University of Bergen
Department of Informatics

Prompting and Causal Fallacy Detection in Large Language Models

Master's thesis (30 ECTS)

Author: Mariann Teigland Lepsøy
Supervisor: Jan Arne Telle

UNIVERSITY OF BERGEN
Faculty of Science and Technology

January, 2026

Abstract

This thesis examines how a large language model evaluates causal fallacies in short, natural language claims. The study addresses three primary questions: how accurately the model detects different types of causal fallacies, how prompting styles influence its judgements, and how well the model's explanations reflect its reasoning. To investigate these questions, the model was evaluated on nine types of causal fallacies and a set of non-fallacious causal claims, using five prompting styles ranging from zero-shot to counterfactual.

The results indicate that the model performs well on causal fallacies with strong and familiar surface patterns, such as post hoc fallacies. However, it struggles with fallacies that require attention to missing causes or multiple interacting factors. The model's performance on non-fallacious claims also depends on the prompt. Some prompts reduce false positives, but may increase missed fallacies.

Analysis of the model's explanations reveals that correct responses have focused justifications, while incorrect responses offer explanations that rationalise the error rather than reveal the reasoning process. These findings suggest that prompting can direct the model's attention but does not change its underlying associative approach to causal reasoning.

In summary, this study identifies the context in which language models can assist with fallacy detection and those in which they are limited. The results demonstrate that language models can help identify potential causal errors, but should not be regarded as independent evaluators of causal arguments. Human oversight remains essential for tasks involving causal assessment. Additionally, the findings suggest directions for future research on language models and causal fallacies.

Acknowledgements

I want to thank my supervisor Jan Arne Telle for suggesting the topic of this thesis and for guiding me throughout the project. The feedback, advice, and support have been indispensable.

I am also grateful to my co-supervisor Dario Garigliotti for his early guidance, including suggestions on relevant literature, data sources, and how the experiment might be approached. His feedback on the thesis draft also helped improve the final version.

Additionally, I would like to thank the institute for funding the OpenAI API usage required to run the experiment.

Mariann Teigland Lepsøy
Tuesday 20th January, 2026

Contents

List of Figures

List of Tables

1 Introduction

Large Language Models (LLMs) are playing an increasing role in tasks that require explanation, judgement and reasoning. They are often treated as tools that can analyse arguments or evaluate the strength of a claim. However, it remains unclear how well they handle the kinds of reasoning errors humans routinely make. Causal fallacies are a clear example. They are persuasive, common in everyday language, and easy to overlook if one does not look at the underlying logic. If LLMs have inherited these tendencies from their training data, their outputs can appear confident even though their reasoning is flawed.

Research has shown that LLMs can recognise some fallacies, but the work so far leaves a couple of gaps. Most studies treat causal fallacies as a single category, which hides differences between the specific types of errors that models can or cannot detect. Also, prompting is known to shape model behaviour, however, we know little about how different prompting styles change the way a model judges causal claims. These gaps mean that we are still missing a clear picture of what the model is judging when it evaluates a causal claim. A closer look at individual fallacy types and the role of prompting techniques can help reveal if the model is reasoning or relying on associative patterns. This thesis takes these concerns as its starting point.

This thesis investigates how the LLM GPT-4.1 mini evaluates individual causal fallacies when given short causal claims and different prompting instructions. GPT-4.1 mini is a lightweight version of OpenAI's GPT-4.1 language model. Larger and more advanced models exist, and model development is progressing quickly, but GPT-4.1 mini is well-suited for a controlled experiment where the goal is to study reasoning behaviour rather than maximising performance. The study offers insight into how consistently the model can separate sound reasoning from common causal errors in short causal statements by comparing its judgements and explanations across prompting styles.

Previous work provides useful insight into fallacy detection and causal reasoning in LLMs, but it does not combine the elements needed to examine how prompting influences specific causal fallacies. This study addresses four gaps in existing literature:

- It evaluates individual causal fallacy types rather than treating them as a single broad category.
- It isolates the effect of prompting by keeping the model fixed and varying only the prompt.
- It analyses differences in accuracy across fallacy types.
- It includes a qualitative analysis of the model's explanations to examine whether the justifications align with its judgements.

These contributions motivate the overall aim of the thesis, which is to evaluate how an LLM judges causal claims. Specifically, the study investigates how well the model detects different causal fallacies, how prompting styles influence its judgement, and what the model's explanations reveal about the reasoning process. These aims lead to the following research questions:

- How accurately does the model detect different types of causal fallacies?
- How do prompting styles affect its fallacy judgement?
- How do the explanations generated by the model reflect its reasoning process?

Before continuing, the reader can evaluate a few causal claims to get a sense of the task the model was asked to solve. The examples and their fallacious and non-fallacious labels were taken from existing sources. The solution can be found in Appendix B.

Decide if the claim's causal reasoning is fallacious.
Answer *Yes* if the reasoning is fallacious and *No* if it is not.

	Claim	Yes	No
1.	The economy has improved greatly because of the new president.	☐	☐
2.	According to the weather bureau, there will be a cyclone in the eastern part of our city in the next two days. In the next two days, there wil be strong winds and rain in the eastern part of our city.	☐	☐
3.	The missile hit the plane's ailerons. The plane rolled out of balance and fell to the ground.	☐	☐
4.	I bought a ticket to win a new car at the mall, since I have never won anything like that in the past.	☐	☐
5.	In cold weather, Tom put on insulated clothing. His heat loss from the skin decreased.	☐	☐
6.	We need to stop allowing colleges to increase tuition every year. The next thing we know, it's going to cost more to attend college for one semester than it is to buy a new home!	☐	☐

Table 1.1: Self-test on causal fallacies. The table presents a set of claims used to illustrate the type of reasoning evaluated in this study.

The following chapters outline how the study is carried out and how the results are analysed. Chapter 2 introduces the theoretical background and includes the different types of causal fallacies used in this study, how prompting can influence model behaviour, and what earlier research has found about LLMs and reasoning. Chapter 3 describes the experimental design, including how the dataset was created, how the prompts were constructed, and how the model responses were evaluated. Chapter 4 presents the quantitative results of the model's fallacy detection accuracy. Chapter 5 discusses what the results suggest about the strengths and limitations of GPT-4.1 mini regarding causal reasoning, expands on the results with a qualitative analysis of the model's explanations, and considers the broader implications of the findings and suggests directions for future work.

2 Theory and Methods

2.1 Causal Reasoning

Causal reasoning is the process of identifying and understanding cause and effect, not just describing statistical associations between events. It shows how changes in one factor affect another, supporting explanations, predictions, and decision-making in complex environments. Pearl (2009) presents a comprehensive theory of causality, which has become central across statistics, artificial intelligence, and cognitive science. The book describes three levels of causal reasoning, called the causal hierarchy, which differ in how deeply they capture causal structure. The first level, association, involves recognising regularities in data and supports predictions. The second level, intervention, concerns what happens when a change or action is made, which is important for planning. The third level, counterfactual reasoning, involves asking what would have happened if conditions were different, allowing for evaluations of outcomes. These levels are important because they outline the kinds of questions a reasoner or model can answer.

Distinguishing between correlation and causation is important to avoid erroneous conclusions. Correlation shows that two variables move together, but it does not reveal why. Causation indicates that one variable directly influences another. Yao et al.'s (2021) overview of causal inference methods shows that this distinction is central to modern causal models and reasoning frameworks. These models formalise how interventions and counterfactuals can be inferred from data through structured representations of causal relationships. The ability to reason this way determines whether we can explain observed patterns, predict the effects of actions, and make informed decisions (Pearl, 2009).

Humans reason about cause and effect intuitively. We often assume that if one event happens before another, or if two events regularly co-occur, one must have caused the other. These shortcuts help us make sense of the world, but also lead to systematic errors

in reasoning. People frequently treat correlation as causation or overlook alternative explanations, such as a common underlying cause. Research in cognitive psychology, including Tversky and Kahneman's (1973) framework on heuristics and biases, shows that causal judgements often rely on simple heuristics rather than deeper structural reasoning. These heuristics increase efficiency but also introduce consistent biases in judgement. This pattern corresponds to several classical fallacies in which reasoning follows surface patterns rather than causal structure, as discussed by Hurley (2012) in his introduction to logic. The same shortcuts that make human judgement fast and intuitive also introduce inconsistencies and biases, a limitation that becomes relevant later when comparing human fallacies with how artificial systems approach causal reasoning.

Artificial intelligence, including LLMs, is primarily trained to model statistical patterns in data, which aligns most closely with the associative level of Pearl's (2009) causal hierarchy. Their goal is to predict the next word in a sequence, encouraging recognition of regularities rather than structured interventions or counterfactual reasoning (Yao et al., 2021). However, studies such as Kıcıman et al. (2024) show that LLMs can generate text that resembles interventional or counterfactual reasoning when prompted. This should not be mistaken for genuine causal understanding, since models may produce explanations that sound causal without engaging in causal reasoning.

When asked to make causal judgements, language models often rely on correlations from their training data. This can lead to reasoning errors similar to those humans make using intuition. In a controlled study on causal inference, Joshi et al. (2024) show that even when LLMs are exposed to structured causal data, their answers often reflect memorised associations rather than general rules of cause and effect. The idea of the "causal parrot", introduced by Zečević et al. (2023), captures this behaviour, suggesting that these models mimic causal language from their training data without real causal understanding. Other research points to the same conclusion. Carro et al. (2024) investigates causal illusion biases in LLMs and find that although generative models can express causal relations in language, they struggle to reason about hypothetical causal situations. This keeps causal reasoning as an open challenge in artificial intelligence.

Because LLMs learn from patterns in text, they often reproduce the same shortcuts that lead to human causal errors. Their reasoning depends on surface features such as event order or co-occurrence, making them prone to assume causation where only correlation exists (Carro et al., 2024). When tasks require distinguishing between genuine causal links and coincidental ones, models frequently default to associative reasoning and repeat familiar explanations (Zečević et al., 2023). These tendencies mirror human causal

fallacies, showing that the models imitate not only language but also patterns of flawed inference. Understanding this overlap is central to this study, which examines how causal fallacies can reveal the specific reasoning errors that occur when large language models attempt to make sense of cause and effect. The next section introduces the fallacies examined in this study and explains how each represents a specific failure of causal inference.

