
A Novel Framework for Testing Causal Reasoning in LLMs: Design, Data Collection, and Evaluation

Abigail E. Thornton, Konstantinos K. Karageorgos, Melissa E. Greenberg, Daisy Alvarez Falcon, Mikaela R. Grace, Taletha M. Callahan-Kanik, Rafael S Villalobos Alva, & Fernando F. Migone

✉ fernando.migone@welocalize.com

✉ abigail.thornton@welocalize.com

🌐 <https://welodata.ai>

Abstract

In this paper, we address significant gaps and challenges in benchmarking the causal reasoning capabilities of large language models (LLMs). We propose a comprehensive and robust evaluation framework for multilingual causal reasoning and present results from our initial assessment. We critically analyze existing benchmarking datasets, highlighting their limitations in originality, complexity, and linguistic diversity, and we illustrate that current evaluations fall short in truly assessing multilingual causal reasoning abilities. We outline the design of our dataset, which involved human experts crafting long and complex prompts, which were translated into Spanish, Japanese, Korean, Turkish, and Standard Arabic. We evaluate model accuracy and consistency across these languages from over 20 models from 10 different developers. Our findings reveal that LLMs exhibit inconsistent causal reasoning with complex, novel prompts in English and underperformance in languages such as Turkish and Arabic. We argue that the evaluation underscores an urgent need for multilingual reasoning assessments, as model performance appears to decline with increased linguistic complexity and decreased availability of resources. We advocate for the development of additional multilingual causal reasoning training data to enhance the fine-tuning of models for improved causal reasoning capabilities.

Introduction

Large Language Models (LLMs) perform inconsistently with reasoning tasks as they lack the logic needed to complete complex, multi-step reasoning problems. Evaluations are typically broad, hindering improvements due to linguistic diversity and prompt complexity limitations. Recent research has also shown that foundational models consistently underperform on Causal Reasoning (Paul et al., 2024, Zečević et al., 2023, Romanou et al., 2023, among others), and it has been argued that when these models do successfully complete these tasks, they are typically accessing and repeating their training data (Zečević et al., 2023). Furthermore, many causal reasoning benchmarks for LLMs only evaluate this capability in English. Robust causal reasoning is a milestone towards Artificial General Intelligence (AGI), as understanding causal relationships is crucial for effective decision-making, problem-solving, and generalization in novel situations (Mumuni & Mumuni 2025). Enhancing LLMs' causal reasoning capabilities addresses these limitations and advances the development of AGI.

The goal of this publication is to present our approach to a deep and comprehensive evaluation that addresses gaps in LLM causal reasoning performance and sets the standard in multilingual causal reasoning research. This multilingual causal reasoning evaluation includes novel prompts which query causality in narrative documents written from the perspective of a participant involved in a series of fact-based events. Since LLMs extract relationships and causality from input text, within the prompt we embed both the questions and the associated documents and measure how accurately and consistently the model identifies causality of the scenario events and in the input texts. Below is an example which shows stories from the perspective of 2 different participants in a story embedded in a prompt with the same question and multiple-choice options. The correct answer is the same for both questions, but models do not consistently answer the same question about the same series of events in the same way.

All prompts are written using the same language as the narrative document. Hence, all prompts in the dataset contain monolingual question and story pairs. The goal of this evaluation is to measure the model's success in identifying causality across languages with varying linguistic characteristics, discover areas of weakness in causal reasoning, and compare the model's performance across multiple languages. All prompts and stories have been written in English and human-translated into Spanish, Japanese, Korean, Turkish, and Standard Arabic.

Figure 1: Two examples of prompts from the perspective of different characters with the same questions and set of response options.

Figure 1a: Welo Data MAS Causal Reasoning: Example 1

Welo Data MAS Causal Reasoning: Example 1

Prompt: You are a helpful assistant for causal relationship understanding. Review the following story and think about the cause-and-effect relationships. Then, answer the question that follows the story.

Story: The Massachusetts Department of Health requires all food-related establishments to adhere to specific rules to ensure food safety, such as refrigeration of all dairy items at all times. A late shift followed by a morning shift at Beans4All is exhausting. Last week, I worked an early shift following my late night shift and I was tired. One of our terrifying regulars, Fran, came in. Her order is simple enough-hot coffee with oat milk-but she is unfriendly. Yuki, the cashier, took the order and gave it to me. I poured the coffee, took the oat milk out of the cooler and added it, and then was on my way back to the counter when I tripped over the kitchen mat, spilling the coffee all over the floor and myself. Yuki rescued me by pouring a new coffee. She grabbed an oat milk out of the refrigerator, added it, and gave Fran the drink.

