second, since ChatGPT likely captures the latent social information (Horton, 2023), we are curious how ChatGPT "interpret" causal language without seeing training examples annotated by domain experts. All prompt designs that we have evaluated are documented in Table 1.

Our process started with a number of carefully crafted, intuitive prompts that include specific instructions. We then selected the best performing prompt as the baseline (BASE) for further comparison with other manually-constructed prompts from four previous studies with minor modifications to suit our task (Huang et al., 2023; Kocoń et al., 2023; Kuzman et al., 2023; Qin et al., 2023).

After that, we augmented the best performing prompt with two additional instructional elements, context of the task and CoT prompts (Reynolds and McDonell, 2021). For the context we include explanations and cue words of causal, correlational, and no relationship from Yu et al. (2020). We examined whether the location of context affects the performance by adding the context before and then after the BASE prompt. We added CoT to the end of the prompt, a usual design, by appending the phrase *"Answer (causal, correlational or no relationship) the question step by step"*, which was inspired from Zhong et al. (2023).

We also conducted additional tests to evaluate whether setting system message affects ChatGPT performance.

3.2 Evaluation Method

A semi-automatic approach was taken to post-process ChatGPT answers, since ChatGPT sometimes does not provide answers in the requested format or even provides invalid answers. We used a set of heuristic rules to map ChatGPT and GPT3.5 answers to the category labels. For instance, if "correlational" is in the answer, but not "causal", the label would be "correlational". See post-processing code in Appendix A Listing 1. Ambiguous answers that cannot be automatically mapped were manually examined and mapped. The number of invalid answers (# of unlabeled) was documented for each experiment. For prompts with CoT, the results were manually examined to verify whether the reasoning is valid.

After the post-processing, the macro f1-score is calculated to measure each model's performance, such as ChatGPT-BASE, against the human-annotated labels. Cohen's Kappa (Cohen, 1960) was also used to evaluate the agreements between different models and prompts. We conjecture that if a prompt shows consistently good performance across GPT3.5 and ChatGPT, the prompt has more robustness than other prompts that perform well on only one of them.

	Prompt
BASE	Read the following sentence - _____ Answer this question as concisely as possible - Does the sentence describe any causal or correlational research finding?
BASE+“conditional”	Read the following sentence - _____ Answer this question as concisely as possible - Does the sentence describe any direct causal, **conditional causal**, or correlational research finding?
BASE+“possible”	Read the following sentence - _____ Answer this question as concisely as possible - Does the sentence describe any direct causal, **possible direct causal**, or correlational research finding?
Huang et al. (2023)	Given Sentence: ‘_____’. Answer causal or correlational if the sentence describes any research finding. Answer as concisely as possible.
Kocoń et al. (2023)	Which of the attributes: “causal”, “correlational”, “no relationship” describe the research finding of a given text? Write your answer in the form of a Python list containing the appropriate attributes. Text: _____
Kuzman et al. (2023)	Please classify the following text describing a research finding and explain your decision. You can choose from the following classes: Causal, Correlational, No Relationship. The text to classify: _____
Qin et al. (2023)	For each snippet of text, label the research finding of the text as causal or correlational or no relationship. The answer should be exact ‘causal’ or ‘correlational’ or ‘no relationship’. Text: _____ Label:
element: **CoT**	Answer (causal, correlational or no relationship) the question step by step.
element: **Context**	Correlational: The statement describes the association between variables, but causation cannot be explicitly stated. Language Cue: association, associated with, predictor, at high risk of... Causal: The statement says that the independent variable directly alters the dependent variable. Language Cue: increase, decrease, lead to, effective in, contribute to, reduce, can... No relationship: The statement is not for current study findings or no correlation/causation relationship is mentioned in the statement.
system message	`You are a helpful assistant that can classify sentences as either causal or correlational research findings.`

Table 1: Different prompt designs

3.3 Data

We utilized two open-access cross-genre datasets that were manually annotated for science claim strength. The first dataset includes a sample of 3,061 research conclusion sentences from structured abstracts in PubMed articles (Yu et al., 2019). The second dataset consists of 2,076 sentences from health-related press releases on EurekAlert!, a major science press release platform (Yu et al., 2020). These sentences were either headlines or the first two sentences in press releases. Both datasets were manually annotated with the same four-category labels including correlational, direct causal, conditional causal, and no relationship.