2.2 Causal Fallacies as Reasoning Errors

Fallacies are mistakes in reasoning that make an argument appear better than it is. They are commonly divided into formal fallacies, errors that arise from the argument's logical form, and informal fallacies, which concern the content of the reasoning rather than its structure (Hurley, 2012). Causal fallacies are a subset of informal fallacies. They involve reasoning errors that assert a causal link without sufficient evidence. In informal logic, causal fallacies are often classified as fallacies of weak induction, where the evidence provides insufficient probabilistic support for the conclusion (Hurley, 2012). This category includes post hoc and cum hoc, which mistake temporal sequence or co-occurrence for causation. The flaw is not distraction or misuse of language, but weak evidential reasoning. Recent computational research groups causal fallacies with other logical fallacies when studying reasoning behaviour in LLMs, as shown in studies by Jin et al. (2022) and Li et al. (2024). Although informal logic and computational studies organise causal fallacies differently, both perspectives identify them as a key form of faulty inference that reveals how easily correlation can be mistaken for causation.

Several recurring patterns show how causal reasoning can fail in both everyday life and analytical contexts. Table 2.1 gives an overview of the causal fallacy types used in this experiment. Each fallacy represents a distinct breakdown in causal inference. They are relevant here because they highlight the types of causal reasoning failures LLMs could reproduce.

Fallacy	Definition	Example
Causal Oversimplification	Also known as the fallacy of the single cause. It occurs when one assumes that only event A caused event B, even though multiple factors contributed.	"Students fail because they're lazy."
Confusing Necessary with Sufficient Condition	Occurs when a sufficient cause is mistakenly assumed to be necessary, or when one assumes that an outcome cannot occur without a particular sufficient cause.	"I don't know why the car won't run; I just filled the gas tank."
Cum Hoc Ergo Propter Hoc	Latin for "with this, therefore because of this". It infers a causal link between events that occur together without evidence of causal direction.	"People who eat yogurt have healthy guts. If I eat yogurt I will never get sick."
Gambler's Fallacy	Mistakenly believing that statistical patterns influence the likelihood of independent events, treating chance patterns as causal.	"That family has had three girl babies in a row. The next one is bound to be a boy."
Neglecting a Common Cause	Falsely inferring causation between two events while overlooking the possibility that both could be caused by a third event.	"The bigger a child's shoe size, the better the child's handwriting"
No Fallacy	The causal statement contains no unsupported or incorrect causal inference.	"The frogs have spawned these days. The pond is teemed with eggs."
Post Hoc Ergo Propter Hoc	Latin for "after this, therefore because of this". It is the fallacy of inferring causation solely because one event follows another.	"I bought new shoes then my head started itching. I must be allergic to the shoes."
Regression Fallacy	Assuming that a change in one variable caused a change in another, when in reality the change is due to regression to the mean.	"A tall father concluded that his tall wife committed adultery because their children were shorter."
Reverse Causation	Inferring a causal direction between correlated events without considering that the true causal direction may be the opposite.	"Lice are benefitial to health because there are rarely any lice on sick people."
Slippery Slope	When an argument is justified by a series of events where each step in the series is not justified.	"Today late for ten minutes, tomorrow late for an hour, and then someday you will simply cease to show up."

Table 2.1: Overview of causal fallacy types and the non-fallacious category included in the experiment.

The fallacy descriptions in Table 2.1 are adapted from standard presentations in Hurley's (2012) book on logic and supplemented with explanations from Wick's website.

Human reasoning often relies on heuristics that simplify judgement but can distort causal inference. People use shortcuts such as the representativeness heuristic, where similarity between events is mistaken for evidence of causation, and the availability heuristic, where easily recalled examples are judged as more likely or typical (Tversky and Kahneman, 1973). These heuristics allow efficient judgements, but also make people more prone to errors such as post hoc reasoning and neglecting alternative explanations. Another tendency is confirmation bias, the preference for evidence that supports existing beliefs while discounting contradictory information, as described by Nickerson (1998). These biases explain why causal fallacies arise naturally in human thinking. They result from efficient but imperfect reasoning strategies, not random mistakes. Because LLMs are trained on human-generated text, they may inherit similar reasoning patterns from their training data. This connection motivates the following experiment, which examines whether the fallacies seen in human reasoning also appear in model outputs.

2.3 Large Language Models and Reasoning

LLMs are advancing quickly, with new versions offering improved reasoning capabilities. This rapid progress is relevant background, but not the focus of this thesis. Instead, this section describes the general principles that shape how current LLMs reason, independent of any model release. Although newer models may improve, the underlying mechanisms remain largely the same. They learn from patterns in data, respond to prompts based on those patterns, and can be guided to generate text that resembles structured reasoning. The discussion here focuses on these characteristics rather than the fast-paced development of newer models.

LLMs perform what we can call linguistic reasoning. They generate responses by predicting the next word based on patterns found in a wide range of human data. Their reasoning comes from reproducing patterns that reflect human reasoning, not from an ability to understand cause and effect. The models link words and concepts that frequently appear together in data, which resembles association more than structured inference. Therefore, they can appear to reason about causality, but it happens through pattern recognition rather than causal understanding, as illustrated by Zečević et al. (2023) and further supported by Jin et al. (2024), who show that models perform poorly

on tasks requiring inference of causal direction alone. Carro et al. (2024) report similar limitations in tasks involving hypothetical or unfamiliar cases. Some studies argue that this limits the models to surface-level reasoning where coherence in language replaces comprehension (Zečević et al., 2023). Other work is more optimistic. In their empirical evaluation of LLMs on structured causal tasks, Kıcıman et al. (2024) find that models can show some forms of causal reasoning when given the right context or prompting. This view opens the possibility that reasoning in models could exist to some degree.

Empirical studies testing causal reasoning in LLMs show that their performance aligns with the pattern-based view of reasoning. They often generate explanations that sound convincing, but rely on patterns of co-occurrence rather than understanding interventions or counterfactuals (Zečević et al., 2023). When tested on causal inference tasks, the model's responses tended to reflect patterns consistent with statistical regularities from training data rather than a genuine understanding of cause and effect (Jin et al., 2024). These findings suggest that models' reasoning is shaped by how causality is expressed in language, not by how causal mechanisms actually work. This pattern mirrors human tendencies to mistake correlation for causation, showing that linguistic patterns can reinforce the same reasoning shortcuts.

An alternative perspective argues that LLMs can sometimes demonstrate causal reasoning under the right circumstances. Research shows that models trained on extensive human-generated text can produce correct causal arguments, even outperforming traditional causal discovery methods on some benchmark tasks (Kıcıman et al., 2024). This ability seems to come from the models' capacity to capture and apply structured human knowledge expressed in language. However, their success is inconsistent across tasks and hard to explain. This view suggests that LLMs can approximate causal reasoning through language generalisation, but their understanding depends on the statistical regularities they learn from text.

Despite some promising findings, most research shows clear limitations in LLMs' reasoning abilities. Factual knowledge and language fluency alone do not guarantee an understanding of causal structure. In their evaluation of how models combine numerical data with contextual knowledge, Cai et al. (2024) find that LLMs often rely on pre-existing associations rather than underlying mechanisms. Their results show that numerical input brings only a minor improvement. Experiments combining textual and numerical input show that models can perform simple causal attributions, but their accuracy drops when the task requires integrating new reasoning across variables. This suggests that LLMs can use numbers to support existing knowledge but still struggle to construct causal relationships from raw data alone. These findings show that even though LLMs can partially

reproduce causal reasoning, their understanding remains context-dependent. This leaves them prone to the same kinds of reasoning errors humans make.

LLMs mostly rely on patterns in text, which can look like causal reasoning without actually being so. They can sometimes give sound causal answers, but this depends on their training data, and the explanations can mirror how humans talk. They are vulnerable to familiar causal slips, especially when correlation is presented as a cause. This can cause the models to drift towards the same fallacies humans make. Therefore, it is possible for the models to sound persuasive while leaning on fallacies.

2.4 LLMs and Fallacies: Empirical evidence

Fallacies let us test how models reason, not just whether they get the correct answer. They expose systematic errors. This is useful for LLMs because their output often sounds persuasive even when the reasoning is flawed. Work on logical fallacies shows that models can mirror human error patterns, making fallacies a practical lens for evaluation.

Several studies frame fallacies as reasoning errors and test whether models can identify them. Jin et al. (2022) performed an empirical study evaluating LLMs on logical fallacy classification. They found that pretrained models showed limited performance, with the best model achieving 53.31 percent F_1. Their designed structure-aware model showed better results, outperforming the best pretrained model by over 5 percent F_1. A separate zero-shot fallacy classification experiment by Pan et al. (2024) found that simple prompting techniques, such as short definitions and multi-round prompting, can improve performance. On some benchmarks, models like GPT-4 achieved F_1 scores close to 80 percent. However, performance varied across datasets and fallacy types, highlighting challenges in relying on LLMs to classify fallacious reasoning. Another empirical study by Li et al. (2024) shows that LLMs struggle with logical reasoning, mainly because they lack an understanding of logical fallacies. They create a what-why-how dataset and show that training LLMs on this dataset enhances their reasoning performance and robustness across logical fallacy types. This suggests that fallacy knowledge can guide the model's decisions. Another study by Payandeh et al. (2024) adds a practical angle, showing how a persuader guides the model's decision using fallacious arguments. This demonstrates susceptibility beyond classification.

Causal fallacies have also been examined in recent studies. Jin et al. (2022) discuss false causality as a fallacy where the argument assumes that correlation implies causation.

