

Automated classification of causal relations. Evaluating different LLM performances.

Giacomo Magnifico

University of Tartu

Institute of Computer Science

Estonia

giacomo.magnifico@ut.ee

Abstract

The search for formal causal relations in natural language faces inherent limitations due to the lack of mathematically and logically informed datasets. Thus, the exploration of causal relations in natural language leads to the analysis of formal-logic-adjacent language patterns. Thanks to the recent advancements of generative LLMs, this research niche is expanding within the field of natural language processing and evaluation. In this work, we conduct an evaluation of 9 models produced by different AI developing companies in order to answer the question “Are LLMs capable of discerning between different types of causal relations?”. The SciExpl dataset is chosen as a natural language corpus, and we develop three different prompt types aligned with zero-shot, few-shot, and chain-of-thought standards to evaluate the performance of the LLMs. Claude 3.7 Sonnet and Gemini 2.5 Flash Preview emerge as the best models for the task, with the respective highest F1 scores of 0.842 (few-shot prompting) and 0.846 (chain-of-thought prompting).

1 Introduction

Causality is a subject deeply related to human perception and nature, and for such reason it is inevitably tied to human bias and variable logical framework (Matute et al., 2015; Henne et al., 2021). The difficulty of the defining task of causality extractions comes from the difficulty in formalising causality itself. Such an endeavour persists in the scientific community within various fields, from social sciences to computer science, from psychology to mathematics. Important frames of reference have been given in works such as Pearl (2009) and Icard et al. (2017), which elaborate on the nature of causal relations and their underlying framework. Numerous works have investigated the mechanisms underlying logical frameworks of causality in various communicative contexts, with

examples such as Henne et al. (2021) and Matute et al. (2015), and from text in natural language (Yang et al., 2022). Some contributions derive from the pioneering work of Lewis (1986) and further the examination of critical aspects of causality and cause-and-effect relationships specifically tied to natural language explanations; other contributions have focused on the interaction between causality and explainability, such as Jacovi et al. (2021) and Halpern and Pearl (2005). In recent years, the interactions between natural language and machine learning models have irreversibly changed with the advent of LLMs and generative models, such as the GPT, Llama, and Mistral lines of generative architectures. The paradigm has shifted to accommodate their existence, and research has begun to investigate the properties of LLMs and test their abilities (Ziyu et al., 2023), including their use for causal extraction within shorter and longer texts (Wang et al., 2024; Chatwal et al., 2025).

Our main research question is thus formulated: *When dealing with short, real-life examples, are LLMs capable of discerning between different types of causal bond within different topic domains?*

The paper moves forward from this research question by outlining relevant literature and evaluating properties of large language models, as reported in Section 2. Then, we present the choice of models produced by different AI leading companies in Section 3, along with the chosen SciExpl dataset of explanatory single-sentences (Magnifico and Barbu, 2025) in order to conduct the analysis. This is followed by a detailed analysis of the parameters selected as a standard for model calls in Section 4, with the details regarding prompt development and instructions being presented in the same section. Sections 5 and 6 close the paper with the results of the evaluation, an analysis of said results, and a summary of the work that has been done, including the limitations and weaknesses.

The code for the entire evaluation pipeline has been rewritten in `ipynb` notebook form, and it is made available to the community through a dedicated repository¹ along with model outputs (both raw and cleaned).

2 Related Work

As research in causal inference has attempted to formalise the relationship between events and causes in the real world, the naturally occurring bias and lack of unified interpretations make it an especially difficult task (Pearl, 2009; Matute et al., 2015; Henne et al., 2021). Multiple formal models and datasets that make use of the markers of logical relations between elements in natural language already exist; some datasets present knowledge pairs with out-of-context information used to test common-sense reasoning (Rein et al., 2024), others make use of underlying causal graphs derived from natural language (Romanou et al., 2023). Others yet focus on in-domain analysis of question-answering and context-driven causation extraction, with one of the most recent examples being Moreno Sandoval et al. (2025). As this paper presents work derived from the causal analysis of natural language data by generative models, the frame of reference for the nature of causal relations is closer in concept to the relativity of causality (Icard et al., 2017). The proximity to Icard’s work is in relation to the variability in causal strength and subjectivity, rather than the precise formality of such relations. This is further expressed in the analysis of the SciExpl dataset (Magnifico and Barbu, 2025) in the following section.