To compare the effectiveness of different prompt designs, we created a sample subset from the PubMed dataset as the development set. In a prior study, Gutiérrez et al. (2022) sample 100 examples for prompt design selection. To ensure an equal representation of each class, we sampled 50 sentences from each class with a total of 200 sentences. The main reason for choosing a relatively small development set is the time cost for post-processing the ambiguous answers. After these experiments, we selected the best prompt design and evaluated it on the entire PubMed and EurekAlert! datasets. Since we are particularly interested in ChatGPT's understanding of conditional causal claims, we conducted two sets of experiments, one with conditional causal category and one without.

3.4 API

OpenAI released a public API for both GPT3.5 and ChatGPT model. For GPT3.5 experiments we use "text-davinci-003" model with temperature set as 0 and max tokens set as 50. The temperature is set to 0 as Gilardi et al. (2023) found that lower temperatures result in more consistent outcomes, ideal for annotation tasks.

For ChatGPT experiments we use "gpt-3.5-turbo" model. We input our prompt designs in the user message as `{"role": "user", "content": prompt}`. When testing the efficacy of system message, we add in the system message `{"role": "system", "content": system message}` prior to the user message.

4 Results

4.1 First Set of Experiments: RQ1 and RQ2

The first set of experiments included conditional causal examples. We started with a prompt based on the human annotation instruction (see "BASE+conditional" in Table 1). The result in Table 2 shows the macro f1-score at .486, much lower than the .881 macro f1-score from a fine-tuned BioBERT model in Yu et al. (2019). Among the four categories, ChatGPT severely underperformed in the conditional causal category with a low .164 macro f1, which prompted us for further investigation.

	No relationship	Direct causal	Conditional causal	Correlational	F1-score	Macro f1-score
No relationship	**29**	1	2	18	0.674	0.486
Direct causal	2	**19**	0	29	0.537	
Conditional causal	4	2	**6**	38	0.164	
Correlational	1	1	0	**48**	0.570	

Table 2: ChatGPT initial confusion matrix: row stands for predicted label and column stands for actual ground truth label

To understand more about how ChatGPT interprets the concept of "conditional causal", we asked ChatGPT *"What is 'conditional causal relationship'?"*. It responded "causal under certain conditions" (see full response in appendix A). We further examined its interpretation by adding CoT - *"Answer (direct causal, conditional causal, correlational, or no relationship) the question step by step."* to the prompt. Again, the response was "causal under certain conditions". These results suggest that the category label "conditional causal" is a misnomer, at least to ChatGPT.

We then attempted to look for an alternative label that would align better with ChatGPT's interpretation. A re-examination of the CoT responses showed that all answers used the given labels "direct causal" or "conditional causal", except for three answers, in which ChatGPT used the terms "possible direct causal", "potential direct causal", and "potential causal". ChatGPT's answer to the question *"What is 'possible direct causal relationship'?"* also showed a better match with the original definition of "conditional causal" (see full answer in appendix A).

We then hypothesized that "possible" or "potential" may be a better term than "conditional" for ChatGPT. We replaced "conditional" with "possible" in the prompt and repeated the evaluation (see "BASE+possible" in Table 1). The results in Table 3 show that the new prompt drastically improved ChatGPT's performance: the f1-score for conditional causal increased from .164 to .578; the f1-scores for the other three categories were also improved slightly; the macro f1 increased from .486 to .631. However, ChatGPT's performance, even with misnomer corrected in the instruction, still falls behind the fine-tuned BioBERT model (RQ1).