Their experiment shows that this fallacy type was one of the most challenging for models to classify, especially when the argument relied on temporal or correlational cues. The results for the false causality fallacy in this study suggest that LLMs treat co-occurrence and sequence as evidence of causation. Joshi et al. (2024) provide empirical evidence that LLMs are prone to causal fallacies, including post hoc. Their experiment on synthetic data shows that models sometimes can detect the absence of a causal relationship based on temporal or spatial cues, but they struggle to infer cause and effect from counterfactual information. The results indicate that models depend on superficial heuristics rather than causal reasoning. This is even when the model is trained with data meant to reduce such bias. Carro et al. (2024) found that LLMs often demonstrate a bias known as the illusion of causality. They often infer causal relationships where none exist and overlook temporal orders that contradict causality. These studies show that causal fallacies are common and persistent errors in LLMs. They appear both when models classify and explain, showing that these errors are part of how LLMs learn and use causal language.

Empirical studies show that LLMs reproduce many of the same fallacies as humans. Causal fallacies in particular remain difficult to avoid, since models rely on surface associations rather than causal structure. Some research suggests that prompting can reduce these errors by guiding the model toward reasoning. The next section will take a closer look at prompting as a possible way to control and evaluate reasoning in LLMs.

2.5 Prompting as a Tool for Reasoning Control

Prompting guides an LLM's output by changing how the input is phrased or structured. It influences how the model reasons, retrieves knowledge, and explains answers. Sahoo et al. (2025) describes prompt engineering as designing inputs that steer a model toward specific behaviours. Small changes in phrasing can affect how a model responds and what kind of reasoning it uses. A well-designed prompt can make LLMs explain their reasoning step by step instead of giving a direct answer, which improves clarity and consistency. Context faithful prompting helps models focus on relevant information rather than repeating memorised facts, as shown in work by Zhou et al. (2023). Structured and multi-round prompting helps models handle ambiguity and misleading information by encouraging more deliberate reasoning instead of relying on surface patterns (Pan et al., 2024). This makes prompting a powerful tool for guiding LLM reasoning without retraining.

Several studies have examined how prompting affects reasoning tasks involving logic and causality. Pan et al. (2024) tested different prompting techniques for fallacy classification and found that both single-round and multi-round prompts improved results. The best improvements came when prompts included short definitions or step by step reasoning, which helped models apply reasoning more consistently across tasks. Zhou et al. (2023) explored prompting methods designed to improve context faithfulness in reasoning tasks. They found that prompts focused on opinions and instructions led models to rely less on memorised knowledge and to improve their reasoning when faced with conflicting or counterfactual cases. Kong et al. (2024) used role-play prompting to improve zero-shot reasoning, showing that assigning the model a clear role or perspective enhanced coherence and accuracy in its explanations. Shanahan et al. (2023) provide a conceptual foundation for this approach, arguing that role prompts can shape how LLMs simulate reasoning by framing their behaviour as role-play rather than understanding. These studies suggest that prompting can act as a form of reasoning control. It encourages models to reason more carefully about context, structure and causal relations.

In this thesis, Chain-of-Thought (CoT) refers specifically to a prompting technique that guides the model to consider a problem step by step before replying. While CoT prompting is widely used to support more deliberate reasoning by making the model use intermediate steps, several studies show that step by step generation does not always lead to deeper reasoning. Sahoo et al. (2025) note that CoT can improve the structure in model outputs in a range of reasoning tasks. In a recent empirical study evaluating LLM behaviour, Moore et al. (2025) find that CoT does not consistently mitigate initial biases. This means the additional reasoning steps could simply elaborate on an initial bias rather than challenge it. Their analysis suggests that CoT tends to think fast unless the prompt explicitly encourages the model to explore alternative interpretations. These findings show that CoT is a good but fragile prompting technique. It can enhance reasoning but also reinforce surface-level heuristics if the model is not properly instructed.

This experiment uses several prompting styles to guide the model in different ways. Zero-shot prompting is used as a baseline and compared with styles designed to support clearer reasoning, such as CoT, role-based, and counterfactual prompts. A prompt written to push the model toward an incorrect answer is also included to test whether the model can resist biased instruction. These prompting styles were chosen because previous studies show that prompting can shift how models approach reasoning tasks. They provide a basis for testing how different forms of guidance influence the causal fallacies examined in this study.

Prompting Style	Description
Zero-Shot	Provides task description to the LLM without any examples on how to perform the task. The model relies solely on its pre-training knowledge to generate responses. It is useful for evaluating how well a model generalises to new tasks with minimal guidance (Sahoo et al., 2025).
Chain-of-Thought	Guides the model to generate intermediate reasoning steps before arriving at a final answer. It enhances interpretability and can improve accuracy on complex reasoning tasks (Sahoo et al., 2025).
Role-Based	A technique where the LLM is assigned a role in the initial prompt before the dialogue with the user. For example it could be asked to portray an expert in a specified field. This guides the model to generate responses consistent with its assigned role (Shanahan et al., 2023).
Counterfactual	Asks the model to reason about what would happen if a part of the situation was different. It introduces a small change in the claim and the model is prompted to judge the outcome under the altered scenario. This helps reveal if the model can see beyond surface patterns and consider how the relation between event changes in a hypothetical case.
Cautious	Introduces a framing where the model is encouraged to avoid false positives when judging fallacies. The instruction states that labeling a valid argument incorrectly as fallacious is more harmful. This allows the experiment to see if the model's detection is sensitive to shifts in perceived cost and risk.

Table 2.2: Overview of prompting styles used in the experiment.

2.6 Challenges in Evaluation and Research Gap

Evaluating how LLMs handle logical fallacies, especially causal fallacies, raises practical and conceptual challenges. Existing benchmarks give useful indications of performance while also highlighting open challenges in measuring fallacy competence.

2.6.1 Limitations of Current Evaluation Methods

Several studies evaluate fallacy detection as a classification task. Work such as Pan et al.'s (2024) zero-shot fallacy classifier study and Jin et al.'s (2022) logical fallacy detection study provide short arguments and ask models to assign a fallacy label. These datasets cover a wide range of informal fallacies and report performance for each type, giving

a clear overview of how models behave across categories. However, the categories are broad. In both studies, causal fallacies are grouped under a single label, false causality, which combines several types of causal errors. Because of this, it is difficult to see how models perform on the specific causal fallacies within the broader category, such as post hoc, cum hoc, and slippery slope.

The datasets used in these classification studies also have constraints. Pan et al. (2024) and Jin et al. (2022) both note that some fallacies appear more frequently in their datasets while others are rare. This reflects how arguments occur in real text and affects evaluation, as models tend to perform more reliably on common categories than on those with fewer examples. Both studies report performance for each fallacy category and an overall score. The per-type results are informative, but the overall score is affected by the dataset distribution. Fallacy types that appear more often contribute more to the combined score than the rare ones. Therefore, the overall score gives only a broad impression of performance and should be read alongside the per-type results.

Other studies explore fallacies in interactive settings rather than through classification. Payandeh et al. (2024) use a debate format in which one model argues and another evaluates the argument. This gives insight into how models react to persuasive or misleading reasoning, but also introduces new sources of variation since the outcome depends on the prompts used and how the models respond to each other. This makes it difficult to isolate fallacy recognition itself, because differences in debate behaviour can be caused by the dialogue setup rather than the model's ability to judge reasoning. Therefore, these studies are informative but cannot be used as a clear measure of fallacy detection. However, the results can complement single-turn classification studies.

There is also research on causal reasoning that is related but not directly comparable. Several studies examine how models judge whether one event causes another using short, structured descriptions (Zečević et al., 2023) or synthetic stories (Joshi et al., 2024). These tasks help isolate specific causal abilities, such as judging temporal order or selecting the most likely cause, but do not ask the model to evaluate whether a line of reasoning is fallacious. Therefore, the results say something about causal inference in controlled settings but not about the informal causal fallacies that appear naturally in text.

Taken together, existing work shows that language models can recognise many kinds of fallacies and that several useful benchmarks already exist. The design choices in these studies also leave open questions. The broad labels used in classification studies provide wide coverage but do not separate the individual causal fallacies within the false causality

category. Interactive debate studies test different abilities and are not directly comparable to single-turn detection. Causal reasoning benchmarks focus on formal causal inference rather than the reasoning patterns found in informal arguments. As a result, we still know little about how LLMs handle specific causal fallacies or how different prompting styles influence this behaviour.

2.6.2 Conceptual and Methodological Challenges

There are also challenges arising from the nature of informal fallacies and their definition. The categories are not always clear, and different sources describe them slightly differently. Hurley (2012) notes that many informal fallacies overlap and can be difficult to separate in practice. This makes annotation uncertain, because an argument that matches one definition can also fit another. This matters for fallacy detection because each claim must be placed in the correct category for the results to reflect genuine differences between the types. If a claim borders on more than one fallacy, a disagreement between the model and the label may come from this ambiguity rather than a clear reasoning error.

Another challenge concerns the role of background knowledge. Language models draw on extensive information from pre-training. Several studies show that models often rely on this stored knowledge even when the task does not require it (Zečević et al., 2023; Joshi et al., 2024). In fallacy detection, this can cause a model to judge a claim based on whether it believes the events are connected in the real world, rather than focusing on the reasoning in the argument. It becomes difficult to tell whether a wrong answer reflects confusion about the causal structure or an incorrect factual assumption from the training data.

Prompting also influences evaluation. Small changes in how the task is phrased can shift how the model approaches the same argument. Studies show that different instructions can lead the model to adopt different reasoning strategies and produce different types of answers (Kong et al., 2024; Zhou et al., 2023). This affects evaluation because the results may reflect the prompt rather than the model's underlying ability. It also raises questions about which prompt is most appropriate when comparing performance across fallacy types.

Together, these challenges show that evaluating causal fallacies is not just a matter of having the right dataset. The definitions are flexible, the model brings prior knowledge

into the task, and the results depend on how the instructions are phrased. These factors should be considered when interpreting fallacy detection performance, especially when comparing different prompting styles.