I went home early due to my burned arms and Yuki cleaned up. However, she forgot to put away the oat milk that I used and it sat out all night. The next day, our manager made a drink using the oat milk that, unbeknownst to him, had sat out all night. That customer got sick from the spoiled oat milk and had to go to the hospital. The health department fined us and Yuki was put on probation, and I feel terrible about it.

Question: What caused Yuki to be put on probation?

Identify the best choice from the list of options A-J. Select the option with the greatest contribution to causing the outcome. Append the option to the json list following the structure below: {"correct_answer": []}.

Options:

- a. Sam was afraid of Fran
- b. Sam was tired from a late shift
- c. Sam dropped the hot beverage
- d. Yuki cleaned up Sam's spill, forgetting about the oat milk
- e. Alex stepped in to help at the counter after Sam went home
- f. Jaime became ill at work
- g. The hospital staff determined Jaime had been food-poisoned
- h. The café was fined by the health department
- i. There are no causal relationships.
- j. There is not enough information.

Figure 1b: Welo Data MAS Causal Reasoning: Example 2**Welo Data MAS Causal Reasoning: Example 2**

Prompt: You are a helpful assistant for causal relationship understanding. Review the following story and think about the cause-and-effect relationships. Then, answer the question that follows the story

Story: The Massachusetts Department of Health requires all food-related establishments to adhere to specific rules to ensure food safety, such as refrigeration of all dairy items at all times, something we take seriously at my job. I worked my way up to manager at Beans4All, a local coffee shop, and take great pride in my work. Keeping the shop clean and following all of the safety rules is my top priority. As I am working one morning, I see one of our regulars Jaime, come in. I take her order of a hot tea with oat milk and quickly make it for her with a container of oat milk that was sitting on the counter. Later, we find out that Jaime got sick and had to go to the hospital. I worry it might have been our fault, and question whether the oat milk was spoiled.

After investigating further, I find the answer. Earlier in the morning, Sam and Yuki were working. Sam had worked a late shift the night before, and he and Yuki said that he was pretty exhausted. As a result, he spilled a hot coffee for one of our regulars, Fran, all over himself and the floor after tripping on the kitchen mat. After cleaning up, Yuki forgot to put the oat milk away and it sat out all day. When I added it to Jaime's tea, I had no idea. As a result, we were fined by the health department and I put Yuki on probation.

Question: What caused Yuki to be put on probation?

Identify the best choice from the list of options A-J. Select the option with the greatest contribution to causing the outcome. Append the option to the json list following the structure below: {"correct_answer": []}.

Options:

- a. Sam was afraid of Fran
- b. Sam was tired from a late shift
- c. Sam dropped the hot beverage
- d. Yuki cleaned up Sam's spill, forgetting about the oat milk
- e. Alex stepped in to help at the counter after Sam went home
- f. Jaime became ill at work
- g. The hospital staff determined Jaime had been food-poisoned
- h. The café was fined by the health department
- i. There are no causal relationships.
- j. There is not enough information.

Problem

It has been argued by Zečević et al., (2023), Bender et al., (2021), and others that LLMs do not actually demonstrate causal reasoning capabilities and instead repeat information from their training datasets. Zečević et al. argue that these models are not only ‘stochastic parrots’ (Bender et al., 2021) but are also sometimes ‘causal parrots.’ By this, Zečević et al. mean that LLMs make correlations about causality based on their training data. Adversarial testing has also shown that LLMs tend to identify correlations in causal reasoning tasks, but these correlations are typically extracted from training data – not the data the models are being assessed on (Kıcıman et al., 2023). To truly evaluate the model’s ability to perform causal reasoning requires novel, human-generated datasets that are distinct from datasets included in the model’s training.¹

Traditionally, causal reasoning benchmarks have been relatively simple, evaluating the relationship between two variables without providing additional context. Many studies have leveraged existing datasets, such as E-CARE (Du et al., 2022, leveraged by Gao et al., 2023), Tübingen cause-effect pairs dataset (Mooij et al., 2016 leveraged by Kıcıman et al., 2023), and Choice of Plausible Alternatives (COPA, Roemmele et al., 2011, leveraged by Ponti et al., 2020 as well as others) by embedding data from these datasets into a prompt with instructions to evaluate and identify which variable most likely caused the outcome or to identify the directionality between 2 variables. Below is an example from the COPA dataset which demonstrates the former type of question and is followed by an example in Table 2 below from the Neuropathic Pain Diagnosis dataset (Kıcıman et al., 2023) which demonstrate the latter type of question.