	No relationship	Direct causal	Possible direct causal	Correlational	F1-score	Macro f1-score
No relationship	**27**	7	8	8	0.675 ↑	
Direct causal	0	**29**	19	2	0.624 ↑	0.631 ↑
Possible direct causal	1	5	**37**	7	0.578 ↑	
Correlational	2	2	14	**32**	0.646 ↑	

Table 3: ChatGPT confusion matrix (revise conditional causal to possible direct causal): row stands for predicted label and column stands for actual ground truth label

To further probe how the one-word switch in the prompt affected ChatGPT's interpretation of conditional causal relationship, we looked into its interpretation of hedges.

Conditional causal relationships are usually expressed by hedges. Actually, the 50 conditional causal examples were covered by six hedge words: "may" 32 times, "appear" 6, "could" 6, "might" 4, "seem" 3, and "unlikely" once. Note the total is 52 since two sentences used two hedge words.

As an example, we examined the six sentences that used "appear(s/ed)" as conditional causal cues. With the "BASE+conditional" prompt, ChatGPT recognized half of them as correlational and the other half as direct causal. With the "BASE+possible" prompt, ChatGPT recognized four as "possible direct causal" and two as "direct causal". It is an improvement, but still not perfect.

Overall, our results provide evidence that ChatGPT has difficulty interpreting hedges in conditional causal claims, even after the prompt instruction was adjusted to match its own interpretation of this concept. This indicates that ChatGPT may have inherited the confusion or bias among human readers regarding conditional causal claims (RQ2).

Note that the misnomer has not been a problem for human annotators since they can adapt their interpretation based on the given definition on "conditional causal", which was semantically equivalent to possible/speculative/qualified causal (Sumner et al., 2014). It is not a problem for fine-tuned BERT models either, since the models learned the concept from training data instead of the category definitions. Despite that, ChatGPT's lack of adaptability may be utilized to design or refine human annotation guidelines to reduce potential misnomers.

4.2 Second Set of Experiments: RQ3 and RQ4

	GPT3.5 Macro f1-score	ChatGPT Macro f1-score	Cohen Kappa
BASE	0.494(7)	**0.743**(3)	0.491
Huang et al. (2023)	0.330	0.504	0.147
Kocoń et al. (2023)	0.514(6)	0.545(6)	0.478
Kuzman et al. (2023)	0.558	0.629	0.530
Qin et al. (2023)	**0.699**	0.735	0.665
BASE + CoT	0.538	**0.772**	0.462
context + BASE	0.695	0.744	0.618
context + BASE + CoT	**0.709**	0.684	0.504
BASE + context	-	0.763	-
BASE + context + CoT	0.364	0.419(1)	0.029
system message + (BASE + CoT)	-	0.726	-

Table 4: Prompt results on a sample of 150 PubMed data. The Cohen's Kappa score is calculated between GPT3.5 and ChatGPT labels. The numbers in parenthesis are unlabeled examples due to invalid answers, such as *"'causal or correlational research finding' neither is mentioned in the sentence"*.

In the remaining experiments we excluded conditional causal examples, shifting focus on ChatGPT's ability in distinguishing direct causal, correlational, or no relationship.

We first compared the performance scores for different prompt designs listed in Table 1. The results are reported in Table 4. Note that the unlabeled examples were excluded when calculating f1. For example, for the BASE prompt, a total of 10 unlabeled examples, 7 from GPT3.5 and 3 from ChatGPT, were excluded, so that the two f1-scores are comparable. When comparing results across prompts, since only three prompts had unlabeled examples, including two with performance at the lower end, the result comparisons below were minimally affected, except that the .743 macro f1 for ChatGPT-BASE should be interpreted with caution.

The first group of results (in rows 2-6) are GPT3.5 and ChatGPT performance with our own BASE prompt and four other prompts inspired from prior studies. The prompt from Qin et al. (2023) performed best with GPT3.5 with .699 macro f1-score, and our own BASE prompt performed best with ChatGPT at .743. We also calculated Cohen's Kappa between GPT3.5 and ChatGPT results with the same prompts, and found that the agreements varied vastly from .147 for the prompt from Huang et al. (2023) to .665 from Qin et al. (2023).