2.6.3 Research Gap

Previous work has given a broad picture of how LLMs handle informal fallacies, but several gaps remain. Broad classification studies treat causal fallacies as one category and do not separate the individual fallacy types within it. Debate studies and causal inference benchmarks provide useful insight, but do not evaluate fallacy detection in natural arguments. There is also limited research on how different prompting styles influence behaviour across specific causal fallacies.

To the best of my knowledge, there is no study that combines the following design choices:

- A focused evaluation of specific causal fallacy types, together with matched non-fallacious causal arguments.
- Keeping the model fixed and varying only the prompting style in order to isolate the effect of the prompt.
- Analysing differences in accuracy across fallacy types, including patterns in false positives and false negatives, rather than only reporting a single overall score.
- A qualitative analysis of the model's explanations, examining how its justifications align or fail to align with the causal structure of the argument.

This experiment is designed to address these gaps. By separating the causal fallacies into specific types and applying several prompting styles to the same model, it is possible to see which fallacies are most challenging, which prompts change the model's behaviour and how these factors interact. The results provide a more detailed view of fallacy detection in causal arguments than is available in existing studies.

3 Methodology

3.1 Experiment Design

3.1.1 Aim of the Experiment

The aim of this study is to see how well an LLM handles claims with causal fallacies. The experiment tests whether the model can detect faulty causal reasoning and whether it accepts such claims as valid. It also examines whether performance changes with different prompting styles. The model is tested on both fallacious and non-fallacious claims to compare its behaviour and identify where it succeeds and fails.

3.1.2 Experimental Setup

The experiment is set up as a detection task. The model is asked to determine whether each claim's causal reasoning is fallacious and to provide a brief justification. The same set of fallacious and non-fallacious data is used for every run. Only the prompting style changes between conditions. All other factors remain fixed, including the claims, their order, the model version, and the model settings. Each claim is given to the model in isolation with a required output format. This allows answers to be compared across prompting styles.

The dataset and Python scripts used to run the experiment are available at: `https://github.com/mariannlepsoy/causal_fallacies`

3.1.3 Model

All experiments used GPT-4.1 mini. This model was chosen because it responds quickly, follows instructions reliably, and does not include a built-in reasoning step found in dedicated reasoning models (OpenAI, 2025). This made it suitable for comparing standard prompting with explicit CoT prompting. The same model version and settings were used in every run. The temperature was set to zero so that the outputs were deterministic, allowing differences in performance to be attributed to the prompt rather than the model.

3.1.4 Evaluation Procedure

Each claim was given to the model with one of the prompting styles. The model returned a yes-or-no judgement and an explanation. The output was scored by checking whether the answer matched the correct label for that claim. Every claim was evaluated once for each prompting style. The results were used to calculate accuracy for each prompt and fallacy type. Only the binary judgement was used to calculate accuracy.

The explanations were used for a qualitative analysis in the discussion chapter, providing insight into how the model justifies its answers and how prompting styles influence these justifications. Since the explanations did not affect correctness, they were not further processed for the quantitative results.

3.2 Dataset and Prompt Design

3.2.1 Dataset

Many of the fallacious claims used in the experiment were taken from the LOGIC dataset (Jin et al., 2022), which contains examples of common logical fallacies. From this dataset, only the entries that they labelled as false causality were selected. To complement these, additional fallacious claims were gathered manually from publicly accessible websites. This approach follows the procedure used by Jin et al. (2022) to constructe the LOGIC dataset, which included examples from student quiz websites and manually collected online sources.

All data used in the final dataset were already labelled as fallacious by their original authors, and most entries were already assigned to a causal fallacy category before being used in this study. In the few cases where a claim was not labelled with a specific causal fallacy type and could plausibly fit more than one causal fallacy category, it was assigned to the category that best matched the causal error in the text. This study follows the practice used in earlier work, where the evaluation is based on pre-labelled fallacy data rather than new annotation. In particular, Pan et al. (2024) also rely on pre-labelled examples when evaluating models on fallacy classification tasks.

A separate process was used to collect comparable non-fallacious data. These were taken from the e-CARE dataset (Du et al., 2022), which contains causal statements with a premise, an instruction asking for cause or effect, and two hypotheses where only one is correct. The correct hypothesis and premise together form a valid causal relation. One hundred entries were selected that matched the length and style of the fallacious entries and did not show signs of fallacious reasoning. For each entry, the premise and correct hypothesis were combined into a single statement without changing the wording. When the instructions asked for an effect, the claim was written as premise followed by hypothesis. When it asked for a cause, the claim was written as hypothesis followed by premise.

The complete evaluation dataset includes the selected entries from the LOGIC dataset, the manually collected fallacious examples, and the non-fallacious data from the e-CARE dataset. The distribution of claims across fallacy types is shown in Table 3.1.

The aim of this thesis is not to reassess whether each claim is universally agreed to be fallacious. Some examples can be debated, and different authors may categorise the same argument differently. A detailed conceptual analysis of each argument is outside the scope of this study. Instead, this study follows the labels and categorisation established in previous work and used in fallacy detection literature.

Fallacy Type	Count	Percentage (%)
Slippery Slope	107	20.4
Post Hoc	87	16.6
Cum Hoc	78	14.9
Gambler's Fallacy	31	5.9
Neglecting a Common Cause	27	5.2
Causal Oversimplification	27	5.2
Reverse Causation	23	4.4
Regression Fallacy	23	4.4
Confusing Necessary/Sufficient	20	3.8
No Fallacy	100	19.1
Total	**523**	**100.0**

Table 3.1: Distribution of claims across fallacy types in the dataset.

3.2.2 Prompt Design

The experiment evaluates how different prompting styles affect the model's ability to recognise causal fallacies. Each prompting style was implemented directly in Python through the script used to run the model. The prompts were constructed to represent different levels of guidance, reasoning support, and bias. The aim was to observe how sensitive the model is to changes in framing and instruction detail.

The model received the same core task across all prompts: decide if the claim's causal reasoning is fallacious. The output was constrained to a JSON object with a binary judgement ("yes" or "no") and a short explanation of up to 50 words.

Instruction constraint used in all prompts

Input

Judge the reasoning only from the text of the claim. Do not use outside knowledge and do not fact check.

This constraint was added because the task is fallacy detection and not fact verification. Causal fallacies are informal fallacies, meaning they can be identified by evaluating whether the content provides sufficient evidential support for the claimed causal link (Hurley, 2012). Therefore, the factual truth of the statement is unimportant in this experiment. An argument could be factually correct but still commit a fallacy if the reasoning is weak. Without this rule, the model could classify a statement as non-fallacious simply because it believes the events are connected in real life. The aim is to have the

model judge whether the argument as written contains a causal fallacy, not whether the model thinks the conclusion matches the real world. Therefore, this rule drives the model to behave similarly to a human evaluating the question: *Is the reasoning fallacious, regardless of whether you personally believe the claim to be true?*

Zero-Shot

Input

Decide if the claim's causal reasoning is fallacious.

The zero-shot prompt serves as a baseline condition. It contains only the task description and rules without additional context or guidance. This allows the experiment to measure how well the model can detect fallacies without hints or examples. The expectation was that this would reflect the model's default behaviour.

Chain-of-Thought

Input

You are assessing whether a claim's causal reasoning is fallacious. Silently think step by step before deciding.

The instruction to think step by step encourages the model to use the internal reasoning mechanism associated with CoT prompting. The goal was not to analyse the internal reasoning itself, but to determine whether prompting the model to think silently increases the accuracy of fallacy detection. This prompt tests if causal fallacy detection benefits from deeper internal reasoning.

Role-Based

Input

You are an expert in causal reasoning and informal logic. You specialise in identifying errors in causal arguments. You only judge the reasoning in the claim, not whether the conclusion is correct in the real world.
Decide if the claim's causal reasoning is fallacious.

Role-based prompting was implemented by assigning the LLM the role of an expert in causal reasoning and informal logic. The purpose was to test whether adopting an expert role would lead the model to apply stricter standards when evaluating causal statements. This choice is supported by findings that models tend to adapt their behaviour to the role described in the prompt, producing responses aligned with the expectations associated with that role (Shanahan et al., 2023).

Counterfactual

Input

First identify the key event in the claim and then identify what the claim presents as the effect. Imagine a version of the situation where the key event is changed or removed. Consider if the effect in the claim would still follow in that imagined version. Then decide if the claim's causal reasoning is fallacious.

This prompt instructs the model to consider whether the claimed effect would still follow if the key event is changed or removed. Several causal fallacies, such as post hoc or neglecting a common cause, can be exposed by counterfactual thinking. Asking the model to perform this check tests whether counterfactual thinking helps it detect faulty reasoning, that might otherwise appeared plausible.

Cautious

Input

In this evaluation, wrongly marking a claim as fallacious is considered more harmful than overlooking a minor reasoning flaw. Be careful about labeling a claim as fallacious unless the causal reasoning error is clear.
Decide if the claim's causal reasoning is fallacious.

In the cautious prompt, the model was told that false positives (incorrectly labelling a valid argument as fallacious) are more harmful than false negatives. This prompt was included to see if the model changes how it judges arguments when instructed to avoid false positives. The task and data remain the same, so the model must still detect the same type of reasoning errors as in the other prompts. Comparing the outputs of this prompt with the others shows whether balancing carefulness with accuracy affects results.

3.3 Data Analysis

All model outputs were returned in a fixed JSON format with two fields: fallacious and explanation. When evaluating correctness, only the fallacious field was used. Since this field always contained an explicit 'yes' or 'no', there was no need for further interpretation of the model's answer.

```
{
    "fallacious": "yes",
    "explanation": "The claim assumes that..."
}
```

For fallacious claims, a response was counted as correct when the model returned *fallacious: yes*. For non-fallacious claims, a response was correct when the model returned *fallacious: no*. These values were used to compute accuracy separately for each fallacy category, each prompting style, and overall. Accuracy was calculated as the proportion of correct answers out of the total number of claims in that category. This produced the accuracy scores shown in Table 4.1.