Table 1: “Examples of forward (Result [R]) and backward (Cause [C]) reasoning from the COPA datasets” (Roemmele et al., 2011).

	Premise	Alternative 1	Alternative 2
R	The man lost his balance on the ladder. <i>What happened as a result?</i>	He fell off the ladder.	He climbed up the ladder.
C	The man fell unconscious. <i>What was the cause of this?</i>	The assailant struck the man in the head.	The assailant took the man’s wallet.

¹ Additionally, some causal reasoning benchmarks evaluate another capability, such as natural language inference (NLI). While Causal Reasoning involves understanding the underlying causes and effects in situations or events, NLI simply refers to logical inferences from text.

Table 2: “Example cause-effect pairs from the Neuropathic pain diagnosis benchmark. ‘Dir.’ refers to the ground-truth causal direction between the variables” (Kıcıman et al., 2023).

Variable A	Variable B	Dir.
Right L1 Radiculopathy	Right adductor tendonitis	→
Pharyngeal discomfort	Right C3 Radiculopathy	←
Right L5 Radiculopathy	Lumbago	→
Left PTA	Left IA Radiculopathy	←
Left T3 Radiculopathy	Toracal dysfunction	→
DLS L5-S1	Right S1 Radiculopathy	→
Left C3 Radiculopathy	DLS C2-C3	←
Left C7 Radiculopathy	Left medial elbow problem	→
Right Ischias	Right L5 Radiculopathy	←
Right Morton trouble	Right L5 Radiculopathy	←

These testing methodologies merely require domain knowledge to correctly answer the questions about causality.²

If current benchmarks only evaluate performance on simple causal reasoning tasks, then models may be limited with respect to identifying causal relationships in novel data or in more complex narratives and pieces of text. Zečević et al., (2023) argue that while LLMs exhibit some ability to recognize and respond to causal relationships, they are far from mastering causality in the comprehensive way that humans do. Zečević et al.’s evaluations across different contexts highlight significant limitations in the models’ ability to generalize causal understanding beyond specific phrasings, recall causal facts accurately amidst linguistic variations, and effectively leverage knowledge base embeddings. These findings underscore the necessity for more robust and nuanced methods for evaluating LLMs’ causal reasoning capabilities.

Given the research on causal reasoning evaluations, we outline the following general challenges in causal reasoning tasks with novel and complex data:

- Discovery of causal relations & identification of non-causal relationships (Romanou et al., 2023, Gao et al., 2023)
- Normality violations and contribution to the final outcome (Nie et al., 2023, Kıcıman et al., 2023)
- Language variation (Zečević et al., 2023, Gao et al., 2023)

² Furthermore, it has been widely suggested that several of these datasets have likely been included in the models’ training datasets as evidenced by their increasing performance on these assessment as well as performance on other tests. For example, Kıcıman et al., (2023), perform a series of memorization and redaction tests as part of their evaluation to determine whether the datasets they used had been included in the training data.

Complexity in Causal Reasoning

To further advance causal reasoning capabilities in LLMs, a gap related to the complexity of the prompts must first be addressed. Zhang et al., (2023) propose a complexity hierarchy for evaluating a model's causal reasoning capabilities. In this hierarchy, there are three types of causal questions – each requiring the model to apply causal reasoning in an increasingly more complex context. Their overview of this hierarchy is provided in Table 3 below.

Table 3: Three types of causal questions (Zhang et al., 2023).