While the search for causal extraction methodologies is nothing new within the area of natural language processing (Khoo et al., 2000; Garcia, 1997), the advent of large language models has undoubtedly shaken the field. The LLM approach to causal inference and extraction in recent years has led to diverse results: while some works suggest that LLMs have only a mimicry of causal inference devoid of actual capabilities (Zečević et al., 2023), others suggest the possibility of causal inference (Kiciman et al., 2023) and the distinction between causation and correlation. It should be noted, however, that while such a distinction is possible and LLMs can benefit from it, LLMs struggle with distinctions between general causal patterns and

non-causal sentences (Jin et al., 2024). A generally agreed claim is that LLMs struggle with complex scenarios involving real-world events (Ashwani et al., 2024; Romanou et al., 2023), as finetuning happens with datasets aligned to specific causal directions in mind. Therefore, it is safe to claim that the use of specific causal datasets can lead to enhanced performance, especially in syntactically-similar tasks (Ashwani et al., 2024). Considering previous evaluation settings for LLMs and causal reasoning, such as Ziyu et al. (2023), this approach leverages the analysis of diverse causal sentences to focus on the capabilities of large architectures to distinguish between causal links of variable strength. Rather than a binary-link identification, the task changes to a classification of a spectrum of clearly identifiable different causal bonds, which can prove challenging for semantic and syntactic similarity. Compared to recent approaches that aim to extract cause-effect relations in a question-answering format (Chatwal et al., 2025), or that focus on document-level causal extraction (Wang et al., 2024), the aim of this work is pointed towards shorter inputs of more variable nature, with the same thorough analysis.

3 Materials

3.1 Data

The main drive for this work was to evaluate the efficiency in causal classification not only between causal and non-causal statements, but within the spectrum of causal bonds of different strength (Icard et al., 2017). As the intention was to evaluate the performance of generative models within the scope of natural language understanding with no forced question-answering format, the choice of dataset fell onto the SciExpl collection made available in Magnifico and Barbu (2025). The available dataset consists of 272 sentences in English, within topic domains within the area of biochemistry, annotated as different “explanation categories” by 120 annotators in total. Each sentence is classified with two labels according to Magnifico and Barbu (2025)’s different categorisation types, one for the explanation type (6 labels) and one for the causal link expressed in the sentence (3 labels). The latter is used for the purpose of this work, as the authors show that the inter-annotator agreement is more robust for that categorisation type (Krippendorff’s alpha value of 0.667) and the category balance is split at 40% - 40% - 20% between the labels. The

¹<https://github.com/gima9552/LLM-Causality-Classification>

following are the definitions for the labels used in the dataset, and an example sentence written ad hoc is provided.

- **Strong** causal links are expressed in sentences that present an explicit cause-effect relation, possibly detailing multiple intermediate steps in the causal process. The original explicit cause, as well as all the presented steps in the causal chain, ultimately lead to the presented effect. *“You are jittery because you drank too much coffee”*.
- **Weak** causal links are used to establish relationships between variables in the form of indirect or implicit causal relations, which may arise from mechanisms such as bias, intrinsic properties of entities, or hypothesised causal influence. While statistical correlations may sometimes point toward such bonds, they do not in themselves imply causality and should not be equated with causal relations. *“Oil paints are difficult to use, and many amateur artists prefer acrylics”*.
- **Contrastive** structures manifest in sentences that present multiple causal links, which could be both/either of the two aforementioned types. As multiple pairs of variables are presented as self-standing cause-effect relations, the different causal interactions are compared to one another to highlight how the differences in origin cause lead to differences in consequential endpoints. Although the contrast itself is not a causal bond, its hierarchical composition as a net of distinct causal structures vouches for a separate category. *“Eating chips makes you thirsty, while eating celery gives opposite results”*.

3.2 Large Language Models

The models chosen for the evaluation step were selected keeping in mind both their performance and their origin. Only one model for each major competitor in the field of generative AI was allowed, restricting the choice to their most popular large architecture according to user ratings². This choice was motivated by the intention to mimic the average user’s choice between the multiple options. The chosen LLMs are the following, presented in alphabetical order along with further

information regarding their overall performance and peculiarities. Of important notice is that all of the performance reported is dated to May 2025, and the presence of newer iterations of the following architectures (as well as different datasets) might make the following information outdated.

Claude 3.7 Sonnet was one of the latest models in the Claude line produced by Anthropic, with less of a focus on mathematical reasoning and more “shifted toward the everyday occurrences” as mentioned in [Anthropic \(2025\)](#). The latest benchmark on GPQA scored 84.8% effectiveness.

Command R7B was one of the smaller models parameter-wise, being at the time the largest ever produced by Cohere. In the latest benchmark available, it was reported to be outperforming both Minstral 8B and Llama 3.1 8B on the GPQA set. ([Cohere, 2024](#)).

DeepSeek V3 0324 was the latest release by DeepSeek, performing very effectively on mathematical reasoning datasets. The benchmark was reported to be around 68.4% on GPQA-Diamond ([DeepSeek-AI et al., 2024](#)).

Gemini 2.5 Flash Preview 04-17 was the current preview iteration of Google’s best-performing model, reported to have an integrated “thinking system” ([Google, 2025](#)). The latest benchmark performance on GPQA was reported at 82.8% .

GPT-4o-mini was one of the more affordable iterations of the GPT model by OpenAI for the general public, and one of their best-performing small models. It had recorded performances around 40% on the GPQA benchmark, but 80% on the MMLU benchmark for linguistic tasks ([OpenAI, 2024](#)).