4 Results

4.1 Overview of Results

The evaluation results are summarised in Table 4.1, which shows accuracy for each fallacy type across the five prompting styles. The table also includes the overall accuracy for each fallacy and prompting style, with accuracy on non-fallacious claims shown separately at the end. The values indicate that performance varies between prompting styles and fallacy types. Some fallacies are consistently easier for the model to detect, while others show lower accuracy across prompts. Prompting styles also do not perform uniformly, with noticeable differences in overall accuracy. This chapter provides an overview of the numerical results, while a more detailed interpretation appears in the next chapter.

Fallacy Type	**Counterfactual**	**Role-based**	**Zero-Shot**	**CoT**	**Cautious**	**Overall**
Neglecting Common Cause	0.96	0.96	0.96	0.96	0.85	**0.94**
Post Hoc	0.95	0.95	0.95	0.93	0.90	**0.94**
Slippery Slope	0.97	0.95	0.94	0.94	0.91	**0.94**
Gambler's Fallacy	0.97	0.90	0.97	0.90	0.77	**0.90**
Cum Hoc	0.92	0.88	0.88	0.83	0.78	**0.86**
Reverse Causation	0.83	0.74	0.83	0.78	0.74	**0.78**
Confusing Necessary/Sufficient	0.95	0.85	0.75	0.65	0.50	**0.74**
Causal Oversimplification	0.85	0.74	0.63	0.63	0.59	**0.69**
Regression Fallacy	0.87	0.74	0.70	0.61	0.43	**0.67**
Overall	**0.94**	**0.90**	**0.89**	**0.86**	**0.79**	**0.83**
No fallacy	0.80	0.75	0.84	0.83	0.89	0.82

Table 4.1: Accuracy of each prompting style across fallacy types.

4.2 Performance across Prompting Styles

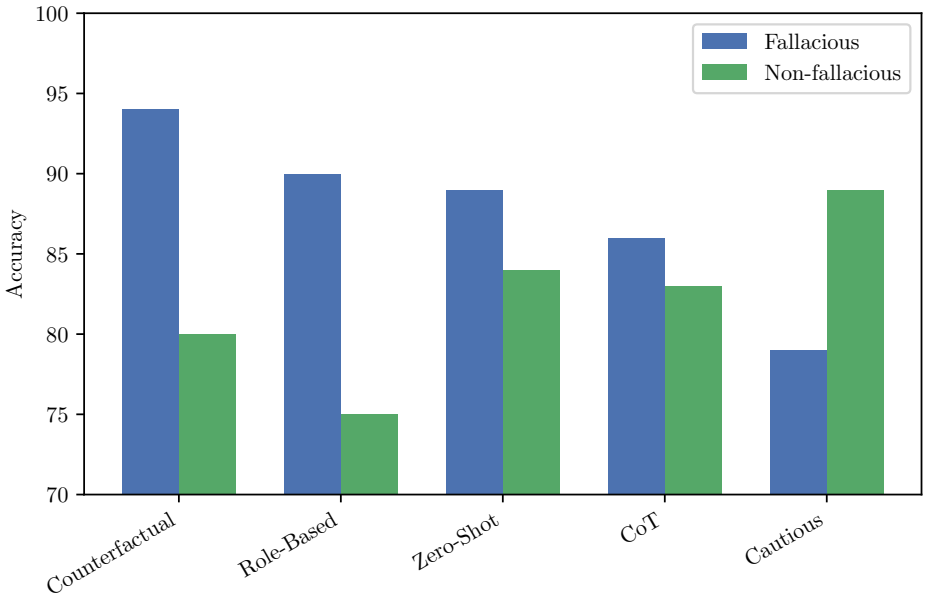

Figure 4.1: Comparison of overall accuracy on fallacious and non-fallacious claims for each prompting style.

There are clear differences in how well the model identifies fallacious arguments across the five prompting styles. Counterfactual achieves the highest overall accuracy, reaching 94 percent on fallacious claims. It also shows consistently high accuracy across fallacy types with less variation than the other styles. Role-based and zero-shot follow with 90 percent and 89 percent accuracy, respectively. CoT does not improve detection, and reduces overall accuracy by 3 percent compared to zero-shot. The lowest overall accuracy for fallacious claims is 79 percent with the cautious prompt.

The ranking differs for non-fallacious claims. Notably, the cautious prompt, which performed worst on fallacious claims, performs best on non-fallacious claims with 89 percent accuracy. The opposite is true for role-based and counterfactual, which were best for fallacious claims but worst for non-fallacious claims, with accuracies of 75 percent and 80 percent. Across most prompting styles, accuracy is lower for non-fallacious inputs than for fallacious ones, with the cautious prompt being the exception.

Overall, the prompting styles differ in effectiveness and in how consistently they detect fallacious reasoning across fallacious and non-fallacious arguments.

4.3 Performance across Fallacy Types

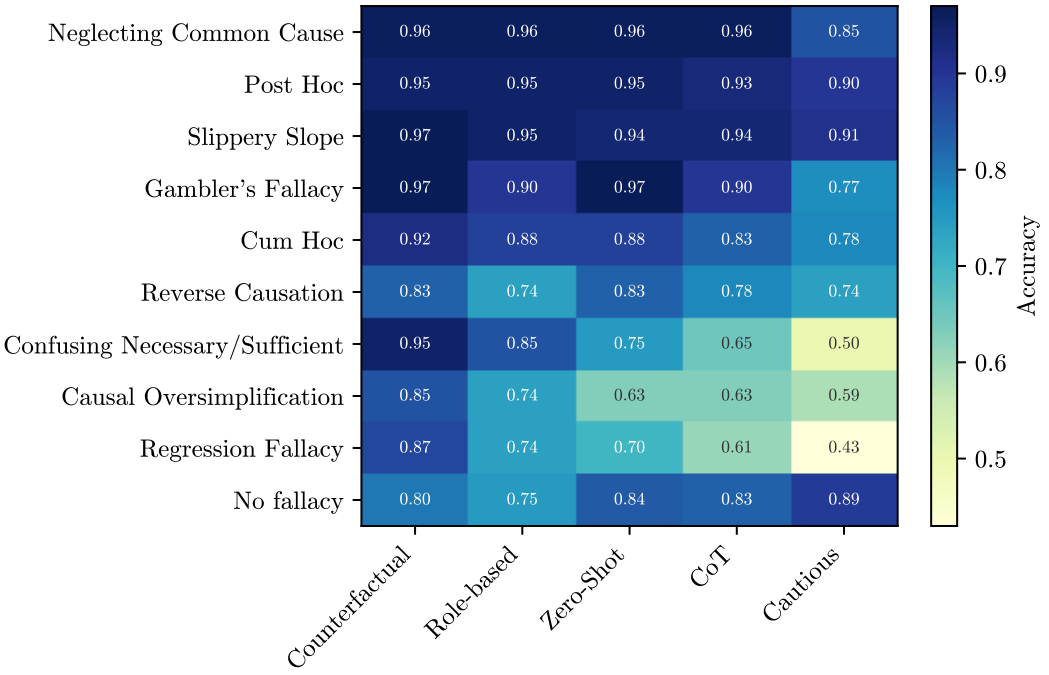

Figure 4.2: Heatmap showing accuracy for each prompting style across all fallacy types.

Accuracy varies considerably between fallacy types across prompting styles, as seen in Figure 4.2. The fallacy types with the highest overall accuracy are neglecting a common cause, post hoc, and slippery slope, each with 94 percent. These fallacies show little variation between prompts, with most prompting styles reaching scores above 90 percent. They are the best-performing categories in the experiment.

Several fallacies show lower performance. The most difficult categories for the LLM to detect are causal oversimplification, regression fallacy, and confusing necessary with sufficient cause, all with overall accuracy below 80 percent. These fallacy types also display larger differences between prompting styles, with the cautious prompt producing the lowest scores. Reverse causation also shows lower overall accuracy, though it varies more evenly across prompts.

The heatmap shows that prompting styles influence fallacy type performance. For example, counterfactual prompting produces consistently high scores across all fallacies, while CoT and cautious show greater variation. Some fallacies that are highly accurate under one prompt drop considerably under others. The confusing necessary with sufficient fallacy reaches 95 percent accuracy with the counterfactual prompt but only 50 percent

with the cautious prompt. The non-fallacious category achieves its highest accuracy under the cautious prompt, a different pattern from the fallacious categories.

Overall, the results show that the model's reliability depends on the specific fallacy category and the prompting style. Certain fallacies are consistently easier for the model to detect, while others present challenges across prompts.

5 Discussion

5.1 Overview

This chapter discusses the experiment's findings and their broader implications. The aim is to interpret the numerical results from the results chapter and explore potential explanations for the patterns. Some findings require more careful interpretation, especially regarding differences between prompting styles. The discussion focuses on how prompting shapes model behaviour. It examines differences between prompting styles, variation across fallacy types, and a qualitative analysis of the model's explanations. Finally, it outlines methodological limitations and suggestions for future work.

5.2 Interpreting the Prompting Style Differences

The experiment showed that prompting style influences how the model judges causal arguments. Some prompts encourage closer attention to the claim's structure, while others appear to reinforce patterns learned during pre-training. The following section discusses how these differences may have shaped the results.

5.2.1 Why Counterfactual and Role-Based Improve Detection

Counterfactual prompting produced the highest overall accuracy in the experiment. A likely explanation is that the counterfactual instruction guides the model to focus more on the relationship between events in the claim. When a prompt instructs the model to consider what could happen if the situation were different, it may become less reliant on familiar patterns from pre-training. This interpretation aligns with the study

by Zhou et al. (2023) on context-faithful prompting, which shows that carefully designed prompts can make LLMs rely more on local context and less on memorised world knowledge. Their experiment shows that counterfactual demonstrations encourage closer attention to the claim itself.

Counterfactual language is closely connected to how causal relations are expressed in natural language. Joshi et al. (2024) note that people often use counterfactual statements to convey causal judgements. An example is the form *If X had not happened, Y would not have occurred*, which naturally leads to a causal interpretation. This relationship between counterfactual framing and causal reasoning helps explain why the counterfactual prompt performs well. The instruction encourages the model's attention to how the events relate to one another.