Type 1	Identifying causal relationships using domain knowledge
	<p>Example 1: Patient: Will my minor spine injury cause numbness in my shoulder?</p> <p>Example 2: Person: I am balancing a glass of water on my head. Suppose I take a quick step to the right. What will happen to the glass?</p>
Type 2	Discovering new knowledge from data
	<p>Example 1: Scientist: In a new scientific experiment. I observe two variables A and B which were from climate data. The observations are as follows: A[...], B[...] Please let me know if A causes B or B causes A.</p> <p>Example 2: Marketing specialist: I plan to launch a new membership program different from our competitors X and Y. There are two ways to design the benefit as members. The first is "buy four and get a fifth one for free," the other is 'get 20 dollar cash return for every 100 dollar spend'. Which one should I choose?</p>
Type 3	Quantitative estimating of the consequences of actions
	<p>Example 1: Sales manager: I have 1000 dealers with the following information about them I can only give membership to 100 of them next year. I want the membership program provides the highest revenue growth. Which 100 dealers should I choose?</p> <p>Example 2: Medical doctor: This is the third time that this patient has returned with lumbago. The epidural steroid injections helped him before, but not for long. I injected 12mn betamethasone the last two times. What is the dose that I should use this time?</p>

Consistent with Zečević et al., (2023), Zhang et al., (2023) argue that LLMs can accurately answer the first type of causal questions by identifying causal relationships using domain knowledge from the model's training data, demonstrating success comparable to human domain experts. However, Zhang et al. argue that the models perform inconsistently with the second type of causal questions as they are unable to

discover and apply new knowledge to make decisions about causality. Furthermore, the third question type requires high precision causal reasoning in context.

Similarly, in their review of existing causal reasoning benchmarks, Yang et al., (2024) argue that the Type 2 and Type 3 questions in Zhang et al., (2023)’s taxonomy require the model to apply causal reasoning with abstraction since the model cannot rely on domain knowledge alone to accurately identify causality. They argue that story-based contextual reasoning as well as graph-based tasks are an appropriate methodology for evaluating these higher levels of causal reasoning. Causal Reasoning Assessment Benchmark (CRAB, Romanou et al., 2023) and BigBench (Srivastava et al. 2024) are two of these types of datasets – each with a different approach to the dataset design.

The CRAB dataset evaluates a model’s ability to identify causal relationships between events described in multiple documents. The dataset includes 173 documents with the top 20 articles about 20 major events. The authors used GPT-3 to automatically extract events from each document and then human annotation was leveraged to annotate event pairs about the same event. This yielded contextual annotations of 2.7k pairs of real-world events which were then grouped into 352 causal chains. The documents and the events were then embedded in a prompt which queries causality as well as degree of causality – shown respectively in the Table 4 and Table 5 below.

Table 4: “Prompt for the Graded Pairwise Inference task” (Romanou et al., 2023).

Prompt: Graded Pairwise Causality – Multiple Choice Question (MCQ)
<p>You are a helpful assistant for causal relationship understanding.</p> <p>Think about the cause-and-effect relationships related to context.</p> <p>Context:</p> <p><DOCUMENTS></p> <p>Event: <EFFECT></p> <p>What is the most likely cause of this event?</p> <p>[A] <CAUSE 1></p> <p>[B] <CAUSE 2></p> <p>[C] <CAUSE 3></p> <p>[D] <CAUSE 4></p> <p>Let's work this out in a step-by-step way to be sure that we have the right answer. Then provide your final answer within the tags, <Answer>A/B/C/D</Answer>.</p>

Table 5: “Prompt for the Pairwise Causality Score Inference task” (Romanou et al., 2023).

Prompt: Pairwise Causality Score
<p>You are a helpful assistant for causal relationship understanding.</p> <p>Think about the cause-and-effect relationships related to context.</p> <p>Context:</p> <p><DOCUMENTS></p> <p>Event 1: <EVENT 1></p> <p>Event 2: <EVENT 2></p> <p>What is the causality score between Event 2 and Event 2 from 0 to 100?</p> <p>Score above 80: Event 1 is definitely responsible for Event 2.</p> <p>Score between 50-80: Event 1 might have been responsible for Event 2.</p> <p>Score lower than 50 Events are somehow related but definitely NOT causally related.</p> <p>Let's work this out in a step-by-step way to be sure that we have the right answer. Then provide your final answer within the tags, <Answer>score</Answer>.</p>

The results of the CRAB evaluation show that the LLMs perform well in the context of explicit causal statements; however, they struggle with applying causal reasoning to new scenarios. Furthermore, when evaluating the LLM’s ability to identify causal structures from groups of events, it was found that the LLM performs worse with events from complex causal structures as opposed to linear causal chains. Finally, the models struggled with appropriately measuring the degree of causality between two events, which the authors hypothesize results from causal chains largely related to temporality. The authors argue that these results indicate that the LLMs exhibit limitations in offering predictions or explanations about causality between two events and hence face challenges related to generalization and accuracy in these contexts.