Llama 3.3 70B Instruct was one of the large-range models developed by Meta, and it had recorded benchmark performances of 50.5% on GPQA-Diamond and 80% on MMLU ([Meta, 2024](#)).

Mistral Nemo was a small, lightweight language model built by Mistral AI in collaboration with NVIDIA, with a 68% benchmark on the MMLU dataset ([AI, 2024](#)).

Nova Lite 1.0 was one of the models from the Nova series by Amazon, with benchmark results comparable to GPT-4o and Claude 3.5 Haiku on both GPQA-Diamond and MMLU datasets ([AWS, 2024](#)).

Qwen3 235B A22B was the latest release from Qwen, with benchmark results comparable to

²<https://openrouter.ai/rankings>,
<https://lmarena.ai/leaderboard/text>

the ones by DeepSeek and Claude on multiple benchmarks (Team, 2025).

In order to ensure a common processing baseline for all the large language models, and as much ease of reproduction of the evaluation as possible, we chose to deploy all the instances of model calls through the OpenRouter API (OpenRouter, 2023). The specifics regarding the parameters chosen for calling the models are presented in the following section.

4 Evaluation Setup

4.1 General Settings

Multiple parameter values had to be taken into account for each model call, resulting in the following decisions for the experimental setup. The *temperature* was set to 0 for each model call, to avoid incongruences between different runs and allow for the highest chance of reproducibility of results. The optional settings *max_tokens* and *response_format* were set to, respectively, 256 (when expecting longer types of output) and “*type*”:“*text*” to ensure that a) multimodal architectures would output information in the correct format and b) limit the possible amount of tokens to avoid unexpected generation-loop issues. Every other setting, apart from the temperature, was left unchanged from the default values of the OpenRouterAPI calling functions. This choice was made in order to streamline the process between data input and output as much as possible, and limit human-biased alterations looking for the “optimal configuration” for each individual model, thus influencing each model’s base performance. However, a fully deterministic output with a chosen *seed* was deemed as too restrictive and non-descriptive of both standard user behaviour and humanlike output; in order to counterbalance the variability in output quality, the reported results were aggregated from the best results from each architecture out of three separate runs.

The standard parameters *top_p*, *top_k*, *frequency_penalty*, *presence_penalty*, *repetition_penalty*, *min_p* and *top_a* fell back to the following values, in order: 1, 0, 0, 0, 0, 1, 0, 0.

4.2 Prompting Techniques

The base intuition was to provide the models with templates following a generally demonstrated curve in performance for LLM evaluations (Ziyu

et al., 2023): from zero-shot equivalent to few-shot equivalent (Liu et al., 2024; Lee et al., 2023), followed by chain-of-thought reasoning (Cheng et al., 2024; Chatwal et al., 2025). Therefore, a set of three different templates was developed with the aforementioned properties in order to properly assess model performance when expecting different kinds of output and input complexity. In order to provide as objective an evaluation as possible, the prompts were designed to be neutral, short and direct. While this might have led to suboptimal results, using different prompts tailored to cater to the strengths of each individual model would have ultimately prevented an objective evaluation. A description of each prompt is available below, and each template is provided within the `ipynb` notebook in the GitHub repository³.

- **Zero-shot Equivalent.** The template included the following information, in order: the model role (“*You are an expert in identifying causal links. Perform classification for an input sentence according to the following categories*”); a list with the definition of each category (with the format “****Category****:*Definition*”); the input sentence to be evaluated; and the instruction to only output the name of the appropriate category for the evaluated sentence (“*Your response must ONLY be the name of the category the sentence belongs to. No other text or explanation*”).
- **Few-shot Equivalent.** The template included the same information contained in the previous one, with the addition of an input and output example for each category of evaluation positioned after the definition list.
- **Chain-of-thought Reasoning.** Similarly to the Few-shot Equivalent, the template included the information contained in the Zero-shot Equivalent. The input and output sentences for each category were also included, with a slight change: instead of the output being only the category label, a sequence of reasoning steps that explained the choice of the category was provided. The final instruction of “only output the category label” was discarded, as it would have been counterproductive.

³<https://github.com/gima9552/LLM-Causality-Classification>

The example sentences were handpicked by the author and chosen for their ease of understanding and average length. All example sentences remained the same for both the few-shot and the chain-of-thought prompts.