Although our BASE prompt achieved the highest macro f1 .743 among the five prompts across GPT3.5 and ChatGPT, the prompt from Qin et al. (2023) shows consistently high performance across GPT3.5 and ChatGPT (.699 and .735 macro f1-scores) and highest inter-model agreement (.665 Cohen's Kappa), demonstrating strong robustness. In comparison, the BASE prompt used the format of a question, while the prompt from Qin et al. (2023) was formatted as a labeling task with stricter formatting instructions. Further studies are needed to examine what design features contributed to the performance differences (RQ3).

We then tested the impact of additional instructional elements, i.e. context and CoT (see results in rows 7-9). We observed a slight improvement in performance when separately incorporating CoT and context to the BASE prompt, resulting in a macro f1-score of .772 and .744 respectively.

Note that adding CoT to the prompt does not guarantee an answer with a reasoning process. We found that only 42% answers to the BASE + CoT prompt included the reasoning process. For the context + BASE + CoT prompt, the response rate increased to 85%. However, a higher CoT response rate did not translate to better performance. Instead, the macro f1-score decreased from .772 to .684.

Our finding that CoT improved ChatGPT performance on a zero-shot setting is consistent with prior literature (Kojima et al., 2022). However, it is worth noting that changing wording in CoT can also impact the results. In our experiments on the development set, we tested two variations of CoT. The first prompt was "BASE + Let's think step by step." This yielded a macro f1 score of .732. The second prompt was "BASE + Answer (causal, correlational or no relationship) the question step by step.", referred to as BASE+CoT in Table 4, which achieved a higher macro f1 score of .772.

We also examined whether the interpretations in CoT responses were faithful, which means ChatGPT's interpretation is consistent with its answer (Jacovi and Goldberg, 2020). After checking all responses in the ChatGPT-Base + CoT experiment, we found that all CoT interpretations were faithful. In other experiments, unfaithful interpretations were occasionally spotted but rare. For example, one answer included a 4-step reasoning process. Step 2 implied that the sentence has no relationship: *step 2: does the sentence describe the research finding as causal or correlational? no*. However, Step 4 changed the final answer to correlational: *step 4: therefore, the answer is "correlational"*.

We also tested if adding context before or after prompt would make any difference. The second and the third group of results in Table 4 show that it did not affect ChatGPT significantly, which performed slightly better with context after prompt (.763 macro f1 for BASE + context vs. .744 for context + BASE). However, stark contrast was observed with GPT3.5, which had a decent performance at .695 macro f1 for context + BASE; however, it failed to output any valid response when context was added after the BASE prompt, indicating that the context after the prompt distracted GPT3.5 away from completing the task. When further adding CoT after the context (i.e. BASE + context + CoT), GPT3.5 performance was still poor at .364 macro f1. Surprisingly, the BASE + context + CoT prompt also dragged ChatGPT performance down to .419. These results suggest that prompt design with additional instructional elements is not always "the more the merrier". The inconsistent performance between GPT3.5 and ChatGPT also indicates the uncertainty when experimenting with prompt engineering across LLMs.

Our last prompt engineering attempt was to add a system message to the best prompt for ChatGPT-BASE + CoT. It did not help as the performance was slightly decreased from .772 to .726.

In summary, context + BASE + CoT resulted in the best GPT3.5 performance prompt at .709 macro f1, and BASE + CoT resulted in the best ChatGPT performance at .772. For ChatGPT, adding CoT helped, but adding both context and CoT hurt. For GPT3.5, adding CoT helped, as well as adding context before prompt, but adding context after prompt distracted it (RQ4).

4.3 ChatGPT Results on Full Data Sets

After finding the best performing prompt on the development set, we applied it to the entire PubMed dataset and EurekAlert! dataset, still excluding the conditional causal examples. The distribution of sentences per label is shown in Table 5. Since ChatGPT consistently outperformed GPT3.5 in previous experiments, we proceeded to test with ChatGPT only. We also repeated the test once a day for three days (April 21-23, 2023) to check the consistency of results among different runs, since ChatGPT cannot guarantee result reproducibility.