Role-based prompting also performs strongly in the results, and several factors can explain why assigning a role performs better than zero-shot and CoT. Kong et al. (2024) show that assigning a model a specific role can stabilise its behaviour by providing a focused perspective. This often leads to clearer reasoning and more reliable use of the input information. Their study shows that when given a defined role, models tend to generate more structured and coherent explanations, which is likely beneficial in a task that requires attention to the argument. Shanahan et al. (2023) make a similar point, noting that being given a role guides the stance the model takes in its reasoning.

Viewed in light of previous research, the improved performance of the role-based prompt may stem from its ability to narrow the model's attention and encourage a more careful style of reasoning than zero-shot or CoT. Zero-shot prompts leave the model free to rely on habits learned during training, while CoT encourages longer explanations without guaranteeing that the underlying reasoning improves. A role-based prompt offers a middle ground providing a direction without forcing the model into unnecessary details. This could explain why the prompt performs consistently well across fallacy types.

These prompts do not add more reasoning steps. Instead, they work because they constrain how the model attends to causal structure.

5.2.2 Why Chain-of-Thought Underperforms Zero-Shot

Unlike the counterfactual and role-based prompts, CoT did not improve performance compared to zero-shot. This may seem surprising since CoT is often assumed to encourage

deeper reasoning. However, more internal reasoning steps do not necessarily improve causal reasoning. Studies show that when models are asked to *think step by step*, they often generate detailed text without examining the argument's structure in depth. Moore et al. (2025) describe this behaviour in their study, where CoT outputs tend to follow the model's initial judgement rather than encouraging deeper reasoning. They also note that step by step reasoning can repeat shortcuts learned in training rather than challenging them. This can result in confident but wrong judgements when the initial judgement is misleading.

Another limitation of CoT prompting is that it does not verify whether the reasoning steps it generates are correct. CoT can produce detailed and coherent explanations, but these steps are not checked against the reasoning demands of the task, since CoT lacks a built-in validation mechanism (Sahoo et al., 2025). In fallacy detection, the model may elaborate the argument rather than examine whether the causal claim is justified. As a result, it can reinforce a misleading interpretation rather than correct it, which contributes to lower accuracy than zero-shot.

5.2.3 Why Cautious Prompt Improves Non-Fallacious Accuracy but Hurts Fallacious Accuracy

The cautious prompt produced the highest accuracy on non-fallacious arguments and the lowest on fallacious ones. This pattern suggests that the instruction shifts how the model weighs uncertainty. The prompt encourages a more conservative approach, asking the model to avoid marking valid arguments as fallacious. The model becomes more likely to treat the argument as valid when unsure rather than risk detecting a fallacy that is not present. This behaviour increases accuracy on non-fallacious claims but also leads to many missed fallacies.

Research by Li et al. (2024) indicates that many LLMs struggle to recognise logically correct statements. In their evaluation, models incorrectly marked several valid sentences as fallacies, which suggests an inclination to over-identify fallacious reasoning even in valid arguments. A cautious prompt counteracts this tendency and encourages the model to judge non-fallacious arguments more carefully. However, this change in behaviour also reduces the model's sensitivity to actual fallacies. These results suggest that emphasising the cost of a mistake makes the model more hesitant to commit to firm conclusions when reasoning is uncertain. The cautious prompt reduces false positives but increases false negatives, which explains its uneven performance across fallacious and non-fallacious arguments.

5.3 Interpreting Patterns Across Fallacy Types

The results show clear differences in how the model responds to various fallacy types. Some are identified with high accuracy across most prompting styles, while others remain difficult regardless of the prompt. These differences offer insight into what kinds of causal mistakes the model finds easier to recognise and which it cannot reliably recognise.

The fallacy types post hoc, slippery slope and neglecting a common cause achieve high accuracy for nearly all prompts. One reason is that these fallacies often contain surface patterns that are easy to recognise. They usually present a simple link between two events or a temporal sequence, which provides strong cues for the model. This interpretation is consistent with findings from fallacy detection research. Jin et al. (2022) observe that models perform better on fallacies that are more distinct and frequently encountered, because these patterns tend to have been seen many times during pre-training. Although their work concerns logical fallacies more broadly, the same idea applies to causal fallacies with an easily recognisable structure.

Lower accuracy appears when fallacy detection depends on information not explicitly stated. Categories such as regression fallacy, causal oversimplification, and confusing necessary and sufficient conditions are harder to detect across prompting styles because they require the model to reason about factors not directly mentioned in the claim. This is an area where LLMs often struggle. Zečević et al. (2023) argue that LLMs rely on learned associations between causal statements and tend to repeat familiar causal narratives rather than reasoning about the underlying causal structure. Therefore, models may give plausible explanations while overlooking missing or unobserved causes. This helps explain why fallacies that depend on hidden structures or multiple causes result in lower accuracy.

The contrast between neglecting a common cause and causal oversimplification helps clarify this limitation. Both fallacies involve missing information, yet the model performs much better on neglecting a common cause. In common cause cases, the missing factor often follows a familiar pattern, such as two events being linked through a third variable. This pattern appears frequently in text and educational material, which likely makes it easier for the model to detect. Oversimplification, on the other hand, requires recognising that there might be several relevant causes and that the argument reduces the explanation to only one. This is a more abstract judgement because the model must recognise that the argument is incomplete without knowing exactly which additional causes are missing.

The weaker performance in this category aligns with prior work showing that language models struggle when causal reasoning depends on complex structures rather than simple observational patterns.

The non-fallacious category shows a different trend compared to the fallacious categories. Accuracy is lower across prompting styles, and the ranking of prompt performance is different. This suggests that confirming that an argument is valid is more challenging than detecting a mistake. Li et al. (2024) report a similar pattern where many models incorrectly mark valid statements as fallacious when the task involves fallacy detection. This aligns with the consistently lower scores seen for the non-fallacious claims.

Together, these observations indicate that fallacies with clear, familiar structures are easier for LLMs to recognise, whereas those that rely on multiple interacting causes or hidden information remain challenging. The differences across categories highlight the limitations of current models in handling complex forms of causal reasoning.

5.4 Biases and Systematic Tendencies

The results reveal several systematic tendencies in how the model evaluates causal arguments. One consistent pattern is a general bias toward identifying fallacies. The model often treats an argument as suspicious when the task is framed around detecting errors. This aligns with previous findings showing that language models often classify correct statements as fallacious when tasked with detecting fallacies (Li et al., 2024). This tendency is visible in this experiment, where accuracy on non-fallacious claims is lower than for fallacious claims. This indicates that the model often treats valid arguments as if they contained an error.

In addition to this bias, the model shows a tendency to favour causal explanations that match familiar narrative forms. Carro et al. (2024) show that language models sometimes judge two events as causally related even when there is no statistical contingency between them. This behaviour is consistent with an illusion of causality, where the model responds to how coherent the story appears rather than to the actual support for the causal link. A similar tendency is visible in the experiment. Several incorrect detections show that the model sometimes accepts causal claims that seem plausible on the surface, even when the arguments do not provide sufficient support for the conclusion.

These behaviours shift depending on how the task is framed. The cautious prompt reduces the general bias toward finding fallacies, while the CoT and zero-shot prompts maintain it to a greater extent. This shows that the model does not follow a single strategy. Instead, its approach depends on how the task is presented. The pattern suggests that part of the model's behaviour is driven by expectations generated by the prompt, rather than solely by the structure of the argument.

These tendencies reflect broader limitations in LLMs. When judging causal arguments, the model balances cues from the prompt, familiar narrative patterns and its own biases in ways that do not always correspond to the structure of the reasoning. Understanding these behaviours helps explain why the model performs well on some fallacies and struggles with others, and highlights the need for approaches that encourage more consistent causal reasoning.

5.5 Analysis of Model Explanations

In addition to the binary fallacy judgement, the model was asked to produce a short explanation for each claim. The same claims were used across prompting styles, and the model generated separate explanations for each prompt. The analysis of the explanations examines two things: how the model justifies its decisions on fallacious and non-fallacious claims, and how the explanations change when different prompting styles give different answers to the same claim. This way, the explanations can be compared across prompts and across correct and incorrect detections.

5.5.1 Behaviour on Non-Fallacious Claims

The non-fallacious dataset provides evidence that explanations adapt to the model's fallacious judgement rather than to the structure of the claim. The explanations are clear and uncontroversial when the model correctly sees that a claim is valid. When the model incorrectly detects a fallacy, the explanations shift to align with the fallacy judgement rather than the claim itself. This behaviour can be seen across prompts.

Claim (No Fallacy)
A cow on the farm is pregnant. Soon the cow will have offsprings.

Zero-shot ✓
The claim correctly infers that a pregnant cow will soon have offspring, which is a reasonable causal relationship without error.

Role ✓
The claim correctly infers that a pregnant cow will soon have offspring, which is a valid causal reasoning based on the biological process of pregnancy leading to birth.

Claim (No Fallacy)
In the summer, the surface of his car is not covered with glass films yet. The glass windows made the heat on the car even more unbearable.

Zero-shot ✗
The claim assumes that 'The glass windows made the heat on the car even more unbearable' without considering other factors, implying a causal link without sufficient evidence, which is a causal fallacy.

Cautious ✗
The claim incorrectly implies that glass windows increase heat inside the car, but 'The glass windows made the heat on the car even more unbearable' is a causal error, confusing correlation with causation without clear reasoning.

Figure 5.1: Comparison of a correctly and an incorrectly detected non-fallacious claim, with model explanations.

An example of a claim the model got correct across all prompting styles is the claim about the pregnant cow (see Figure 5.1). All prompting styles answered that the claim was non-fallacious, and all explanations agreed that the claim described a normal biological process. The explanations were similar, noting that pregnancy leads to offspring and providing no further analysis. This is typical for correct outputs. When the model does not detect a fallacy, it provides a brief, neutral justification that closely matches the wording of the claim. The explanations are mostly paraphrases and do not introduce interpretation beyond the claim's clear meaning.