5 Results and Analysis

In the previous section, the experimental setup was established and presented alongside the LLMs and the prompts used for each analysis setup. Here, we discuss the result obtained through a comparison between the labels produced by the LLMs and the human-annotated ground truth labels given in the dataset. In the case of chain-of-thought prompting, the answers were manually cleaned by the author, and only the final explicitly assigned label was used for the evaluation of model accuracy. Empty outputs, broken sentences, and additional hallucinated inputs (examples of which are provided in Table 2) were labelled as “no explanations”, whereas hypotheticals that presented an explicit label were categorised accordingly. The results are reported by increasing prompt complexity, from *zero-shot* equivalent to *few-shot* equivalent and concluding with *chain-of-thought* equivalent. In Table 1, the F1 scores for the performance of all models are reported, divided by template.

model	zero	few	c-o-t
claude-3.7-sonnet	0.688	0.842	0.816
command-r-08-2024	0.504	0.654	0.658
deepseek-chat-v3-0324	0.588	0.684	0.710
gemini-2.5-flash-preview	0.654	0.827	0.846
gpt-4o-mini	0.596	0.676	0.721
llama-3.3-70b-instruct	0.636	0.724	0.746
mistral-nemo	0.551	0.735	0.353
nova-lite-v1	0.614	0.651	0.713
qwen3-235b-a22b	0.368	0.331	0.070

Table 1: Results of model evaluation expressed via micro-F1 score. The models are presented in alphabetical order, and the two highest scores per template are in bold.

Despite the difference in size, production date, knowledge cutoff and performance on other tasks, almost all LLMs performed above the 0.50 threshold when prompted with the zero-shot equivalent template. The only notable exception was Qwen3, which encountered issues with the token-generation limit and produced enough empty outputs to underperform compared to the average per-

formance value. As the results provided in Table 1 show, Claude and Gemini were the best-performing models within this category, closely followed by Llama 3.3 70B.

With the inclusion of example sentences in the prompts, the performance of nearly all LLMs improved by a minimum of 0.037 (Nova Lite), with the largest margin of improvement shown by Mistral Nemo (0.184). However, compared to the previously clean outputs, this prompt template led to the occasional generation of input sentences along with the required output. A striking example was Nova Lite hallucinating 8 extra input sentences, which were recognised as additions upon reading the model output, and failing to score them appropriately. Both Command and Deepseek produced empty output lines, which were labelled as “no explanation”, and Qwen3 presented the same generation-limit issues previously reported. We decided not to alter the token limitation, as the overarching rule of the template of “only providing the category label” as an output was still standing. On the other side of the output spectrum, both GPT-4o and Claude generated outputs that most closely followed the given directions.

As the prompt-induced rule of “only providing the category label” was removed with the *chain-of-thought* equivalent template, all the outputs required manual postprocessing before an automated evaluation to determine the F1 score. Where Claude, Command, Gemini, GPT, and Nova provided a single-sentence output as presented in the template, DeepSeek, Mistral, and Llama 3.3 produced long token sequences that did not adhere to the suggested pattern. Furthermore, a common occurrence for both Deepseek and Nova was to stray from the “single label” instruction by proposing alternatives and hypotheses (e.g. “However, if the sentence focused on [...] the label should be [...]”). Regarding specific types of hallucinations, the last entry in Table 2 is peculiar, as Deepseek provided two different answers with two different ratings without any instruction to do so. Ultimately, it seemed that the *chain-of-thought* prompt caused the highest rate of issues and hallucinations by the models, including:

- the information presented in the chain-of-thought output not corresponding to the model-assigned label;
- multi-labelling, with assigned labels being contradictory at times;

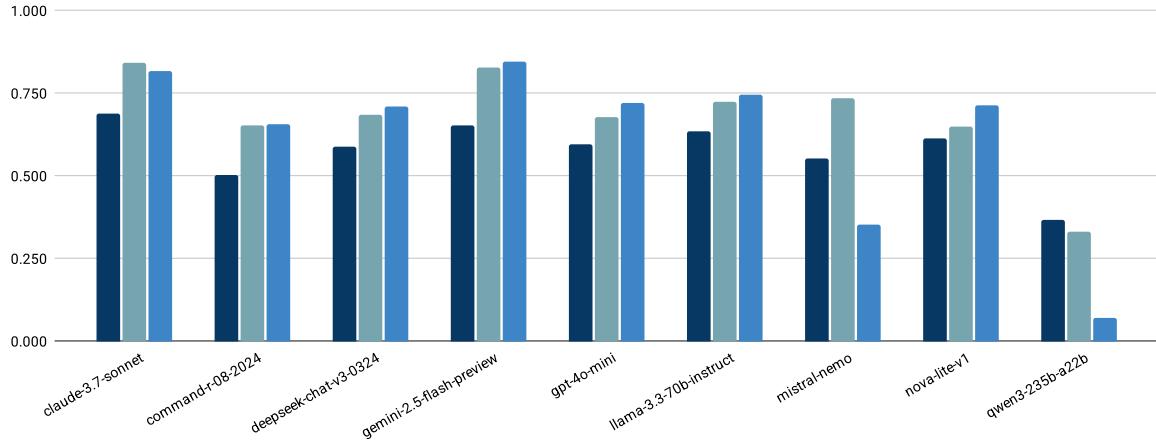


Figure 1: Chart presenting the F1 score for all models in alphabetical order. Each column represents a different prompt, left to right: zero-shot, few-shot, chain-of-thought.