	PubMed	**EurekAlert**
No relationship	1,353	486
Causal	494	568
Correlational	995	738
Total	2,842	1,972

Table 5: Dataset description for each label

Table 6 shows that the macro f1-scores for both datasets decreased from the best performance (.772) on the development set to the range of .695 to .698 for the PubMed dataset and the range of .628 to .638 for EurekAlert!. Overall the unlabeled examples are not a major issue with its ratios all below 0.5%. However, the results among the three runs disagreed to some extent, as measured by average Kappa values at .813 and .701 respectively, raising concerns for result reproducibility if used as an off-the-shelf text classification model.

	PubMed			**Eureka**		
	Macro f1-score	**# of unlabeled**	avg_k	**Macro f1-score**	**# of unlabeled**	avg_k
1st	0.698	5	0.813	0.628	2	0.701
2nd	0.695	14		0.638	6	
3rd	0.695	14		0.634	6	

Table 6: ChatGPT performance on entire PubMed dataset and EurekAlert! dataset for 3 days. The avg_k represents the average Cohen Kappa value.

4.4 Performance of Ensemble Models

Despite the promising performance of GPT3.5 and ChatGPT with various prompts, they are still relatively weak models with macro f1-scores below 0.8. The correlations among these models were also in low to mid range, as measured by Cohen's Kappa. For example, the Kappa values between GPT3.5 and ChatGPT range from .147 to .665 in Table 4. The Kappa values among the ChatGPT results with the five different manual prompts range from .199 to .656. These observations suggest the possibility of constructing an ensemble model through simple majority vote (Dietterich, 2000). Therefore we tried two ensemble models (1) combining five ChatGPT models with the five manual prompts, and (2) combining ten models, five from GPT3.5 and five from ChatGPT.

We used a straightforward majority voting approach to ensemble each model's outcomes. In case of a tie, we used a weighted voting approach that takes the macro f1-score of each model as the weight, favoring the better-performing models.

The result in Table 7 shows that the ensemble of five ChatGPT models with weighted tie-breaking resulted in .743 macro f1, which did not beat the .772 best performance with BASE + CoT. The ensemble of both GPT3.5 and ChatGPT models performed even worse, at .705 macro f1. In summary, the simple majority vote ensemble did not lead to a better-performing model.

	ChatGPT	**GPT3.5+ChatGPT**
No relationship	0.788	0.724
Causal	0.691	0.684
Correlational	0.748	0.707
Macro f1-score	0.743	0.705
# of ties	5	6

Table 7: Ensemble results. ChatGPT refers to an ensemble of ChatGPT models on five manual prompts and GPT3.5+ChatGPT refers to an ensemble of both GPT3.5 and ChatGPT.

5 Conclusion and Discussion

Causal language is an important rhetorical device in science communication. However, subjectivity in causal language use and understanding is a challenge for science writing and reading. Since ChatGPT captures latent social information to some extent, this study evaluated its ability to understand causal language in science papers and news by testing their accuracy in a task of claim strength classification. The results show that (1) ChatGPT is still behind the existing fine-tuned BERT models by a large margin; (2) ChatGPT seems to have inherited the confusion observed among average human readers when judging the strength of conditional causal claims that were mitigated by hedges; (3) ChatGPT performance varied substantially with semantically-similar prompts and across different model versions; (4) CoT responses were faithful and helpful. ChatGPT was able to reproduce its results at the level of 0.7-0.8 measured by Cohen's Kappa. However, the inconsistency in performance across model versions and semantically-similar prompts suggests caution when generalizing prompt engineering results across tasks and models.

While we were conducting our experiment, another study posted to arxiv (Chen et al., 2023) reported their ChatGPT evaluation on the PubMed data set. Both studies shared the findings that prompt engineering required significant investment and a slight difference in prompts could lead to substantial change in performance. Both studies on the same task found that CoT helped performance. While both zero-shot and few-shot settings were tested in Chen et al. (2023), we tested the zero-shot setting only. Our study has better performance under the zero-shot setting. Comparing the prompts, we hypothesize that explicitly asking the causal relationship in the prompt may have helped. However, a systematic method is still lacking to infer causality between word choices in prompts and the performance.