The pattern changes in cases the model incorrectly labels as fallacious. This can be seen with the claim about heat inside a car during summer (see Figure 5.1). The claim describes a normal heat buildup effect, but even though the dataset labels it as non-fallacious, every prompting style answered that it was fallacious. A couple of the explanations, zero-shot and counterfactual, state that the claim does not consider other factors, but the claim never suggests that the glass windows are the only cause of the heat. It merely says that the windows made the heat "even more unbearable". The model still interpreted this as a faulty causal jump.

Claim (No Fallacy)
The witness considerated for two minutes before answering the judge's question in court. The judge waited for two minutes.

Zero-shot ✓
The claim states a sequence of events without asserting a causal relationship; it does not claim the judge waited because the witness considered for two minutes.

CoT ✗
The claim implies causation without evidence: 'The judge waited for two minutes' is presented as caused by the witness's consideration, but this is merely correlation, not proven causation.

Figure 5.2: Example of a non-fallacious claim where prompts disagree.

Several claims show disagreement: one prompting style answers yes, and others answer no. When a prompt incorrectly marks a claim as fallacious, it constructs a fallacy narrative around a phrase that the model treats as a neutral description or valid causal reasoning with other prompts. Figure 5.2 illustrates this with the claim about the witness and the judge. Four prompting styles, including zero-shot, treat the claim as a neutral sequence of events, while CoT treats the same claim as an unsupported causal claim. The explanations, therefore, adapt to the binary decision rather than to the structure of the claim, showing how easily the model's reasoning can shift with a different prompt.

A consistent pattern is that the explanations follow the model's judgement. When the answer is correct, the explanations remain close to the text. When it is incorrect, the explanation reshapes the claim to fit a fallacy the model believes is present. This is most clearly seen in disagreement between prompting styles. The same claim can be treated as either harmless or fallacious depending on the prompt, and the explanation adjusts to justify the chosen label. This suggests that the explanations do not provide independent insight into the model's reasoning. Instead, they reflect how the prompt frames the task and how the model commits to a decision.

5.5.2 Behaviour on Fallacious Claims

Claim (Post Hoc)
The party was going great! Then Gerald showed up and little by little people started collecting their coats and heading home. Gerald ruined the party.

Zero-shot ✓
The claim commits post hoc ergo propter hoc by assuming 'Gerald showed up' caused people to leave, as indicated by 'Then Gerald showed up and little by little people started collecting their coats and heading home.'

Cautious ✓
The claim assumes Gerald caused the party to end without evidence, shown in 'Gerald ruined the party.' This is a post hoc fallacy, attributing causation solely because one event followed another.

Claim (Regression Fallacy)
I had a real bad headache, then saw my doctor. Just by talking ... all better the next day. It was well worth the $200 visit fee.

Zero-shot ✗
The claim suggests a causal link between talking with the doctor and headache relief without asserting a fallacy; it is a plausible cause-effect statement based on the sequence of events described.

Counterfactual ✗
The claim suggests talking with the doctor caused the headache to subside, which is plausible as a therapeutic effect. No clear error in causal reasoning is present in the claim's context.

Figure 5.3: Comparison of a correctly and an incorrectly detected fallacious claim, with model explanations.

In contrast to the non-fallacious cases, explanations for fallacious claims follow a different but equally consistent pattern. When the model correctly identifies a fallacy, the explanations usually point to the right part of the claim and name the general type of mistake. They often highlight temporal order, missing causes or an invalid link between events. Even though the explanations are brief, they often capture why the claim is considered fallacious and align with the fallacious label. This pattern is clear in the post hoc example shown in Figure 5.3, where all prompts detect the fallacy and the explanations highlight the flawed reasoning.

When the model fails to detect the fallacy, the explanations present the claim as valid rather than faulty. In Figure 5.3, the regression fallacy claim of the headache and doctor's visit shows this. Instead of questioning the reasoning step that makes the claim fallacious, it reframes the claim as an acceptable causal statement. For example, claims that rely on temporal coincidence are considered valid cause-and-effect descriptions, and claims that omit key factors are considered reasonable conclusions. The explanations work to defend the model's incorrect answer rather than reason about the flaw.

Claim (Cum Hoc)
I'm moving to Connecticut because it is the richest state in the nation and I'm tired of being poor

Counterfactual ✓
The claim assumes moving to Connecticut will directly solve poverty ('I'm moving to Connecticut because it is the richest state... and I'm tired of being poor'), which is a causal fallacy ignoring other factors affecting personal wealth.

Zero-shot ✗
The claim expresses a personal decision based on a causal reason without faulty logic; moving to a richer state to improve one's financial situation is a reasonable causal connection.

Figure 5.4: Example of a fallacious claim where prompts disagree.

This pattern becomes especially clear when prompts disagree. With several of the fallacious claims, such as the cum hoc example in Figure 5.4, the prompting style leads to disagreement: one prompt identifies the reasoning as faulty, while another accepts it as valid. The explanations differ noticeably in these cases. One prompt highlights the problematic reasoning step and the other reframes it as sensible. This mirrors the pattern seen in the non-fallacious results, but in the opposite direction. Instead of inventing a fallacy, the explanation now smooths over one that is actually present.

Two explanation modes emerge when examining the fallacious claims. Correct answers come with short descriptions that match the fallacy type, while incorrect answers reinterpret the claim in a way that makes the reasoning seem legitimate. As in the non-fallacious cases, the explanations here mirror the model's decision and not the reasoning error. This limits how much the explanations can be used to understand how the model evaluates causal arguments.

Analysing the explanations is important because it shows not only what the model answers, but also which cues it uses to justify those answers. Across both fallacious and non-fallacious claims, the explanations show a consistent pattern. They rely on a small set of templates and shift to support the detection already made. Identical claims can receive opposite explanations solely due to the prompting style. These findings show that the explanations offer insight into surface patterns the model uses, but provide limited evidence of genuine causal reasoning. In the context of the research question, the results indicate that the explanations are unreliable indicators of causal reasoning. They tell us what decision the model made, not how it arrived there.

5.6 Methodological Considerations and Limitations

Several methodological limitations should be considered when interpreting the findings of this experiment. One central limitation is the uneven distribution of examples across fallacy types. Some types, particularly slippery slope and post hoc, contain significantly more items than other categories. See Table 3.1 for the full distribution. This imbalance affects the statistical precision of the reported accuracies. Categories with less data have wider uncertainty intervals, so small differences in accuracy should be treated with caution. In inductive reasoning, conclusions from small samples are weaker because they provide a less reliable basis for generalisation (Hurley, 2012).

Another limitation is the dataset's construction. Although the dataset combines items from established sources and publicly available fallacy collections, the claims' fallaciousness was not independently re-annotated for this study. Most labels were inherited from their source material, and for manually collected examples, classification relied on the authors' descriptions. This means that any upstream misclassifications carry into the evaluation. This is a limitation when assessing the validity of the model's errors.

Additionally, the non-fallacious subset was initially screened using GPT-5.1 before manual verification. Although each suggestion was reviewed and corrected if necessary, this introduces a form of model involvement in dataset selection. The impact is likely small since the final decisions were made by me, and all items in the subset originated from an annotated dataset (e-CARE), but it still warrants an acknowledgement.

This study also only evaluates a single model, GPT-4.1 mini. While testing across multiple prompting styles provides insight into prompt sensitivity, the results reflect a single model's behaviour. Recent work indicates that LLMs vary widely in their susceptibility to fallacies (Pan et al., 2024; Payandeh et al., 2024). Therefore, the generalisability of the results is limited by the model choice.

Finally, this experiment evaluates the detection of isolated claims. In the real world, fallacies often appear across several sentences in broader argumentative settings. Providing short, context-free statements gives a clear, controlled test environment, but it reduces external validity. Good performance on this data does not automatically mean reliable behaviour in more realistic situations. Previous work, such as the LOGICOM benchmark by Payandeh et al. (2024), shows that model behaviour can vary once the task shifts to more complex settings.

These limitations are important to acknowledge because they help clarify what the experiment can and cannot claim. They show where the findings rest on strong evidence and where the evidence is weaker. For example, uneven sample sizes and inherited labels introduce some uncertainty, and evaluating isolated statements does not capture all the ways fallacies appear in real communication. Being transparent about this makes the study more reliable since it shows how the results were produced.

At the same time, the limitations do not undermine the results. The overall pattern remains clear. The experiment consistently shows differences between prompting styles and between fallacy types, which appear across numerous items and multiple prompts. Rather than calling the results into question, the limitations help place them in the right context. The study describes how the model behaves under controlled conditions and defined tasks, rather than making broad claims about all forms of causal reasoning.

5.7 Future Work

Several directions for future research follow from this study. First, it could be valuable to replicate the experiment with multiple language models. This thesis evaluates only one model configuration, and other configurations could yield different results. Comparing different models would distinguish general behaviour patterns from model-specific behaviour.

A second direction concerns the dataset. Expanding the number of examples, especially for fallacy types that were sparsely represented, could strengthen the statistical basis for evaluating model accuracy. A more balanced dataset would also enable analysis of more subtle differences between fallacy types.

Future studies could also extend the experimental design. This study focuses on short statements presented out of context, but causal fallacies in everyday life often depend on narrative framing. Evaluating models on longer arguments or conversational exchanges could provide a more complete picture of how they handle fallacies in a realistic setting.

Another direction for future work is to compare model performance with humans. Having humans respond to the same set of claims would provide a baseline for evaluating the model's strengths and weaknesses. It could help distinguish between fallacies that are challenging for both humans and models, and those where human and model judgement diverge.

Further work might also explore more advanced prompting strategies. One option is to use prompts that guide the model through a more explicit reasoning process or to question its own reasoning before giving a final answer. Another option is to test few-shot prompting by providing solved examples as part of the instruction. Future studies could also explore combining different prompting styles, such as role-based and few-shot or cautious and counterfactual, to see if the combination improves performance. These techniques could help reveal if stronger guidance helps the model avoid fallacious judgements.