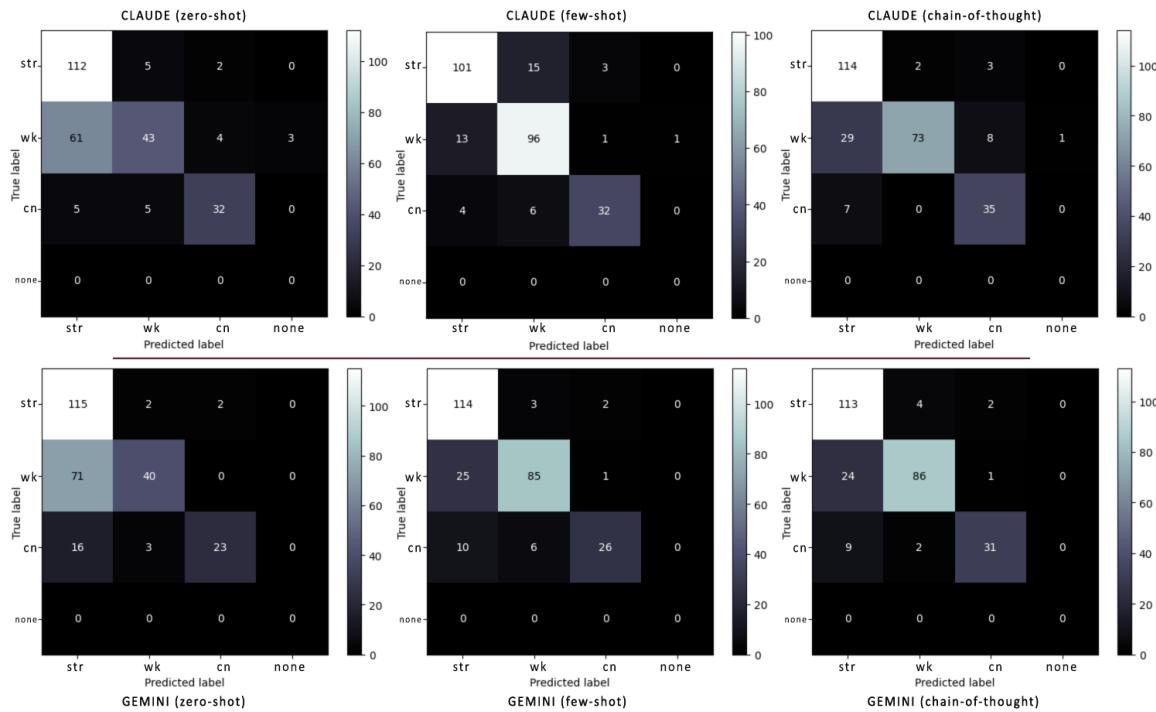


Figure 2: Confusion matrices reporting precision/recall for the two best-performing models, Claude and Gemini, divided by prompt type. The labels *str*, *wk*, *cn*, *none* correspond to *strong*, *weak*, *contrastive*, *no explanation*.

- empty outputs, likely due to token-generation loops;
- half-sentence outputs / well-formed sentences but no labels.

An example list of errors derived by the chain-of-thought prompts is available in Table 2, presenting the model output on the right side and the type of error defined in the middle.

However, as presented in Table 1 and shown in Figure 1, the performance of the majority of the models did not deteriorate despite the presence

of prompt-related issues, showing in fact a positive trend. The exceptions to this were the Claude architecture with a 0.028 minor loss, and both Mistral (0.382 drop) and Qwen3 (0.261 drop) were plagued by empty outputs and broken sentences, possibly due to token-generation issues due to the longer prompt and the lack of text-explicit limitation. Overall, the results appeared in line with general knowledge and previous surveys of LLM evaluations, the use of *chain-of-thought* prompting generally favoured more complete and understandable outputs, with the drawback of context-loss and