BigBench (Srivastava et al., 2022), a series of datasets used to evaluate LLMs and their future capabilities, also includes a causal judgment dataset which aims to evaluate LLM performance on causal reasoning with moral, intentional, and counterfactual analysis. The dataset includes 190 stories with multiple cause-effect events from 30 papers published in the span of 1989 to 2021. Table 6 below is an example of a story and a pair of questions asked about the story.

Table 6: Example of a causal story and pair of questions asked about the story (Srivastava et al., 2024)

<p>Story: Mr. Wallace is highly influential in the organized crime scene. Although he commands the respect of many, there are also a number of people who stand to benefit from his death. Today, he was having lunch with one of his associates, Mr. Vincent. At one point during their meeting, Mr. Wallace left the table to go to the toilet. While he was gone, Mr. Vincent managed to slip a lethal dose of poison into his martini. The poison is known to take one hour (give or take 10 minutes, depending on the victim's body weight) to kill and would go unnoticed for the first half hour after consumption. When Mr. Wallace returned to the table, he finished his lunch and then drank the rest of his martini. The two men concluded their meeting a few minutes later. Mr. Wallace paid the bill and they both left the restaurant going in different directions. Mr. Wallace had another important business meeting about 15 minutes away in a remote part of the countryside. On an isolated stretch of road, he noticed that he was suddenly being followed by a van that had pulled out from the roadside. The van accelerated and began hitting his car from behind. Mr. Wallace recognized that the driver was Mr. Bruce, an associate of one of Mr. Wallace's rivals. He tried to evade the van but there was no time. The van relentlessly hit his car and pushed it off the side where it fell into a ravine and exploded. The remains of Mr. Wallace's body were discovered later that day. The chief coroner's report later revealed that Mr. Wallace had received fatal burns in the car explosion. The report also indicated, however, that a lethal dose of poison was found in Mr. Wallace's blood.</p>
<p>Question 1: Did the crime life cause Mr. Wallace's death?</p>
<p>Question 2: Did the car crash cause Mr. Wallace's death?</p>

The causal judgment task attempts to measure the models' ability to reason over multiple potential causes and effects and then evaluate how the models' responses compare to those of humans. Srivastava et al., (2022) refer to Knobe, J. (2003)'s series of experiments evaluating how humans identify causality of the same scenario when the question is phrased slightly differently with respect to intentionality. Knobe's results showed that humans have different responses to different versions of the same story depending on how the question of causality is phrased. Subsequent research suggests that when identifying causality, humans often factor in other considerations, such as norm violation, intentionality, morality, etc.

Yang et al., (2024) further describe limitations with the approaches for evaluating more complex types of causal reasoning. They argue that LLMs might only perform well on these tests since the causal path is often described in the same sequence as the 'temporal or spatial' order of events. That is, we often describe a chain of events in linear order, and models may be inferring causality from that sequence of events alone. The advent of more sophisticated datasets like CRAB and BigBench represents a crucial step forward, challenging LLMs to extract and identify causal relationships within extended texts and multifaceted contexts. The insights gained from these evaluations reveal that while LLMs exhibit proficiency in identifying explicit causal statements, they still encounter difficulties in generalizing causal reasoning to novel scenarios and complex causal structures. Furthermore, the alignment of LLMs with human reasoning, especially in contexts involving intentionality, morality, and norm violations,

underscores the complexity and variability of causal reasoning. These findings highlight both the progress and the ongoing challenges in enhancing the causal reasoning capabilities of LLMs, emphasizing the need for continuous advancements in dataset design and model training methodologies.

Lack of Linguistic Diversity

There is also a general lack of linguistic representation in causal reasoning benchmarks. For example, the XCOPA dataset was created from the COPA dataset by translating and re-annotating the data into 11 languages from 11 different language families. These languages were purposefully selected to have a balanced distribution across language families as well as geographical locations (Ponti et al., 2020). The table below provides an overview of the languages included in the dataset.

Table 7: Languages covered in the XCOPA dataset (<https://github.com/cambridgeltl/xcopa>).