Despite our effort for a systematic review of ChatGPT's understanding of causal claims, our study design has some limitations as the evaluation methodology for prompt engineering is still under development in the NLP community. We arbitrarily decided on the size of the development set. Our study focused on zero-shot setting with the purpose of evaluating the latent understanding on causal claims within ChatGPT. Further exploration could be conducted to investigate the impact of few-shot settings by carefully selecting examples based on recent progress in few-shots prompting methods (Lu et al., 2022; Liu et al., 2022).

We conclude that ChatGPT has a promising but still limited ability in understanding causal language in science writing. CoTs improved prompt performance, but finding the optimal prompt is difficult with inconsistent results and the lack of effective methods to establish cause-effect between prompts and outcomes. Following instruction is an important prerequisite for using ChatGPT as a text classification tool, to avoid high labor cost for post-processing its answers. However, ChatGPT provides a new, simulation-style approach for designing and evaluating human annotation guidelines.

Acknowledgements

This research is supported by the US National Science Foundation under grant 1952353.

References

Rachel C Adams, Petroc Sumner, Solveiga Vivian-Griffiths, Amy Barrington, Andrew Williams, Jacky Boivin, Christopher D Chambers, and Lewis Bott. 2017. How readers understand causal and correlational expressions used in news headlines. *Journal of experimental psychology: applied*, 23(1):1.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. *arXiv preprint arXiv:2302.04023*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Shan Chen, Yingya Li, Sheng Lu, Hoang Van, Hugo JWL Aerts, Guergana K Savova, and Danielle S Bitterman. 2023. Evaluation of chatgpt family of models for biomedical reasoning and classification. *arXiv preprint arXiv:2304.02496*.

Stacey S Cofield, Rachel V Corona, and David B Allison. 2010. Use of causal language in observational studies of obesity and nutrition. *Obesity facts*, 3(6):353–356.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Thomas G Dietterich. 2000. Ensemble methods in machine learning. In *Multiple Classifier Systems: First International Workshop, MCS 2000 Cagliari, Italy, June 21–23, 2000 Proceedings 1*, pages 1–15. Springer.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*.

Bernal Jiménez Gutiérrez, Nikolas McNeal, Clay Washington, You Chen, Lang Li, Huan Sun, and Yu Su. 2022. Thinking about gpt-3 in-context learning for biomedical ie? think again. *arXiv preprint arXiv:2203.08410*.

Noah Haber, Emily R Smith, Ellen Moscoe, Kathryn Andrews, Robin Audy, Winnie Bell, Alana T Brennan, Alexander Breskin, Jeremy C Kane, Mahesh Karra, et al. 2018. Causal language and strength of inference in academic and media articles shared in social media (claims): A systematic review. *PloS one*, 13(5):e0196346.

John J Horton. 2023. Large language models as simulated economic agents: What can we learn from homo silicus? *arXiv preprint arXiv:2301.07543*.

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. *arXiv preprint arXiv:2302.07736*.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? *arXiv preprint arXiv:2004.03685*.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. 2020. How can we know what language models know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Jan Kocoń, Igor Cichecki, Oliwier Kaszyca, Mateusz Kochanek, Dominika Szydło, Joanna Baran, Julita Bielaniewicz, Marcin Gruza, Arkadiusz Janz, Kamil Kanclerz, et al. 2023. Chatgpt: Jack of all trades, master of none. *arXiv preprint arXiv:2302.10724*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*.

Taja Kuzman, Igor Mozetic, and Nikola Ljubešic. 2023. Chatgpt: Beginning of an end of manual linguistic data annotation? use case of automatic genre identification. *ArXiv, abs/2303.03953*.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pretrain, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Judea Pearl and Dana Mackenzie. 2018. *The book of why: the new science of cause and effect*. Basic books.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? 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)*, pages 2463–2473.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7.

Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. 2023. In chatgpt we trust? measuring and characterizing the reliability of chatgpt. *arXiv preprint arXiv:2304.08979*.

Petroc Sumner, Solveiga Vivian-Griffiths, Jacky Boivin, Andy Williams, Christos A Venetis, Aimée Davies, Jack Ogden, Leanne Whelan, Bethan Hughes, Bethan Dalton, et al. 2014. The association between exaggeration in health related science news and academic press releases: retrospective observational study. *Bmj*, 349.

Fiona Anting Tan, Devamanyu Hazarika, See Kiong Ng, Soujanya Poria, and Roger Zimmermann. 2021. Causal augmentation for causal sentence classification. In *Proceedings of the First Workshop on Causal Inference and NLP*, pages 1–20.

Zengzhi Wang, Qiming Xie, Zixiang Ding, Yi Feng, and Rui Xia. 2023. Is chatgpt a good sentiment analyzer? a preliminary study. *arXiv preprint arXiv:2304.04339*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Dustin Wright and Isabelle Augenstein. 2021. Semi-supervised exaggeration detection of health science press releases. *arXiv preprint arXiv:2108.13493*.

Seonghyeon Ye, Hyeonbin Hwang, Sohee Yang, Hyeongu Yun, Yireun Kim, and Minjoon Seo. 2023. In-context instruction learning. *arXiv e-prints*, pages arXiv–2302.

Bei Yu, Yingya Li, and Jun Wang. 2019. 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)*, pages 4664–4674.

Bei Yu, Jun Wang, Lu Guo, and Yingya Li. 2020. Measuring correlation-to-causation exaggeration in press releases. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4860–4872.

Qihuang Zhong, Liang Ding, Juhua Liu, Bo Du, and Dacheng Tao. 2023. Can chatgpt understand too? a comparative study on chatgpt and fine-tuned bert. *arXiv preprint arXiv:2302.10198*.

A Appendix

ChatGPT’s response to the question “What is ‘conditional causal relationship’?”

A conditional causal relationship is a type of relationship between two variables, where one variable (the cause) has a causal effect on the other variable (the effect) under certain conditions. In other words, the causal effect of the cause variable is dependent on the presence or absence of a third variable, known as the "conditioning variable."

For example, let’s say that we are interested in studying the effect of a new medication (the cause) on blood pressure (the effect), but we know that the effect of the medication on blood pressure may depend on other factors such as age or body weight (the conditioning variables). In this case, we might say that there is a conditional causal relationship between the medication and blood pressure, where the causal effect of the medication on blood pressure is conditional on the values of the conditioning variables.

In summary, a conditional causal relationship is a relationship between two variables where the causal effect of one variable on the other is conditional on the presence or absence of a third variable.

ChatGPT’s response to the question “What is ‘potential causal relationship’?”

A potential causal relationship refers to the possibility that one variable or event may be responsible for causing another variable or event. It suggests that there may be a cause-and-effect relationship between two phenomena, but it does not necessarily prove that one caused the other.

Establishing a potential causal relationship typically involves observing an association between two variables and then conducting further research or analysis to determine whether there is a causal link between them. This can be challenging because there may be other factors that are responsible for the observed association, or there may be reverse causation (i.e., the effect is actually causing the cause).

In order to establish a true causal relationship, it is often necessary to use rigorous scientific methods such as randomized controlled trials or longitudinal studies, which can help to rule out alternative explanations and establish a causal relationship with greater certainty.

Listing 1: Post-processing code

```
def chatgpt_label(answer):
   if ’no relationship’ in answer:
      return ’norel’
   if ’correlational’ in answer and
       ’causal’ not in answer:
      return ’correlational’
   elif ’causal’ in answer and
       ’correlational’ not in answer:
      return ’causal’
   elif ’causal’ not in answer and
       ’correlational’ not in answer:
      if answer.startswith(’no’) or
          answer.startswith(’unclear’):
         return ’norel’
      return ’???’
   else:
      return ’???’
```