model	error type	example
qwen3-235b-a22b	broken sentence	The sentence explains a cause-effect relation where the spines of cacti (cause) directly
nova-lite-v1	extra tokens	<p>**Output**: The sentence explains that broad spectrum antibiotics are used due to the unavailability or delay in obtaining culture data that identifies the infecting bacteria. This indicates a relationship between the unavailability of specific information and the decision to use broad spectrum antibiotics. The category is: Weak Causation</p> <p>**Explanation**: The sentence provides a reason for using broad spectrum antibiotics, but it does not detail a direct cause-effect relationship or multiple steps in the causal process. Instead, it emphasizes the role of the unavailability of specific information, which is an indirect cause.</p> <p>**Explanation**: The sentence provides a reason for using broad spectrum antibiotics, but it does not detail a direct cause-effect relationship or multiple steps in the causal process. Instead, it emphasizes the role of the unavailability of specific information, which is an indirect cause.</p>
mistral-nemo	misclassification, extra tokens	<p>The sentence presents a contrast between two different environments: temperate forest and desert. It explains the outcomes for the fox population in each environment (thriving in the first, going extinct in the second) due to the availability of resources, which is a causal relation. However, the sentence does not go into the details of the causal process like the previous strong causation example. Therefore, the category is: **Weak Causation**.</p> <p>In detail:</p> <ul style="list-style-type: none"> - The sentence presents a contrast between two environments: temperate forest and desert. - The outcome for the fox population in each environment is explained: thriving in the temperate forest and extinction in the desert. - The reason for these outcomes is the availability of resources in each environment, presenting a causal relation. - The sentence does not delve into multiple steps in the causal process, unlike the strong causation example, making it weak causation.
llama-3.3-70b-instruct	extra tokens, hypothetical	<p>The sentence explains a cause-effect relation between the climate and the migration patterns of birds. The birds in the northern hemisphere migrate due to the lack of food and suitable breeding grounds during winter, whereas the birds in tropical regions do not migrate because the consistent climate provides stable resources. The category is: **Strong Causation**.</p> <p>However, it can also be argued that the sentence presents a comparison between two different situations (birds in the northern hemisphere vs birds in tropical regions), which could classify it as **Contrastive**. But since the primary focus is on explaining the cause-effect relation between climate and migration patterns, the more suitable category would be **Strong Causation**.</p>
deepseek-chat-v3-0324	misclassification, extra tokens	<p>**Output**: The sentence explains a cause-effect relation between sunlight and plant growth, detailing how ample sunlight enhances photosynthesis and promotes growth, while insufficient light limits energy production and growth. It also contrasts two scenarios (sunny field vs. shaded forest) to highlight the opposing outcomes. The primary focus is on the causal mechanism, with a secondary contrastive element. The category is: **Strong Causation**.</p> <p>*(Rationale: The sentence primarily describes a direct causal process (sunlight → photosynthesis → growth), which fits “Strong Causation.” While it includes a contrast (sunny vs. shaded), the contrast serves to reinforce the causal explanation rather than being the main focus.)*</p>
deepseek-chat-v3-0324	misclassification, extra tokens, hypothetical, input duplication	<p>**Output**: The sentence explains a cause-effect relation between the degeneration of neurons in the hippocampus and the impairment of memory and spatial navigation in Alzheimer’s disease, while contrasting it with the functionality in healthy brains. The category is: **Strong Causation** (due to the clear mechanistic explanation of neurodegeneration leading to cognitive deficits, with a secondary contrastive element). Alternatively, if the **contrastive** aspect is emphasized more strongly: The sentence compares the effects of hippocampal degeneration in Alzheimer’s disease (impaired memory) with the preserved function in healthy brains (clear memory processing), presenting two opposing scenarios. The category is: **Contrastive**.</p> <p>*(The classification depends on whether the primary focus is the causal mechanism within Alzheimer’s or the comparison between disease and health. Here, the causal chain is more detailed, so “Strong Causation” is likely the better fit.)*</p>

Table 2: Examples of erroneous outputs from chain-of-thought prompting, with the types of error explicated in the middle column.

heightened focus of the generation of text based on specific token-cues from the prompt with the lengthening of the outputs.

By scrutinising the performance of the Gemini 2.5 and Claude 3.7 Sonnet models, the two best-performing ones, it is possible to highlight the weak points in causal categorisation previously assessed in Jin et al. (2024). As seen in Figure 2, for both models, there was consistent misclassification of *weak causation* and *contrastive* labels as *strong causation*, represented by *wk*, *cn* and *str*, respectively. While this effect might lead to think that the dataset is unbalanced, the split between the labels is 40/40/20% with the most represented categories being *strong causation* and *weak causation*; therefore, the misclassification is probably indicating a lack of effectiveness by the architectures in discerning the less syntactic-oriented types of causal links. It is worth mentioning that the few-shot performance of Claude 3.7 Sonnet (upper centre in Figure 2) presents a generalisation of the issue on both sides of the causal-strength spectrum, as there is almost equal misclassification of strong and weak causal bonds. Despite this weak point still being present in the evaluated LLMs, it is clearly visible in the provided confusion matrices that the accuracy in the classification of different causal bonds with varying strength is promisingly high.

6 Conclusions

This paper presented the evaluation of 9 different Large Language Models, of diverse proprietary nature, as classifiers of causal bonds between sentences in a natural language dataset. Three prompt templates were developed and used to gradually increase context and provide the models with additional information and point verbal reasoning capabilities in the appropriate direction. When tested with zero-shot, few-shot, and chain-of-thought-based prompts, the best models performed with an F1 score of 0.688, 0.842 (Claude 3.7 Sonnet zero and few-shot), and 0.846 (Gemini 2.5 Flash chain-of-thought). The pipeline code has been rewritten as a `ipynb` notebook and made available at a dedicated repository⁴.