These directions could deepen our understanding of how LLMs interpret causal fallacies and give a clearer view of their strength and limitations in reasoning-focused tasks.

6 Conclusion

This thesis examined how GPT-4.1 mini evaluates causal fallacies and how different prompting styles influence its ability to do so. The study focused on three questions. First, how accurately the model detects different fallacy types. Second, how prompting styles affect its judgement. Third, what the model's explanations tell us about how it came to its conclusions.

The results cleary answer these questions. The model can detect some causal fallacies, especially those with strong surface patterns. Post hoc, slippery slope, and neglecting a common cause are detected with high accuracy. However, fallacies that depend on missing information, multiple causes, or more abstract causal structures, such as causal oversimplification and regression fallacy, are harder for the model to detect. Detection of non-fallacious arguments is also lower than for fallacious ones. This shows the model is better at spotting some stereotypical causal errors than at confirming valid causal reasoning.

Prompting styles influence this behaviour but do not change its limitations. Counterfactual and role-based prompts improve detection for fallacious claims the most, and make the model focus more on the relationship in the claim. The cautious prompt reduces false positives for non-fallacious claims but misses the most fallacies of all prompting styles. CoT does not outperform zero-shot prompting, suggesting that instructing the model to think step by step does not overcome its associative habits. Prompting techniques can change how the model makes decisions and highlight certain aspects of the claim, but they do not give the model new causal competence.

The analysis of explanations makes the pattern clearer. Correct detections come with focused and reasonable explanations, while incorrect ones tend to justify the model's mistake. These errors rarely show the reasoning process that led to the wrong answer. Instead, the model often introduces fallacies that are not present or ignores the causal

structure. This shows the explanations follow the decision rather than guide it. Identical claims can also receive completely different explanations depending on the prompt. The explanations reflect the model's tendency to rely on familiar templates instead of a consistent understanding of cause and effect.

These findings address the aims and research questions of the thesis. The starting point was the concern that LLMs mainly operate at an associative level and can reproduce human causal fallacies rather than overcome them. The experiments confirm this concern in more detail. By testing an LLM on nine causal fallacy types, this study shows that good performance is limited to fallacies that match patterns common in the training data, while performance drops when the patterns are less familiar, such as those that depend on hidden factors. This aligns with other work describing how LLMs mimic causal language without understanding the underlying logic.

The results also complement and nuance previous research on fallacies and causal reasoning. Broader fallacy benchmarks have shown that models struggle with false causality and that performance varies widely across datasets and prompts. This thesis adds a focused perspective by showing how a single model behaves across specific causal fallacies with controlled prompts and by connecting quantitative accuracy patterns to a qualitative analysis of explanations. It gives a more detailed view of how current models reflect informal logic.

Four central contributions stand out from this analysis. A controlled evaluation of how five prompting styles affect causal fallacy detection in a defined task. A comparison of nine causal fallacy types, and highlighting which are consistently detected and which remain challenging. A qualitative analysis of the model's explanations, showing how often they rely on surface-level cues. A demonstration that prompting can influence the model and sometimes improve accuracy, but it does not resolve deeper limitations from its training.

The findings underline an important distinction for anyone using language models in settings that rely on causal interpretation or argumentative clarity. The model can help surface potential fallacies and highlight patterns in reasoning, and with careful prompting it can support human review. However, they should not be treated as independent evaluators of causal arguments. Confident explanations can hide that its judgements rely on associative patterns. Human oversight is still essential for tasks where accuracy matters.

The central message of this thesis is clear. LLMs can detect many familiar causal errors and can be guided by prompting to do so more consistently. However, they do not yet reason about cause and effect the way causal inference and informal logic require. By making this gap more explicit, this thesis helps clarify where LLMs can be trusted and where future work is needed if we want models to do more than repeat our own causal fallacies.

Bibliography

Hengrui Cai, Shengjie Liu, and Rui Song. Is knowledge all large language models needed for causal reasoning?, 2024. URL https://arxiv.org/abs/2401.00139.

Maria Victoria Carro, Francisca Gauna Selasco, Denise Alejandra Mester, Margarita Gonzales, Mario A. Leiva, Maria Vanina Martinez, and Gerardo I. Simari. Do large language models show biases in causal learning?, 2024. URL https://arxiv.org/abs/2412.10509.

Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. e-CARE: a new dataset for exploring explainable causal reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 432–446, Dublin, Ireland, May 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.acl-long.33.

Patrick J Hurley. *A concise introduction to logic*. Wadsworth Cengage Learning, Boston, MA, 11th edition, 2012. ISBN 9780840034175.

Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schölkopf. Logical fallacy detection, 2022. URL https://arxiv.org/abs/2202.13758.

Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona Diab, and Bernhard Schölkopf. Can large language models infer causation from correlation?, 2024. URL https://arxiv.org/abs/2306.05836.

Nitish Joshi, Abulhair Saparov, Yixin Wang, and He He. Llms are prone to fallacies in causal inference, 2024. URL https://arxiv.org/abs/2406.12158.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. Better zero-shot reasoning with role-play prompting. In

Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4099–4113, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.228. URL https://aclanthology.org/2024.naacl-long.228/.

Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. Causal reasoning and large language models: Opening a new frontier for causality, 2024. URL https://arxiv.org/abs/2305.00050.

Yanda Li, Dixuan Wang, Jiaqing Liang, Guochao Jiang, Qianyu He, Yanghua Xiao, and Deqing Yang. Reason from fallacy: Enhancing large language models' logical reasoning through logical fallacy understanding, 2024. URL https://arxiv.org/abs/2404.04293.

Kyle Moore, Jesse Roberts, Thao Thi Minh Pham, and Douglas Fisher. Chain of thought still thinks fast: Apricot helps with thinking slow. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 47, 2025. URL https://escholarship.org/uc/item/18x411vv.

Raymond S. Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2):175–220, 1998. doi: 10.1037/1089-2680.2.2.175. URL https://doi.org/10.1037/1089-2680.2.2.175.

OpenAI. Gpt-4.1 mini. https://platform.openai.com/docs/models/gpt-4.1-mini, 2025. Accessed: 2025-11-18.

Fengjun Pan, Xiaobao Wu, Zongrui Li, and Anh Tuan Luu. Are LLMs good zero-shot fallacy classifiers? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14338–14364, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.794. URL https://aclanthology.org/2024.emnlp-main.794/.

Amirreza Payandeh, Dan Pluth, Jordan Hosier, Xuesu Xiao, and Vijay K. Gurbani. How susceptible are LLMs to logical fallacies. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8276–8286, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.726/.

Judea Pearl. *Causality*. Cambridge University Press, 2 edition, 2009. ISBN 9780521895606.

Pranab Sahoo, Ayush Kumar Singh, Sriparna Saha, Vinija Jain, Samrat Mondal, and Aman Chadha. A systematic survey of prompt engineering in large language models: Techniques and applications, 2025. URL https://arxiv.org/abs/2402.07927.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. Role play with large language models. *Nature*, 623:493–498, 2023. doi: 10.1038/s41586-023-06647-8. URL https://doi.org/10.1038/s41586-023-06647-8.

Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. Technical Report AD-767 426, Oregon Research Institute, Eugene, Oregon, August 1973. ONR Technical Report.

George Wick. Causal & inductive fallacies. URL https://amateurlogician.com/causal-inductive-fallacies/. Accessed: 2025-09-11.

Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. A survey on causal inference. *ACM Trans. Knowl. Discov. Data*, 15(5), May 2021. ISSN 1556-4681. doi: 10.1145/3444944. URL https://doi.org/10.1145/3444944.

Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. Causal parrots: Large language models may talk causality but are not causal, 2023. URL https://arxiv.org/abs/2308.13067.

Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. Context-faithful prompting for large language models, 2023. URL https://arxiv.org/abs/2303.11315.

Appendix A

Statement on the Use of AI Tools

Technical Assistance
I used ChatGPT 5 and ChatGPT 5.1 for technical support with LaTeX and code. This included help with tasks such as setting up tables, creating figures, fixing formatting issues, and suggesting solutions when something did not compile. I treated the suggestions as guidance and made my own adjustments before including anything in the thesis.

Dataset Screening
When constructing the non-fallacious dataset from the e-CARE dataset, I used ChatGPT 5.1 to narrow down potential examples from thousands of entries. I manually reviewed the suggestions and made substitutions as needed. The final selection reflects my own judgement.

Language Editing and Proofreading
I used Grammarly, ChatGPT 5, and ChatGPT 5.1 to improve the readability of my writing. I wrote the text myself first, then used these tools to identify spelling mistakes, unclear or repeated language, and possible improvements in phrasing. I did not copy the suggested edits directly. I reviewed them and rewrote the text to keep my intended meaning.

Concept Clarification
I used Perplexity and ChatGPT 5 as discussion partners to check my understanding of ideas and concepts when reading research papers. I read and interpreted the papers myself before using these tools to clarify details I found challenging.

The ideas, arguments, references, and conclusions are my own.

I am aware that I am responsible for all content of this master's thesis.

Appendix B

Answers To Self Test

Decide if the claim's causal reasoning is fallacious.
Answer *Yes* if the reasoning is fallacious and *No* if it is not.

	Claim	Yes	No
1.	The economy has improved greatly because of the new president.	☑	☐
2.	According to the weather bureau, there will be a cyclone in the eastern part of our city in the next two days. In the next two days, there wil be strong winds and rain in the eastern part of our city.	☐	☑
3.	The missile hit the plane's ailerons. The plane rolled out of balance and fell to the ground.	☐	☑
4.	I bought a ticket to win a new car at the mall, since I have never won anything like that in the past.	☑	☐
5.	In cold weather, Tom put on insulated clothing. His heat loss from the skin decreased.	☐	☑
6.	We need to stop allowing colleges to increase tuition every year. The next thing we know, it's going to cost more to attend college for one semester than it is to buy a new home!	☑	☐

Table B.1: Self-test solution