Our evaluation seems to reinforce the hypothesis that LLMs can discern different types of sentence-internal causal bonds, more reliably so if provided with example-based prompts (few-shot) that pro-

vide an explicit definition of such causal properties. This implementation suggests that there is no real need for more complicated chain-of-thought prompts when it comes to the analysis and explanation of sentences from a classifying standpoint; however, manual analysis of the chain-of-thought outputs reveals that for the best-performing models, the generated “line of thought” is correct on average despite the occasional hallucination. The proposed idea in previous works that LLMs struggle with precise distinction between correlation and causation is reinforced by confusion-matrix analysis, as the misclassification of the two is manifest even in the best output cases. The dataset taken into consideration, albeit simple on a semantic level and of reduced size, is still based on the general communication patterns that humans use; this implies some generalizable proficiency by the LLMs to operate effectively in everyday causal-analysis situations.

Further research work is required to reinforce the hypotheses confirmed by this paper, especially relating to the ability of LLMs to truly identify underpinning causal links. Possible avenues of research should be directed to the analysis of causal generation in a contextless environment, more so than testing models on what could be a consequence of mere token-context performance. The use of effective state-of-the-art models, rather than the popular ones due to price or ease of implementation, should also be reinforced in future research. Furthermore, the use of model-tailored prompts instead of general ones, with more stringent instructions, could help prevent hallucinations as much as possible. Possibly, all models should be tested on multiple natural-language causal datasets, with the best case scenario allowing for different domains to avoid topic specificity, and several runs of the same model on standardised settings would be necessary.

As can be inferred from above, this work presents some clear limitations: the choice of a semantically simple dataset, the choice of models based on user popularity, the use of single-sentence classification examples, and the grounding of the causal analysis in natural language relations. The latter, especially, is an intrinsic hurdle when applying concepts of causality to information disjointed from the statistically-informed causal graphs and relations. Furthermore, the implementation of more in-depth chain-of-thought prompt choices and more complex methods of analysis might have led to a

⁴<https://github.com/gima9552/LLM-Causality-Classification>

different set of results, alongside the presentation of fully context-free samples.

References

- Mistral AI. 2024. [Mistral nemo](#).
- Anthropic. 2025. [Claude 3.7 sonnet and claude code](#).
- Swagata Ashwani, Kshiteesh Hegde, Nishith Reddy Mannuru, Dushyant Singh Sengar, Mayank Jindal, Krishna Chaitanya Rao Kathala, Dishant Banga, Vinija Jain, and Aman Chadha. 2024. [Cause and effect: Can large language models truly understand causality?](#) *Proceedings of the AAAI Symposium Series*, 4(1):2–9.
- Amazon AWS. 2024. [Amazon nova foundation models](#).
- Pulkit Chatwal, Amit Agarwal, and Ankush Mittal. 2025. [Enhancing causal relationship detection using prompt engineering and large language models](#). In *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, pages 248–252, Abu Dhabi, UAE. Association for Computational Linguistics.
- Xiaoxue Cheng, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. 2024. [ChainLM: Empowering large language models with improved chain-of-thought prompting](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2969–2983, Torino, Italia. ELRA and ICCL.
- Cohere. 2024. [Introducing command r7b: Fast and efficient generative ai](#).
- DeepSeek-AI, :, Xiao Bi, Deli Chen, Guanting Chen, Shanhua Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, Huazuo Gao, Kaige Gao, Wenjun Gao, Ruiqi Ge, Kang Guan, Daya Guo, Jianzhong Guo, Guangbo Hao, Zhewen Hao, Ying He, Wenjie Hu, Panpan Huang, Erhang Li, Guowei Li, Jiashi Li, Yao Li, Y. K. Li, Wenfeng Liang, Fangyun Lin, A. X. Liu, Bo Liu, Wen Liu, Xiaodong Liu, Xin Liu, Yiyuan Liu, Haoyu Lu, Shanghao Lu, Fuli Luo, Shirong Ma, Xiaotao Nie, Tian Pei, Yishi Piao, Junjie Qiu, Hui Qu, Tongzheng Ren, Zehui Ren, Chong Ruan, Zhangli Sha, Zhihong Shao, Junxiao Song, Xuecheng Su, Jingxiang Sun, Yaofeng Sun, Minghui Tang, Bingxuan Wang, Peiyi Wang, Shiyu Wang, Yaohui Wang, Yongji Wang, Tong Wu, Y. Wu, Xin Xie, Zhenda Xie, Ziwei Xie, Yiliang Xiong, Hanwei Xu, R. X. Xu, Yanhong Xu, Dejian Yang, Yuxiang You, Shuiping Yu, Xingkai Yu, B. Zhang, Haowei Zhang, Lecong Zhang, Liyue Zhang, Mingchuan Zhang, Minghua Zhang, Wentao Zhang, Yichao Zhang, Chenggang Zhao, Yao Zhao, Shangyan Zhou, Shunfeng Zhou, Qihao Zhu, and Yuheng Zou. 2024. [Deepseek llm](#):
- Scaling open-source language models with longtermism.
- Daniela Garcia. 1997. Coatis, an nlp system to locate expressions of actions connected by causality links. In *Knowledge Acquisition, Modeling and Management*, pages 347–352, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Google. 2025. [Gemini 2.5 flash documentation](#).
- Joseph Y Halpern and Judea Pearl. 2005. Causes and explanations: A structural-model approach. part ii: Explanations. *The British journal for the philosophy of science*.
- Paul Henne, Kevin O'Neill, Paul Bello, Sangeet Khemlani, and Felipe De Brigard. 2021. [Norms affect prospective causal judgments](#). *Cognitive Science*, 45(1):e12931.
- Thomas F. Icard, Jonathan F. Kominsky, and Joshua Knobe. 2017. [Normality and actual causal strength](#). *Cognition*, 161:80–93.
- Alon Jacovi, Swabha Swayamdipta, Shauli Ravfogel, Yanai Elazar, Yejin Choi, and Yoav Goldberg. 2021. [Contrastive explanations for model interpretability](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1597–1611, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Zhijing Jin, Jiarui Liu, Zhiheng LYU, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona T. Diab, and Bernhard Schölkopf. 2024. [Can large language models infer causation from correlation?](#) In *The Twelfth International Conference on Learning Representations*.
- Christopher S. G. Khoo, Syin Chan, and Yun Niu. 2000. [Extracting causal knowledge from a medical database using graphical patterns](#). In *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics, ACL '00*, page 336–343, USA. Association for Computational Linguistics.
- Emre Kıcıman, Robert Osazuwa Ness, Amit Sharma, and Chenhao Tan. 2023. [Causal reasoning and large language models: Opening a new frontier for causality](#). *Trans. Mach. Learn. Res.*, 2024.
- Unggi Lee, Haewon Jung, Younghoon Jeon, Younghoon Sohn, Wonhee Hwang, Jwoong Moon, and Hyeyoncheol Kim. 2023. [Few-shot is enough: exploring chatgpt prompt engineering method for automatic question generation in english education](#). *Education and Information Technologies*, 29(9):11483–11515.
- David Lewis. 1986. Causal explanation. In David Lewis, editor, *Philosophical Papers, Volume II*, pages 214–240. Oxford University Press.
- Yilun Liu, Shimin Tao, Weibin Meng, Feiyu Yao, Xiaofeng Zhao, and Hao Yang. 2024. [Logprompt](#):

- Prompt engineering towards zero-shot and interpretable log analysis. In *Proceedings of the 2024 IEEE/ACM 46th International Conference on Software Engineering: Companion Proceedings, ICSE-Companion '24*, page 364–365, New York, NY, USA. Association for Computing Machinery.
- Giacomo Magnifico and Eduard Barbu. 2025. Emerging categories in scientific explanations.
- Helena Matute, Fernando Blanco, Ion Yarritu, Marcos Díaz-Lago, Miguel A. Vadillo, and Itxaso Barbería. 2015. Illusions of causality: how they bias our everyday thinking and how they could be reduced. *Frontiers in Psychology*, Volume 6 - 2015.
- Meta. 2024. Llama 3.3: Model cards & prompt formats.
- Antonio Moreno Sandoval, Jordi Porta, Blanca Carballo Coronado, Yanco Torterolo, and Doaa Samy. 2025. The financial document causality detection shared task (FinCausal 2025). In *Proceedings of the Joint Workshop of the 9th Financial Technology and Natural Language Processing (FinNLP), the 6th Financial Narrative Processing (FNP), and the 1st Workshop on Large Language Models for Finance and Legal (LLMFinLegal)*, pages 214–221, Abu Dhabi, UAE. Association for Computational Linguistics.
- OpenAI. 2024. Gpt-4o mini: advancing cost-efficient intelligence.
- OpenRouter. 2023. Openrouter documentation.
- Judea Pearl. 2009. Causal inference in statistics: An overview. *Statistics Surveys*, 3(none):96 – 146.
- David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.
- Angelika Romanou, Syrielle Montariol, Debjit Paul, Leo Laugier, Karl Aberer, and Antoine Bosselut. 2023. CRAB: Assessing the strength of causal relationships between real-world events. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15198–15216, Singapore. Association for Computational Linguistics.
- Qwen Team. 2025. Qwen3: Think deeper, act faster.
- Zimu Wang, Lei Xia, Wei Wang, and Xinya Du. 2024. Document-level causal relation extraction with knowledge-guided binary question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16944–16955, Miami, Florida, USA. Association for Computational Linguistics.
- Jie Yang, Soyeon Caren Han, and Josiah Poon. 2022. A survey on extraction of causal relations from natural language text. *Knowl. Inf. Syst.*, 64(5):1161–1186.
- Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. 2023. Causal parrots: Large language models may talk causality but are not causal. *Transactions on Machine Learning Research*, 2023(8).
- Zhuang Ziyu, Chen Qiguang, Ma Longxuan, Li Mingda, Han Yi, Qian Yushan, Bai Haopeng, Zhang Weinan, and Ting Liu. 2023. Through the lens of core competency: Survey on evaluation of large language models. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 2: Frontier Forum)*, pages 88–109, Harbin, China. Chinese Information Processing Society of China.