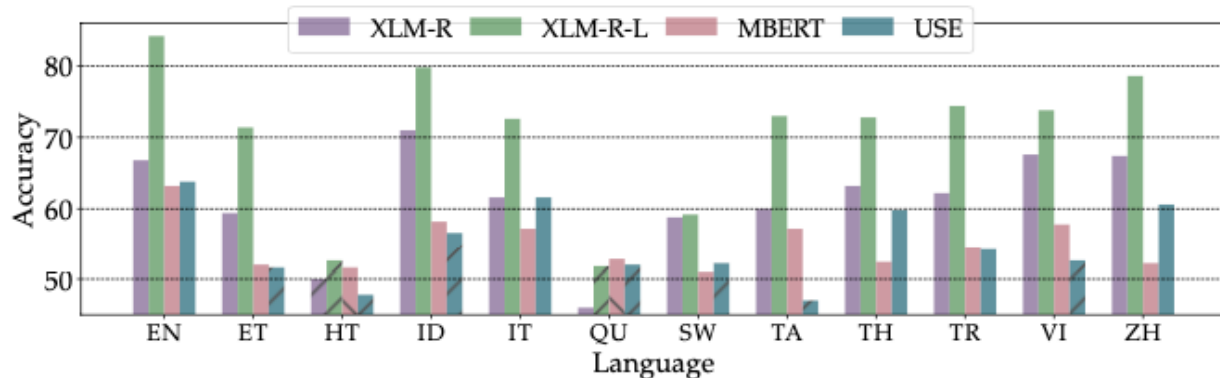
ISO 639-2	Name	Family	Area
et	Estonian	Uralic	Northern Europe
ht	Haitian Creole	French Creole	Caribbean
id	Indonesian	Austronesian	Southeastern Asia
it	Italian	Indo-European	Southern Europe
qu	Southern Quechua	Quechuan	Southern America
sw	Swahili	Niger-Congo	Eastern Africa
ta	Tamil	Dravidian	Southern Asia
th	Thai	Kra-Dai	Southeastern Asia
tr	Turkish	Turkic	Western Asia
vi	Vietnamese	Austroasiatic	Southeastern Asia
zh	Mandarin Chinese	Sino-Tibetan	Eastern Asia

In addition to these attributes, Ponti et al. (2020) note that some of these languages have a low degree of resource availability in comparison to the others.³ They evaluate a series of pretrained multilingual

³ It has been noted by Joshi et al., (2020) that there is a general lack of linguistic diversity in NLP – especially with respect to publicly available training data. They create a resource-level taxonomy and categorize almost 2.5k languages on a scale of 0-5 where 0 indicates there is no public training data and 5 indicates that there is a high-level of training data due to dominant online presence, annotated datasets, etc. Consequently, level 5 languages are well-represented in the training data and hence are well-supported with each advance in NLP whereas other languages lack sufficient training data, resulting in poor support and lagging behind in the latest advancements in NLP technology.

models, including XLM-R (Conneau et al., 2020), MBERT (Devlin et al., 2019) and multilingual USE (Yang et al., 2019) and show that languages not included in the training data perform worst (Haitian Creole and Southern Quechua). Figure 2 is the graph shared in their paper with an overview of each model's performance across each language. Furthermore, Ponti et al. observe that some models perform better on languages other than English. They hypothesize reasons for this including transfer from English as well as distinct scripts not used in other languages.

Figure 2: Per Language XCOPA Results (from Ponti et al. 2020)



Ponti et al.'s findings highlight the varying performance of multilingual models across different languages, which brings us to an important consideration regarding the types of datasets used to evaluate these models. As observed earlier in this section, COPA is a relatively simplistic dataset and success on COPA demonstrates simple causal reasoning abilities – not more complex ones. Thus, success with XCOPA indicates the same, demonstrating that there is a need for more complex causal reasoning evaluations for other languages in addition to English.

Methodology

In this section, we outline our approach for evaluating LLM causal reasoning capabilities in multiple languages. We i) design a completely novel dataset; ii) leverage human experts to create all the data in the dataset; and iii) create complex prompts that address various aspects of causality. This yields a comprehensive and robust evaluation of multilingual causal reasoning in LLMs.

There are three parts to our dataset: i) fact-based scenarios, ii) scenario-based narratives, and iii) question & answer pairs. We asked domain experts to generate novel, fact-based and domain-specific scenarios that utilize terminology and jargon from the respective field or industry. Then, we instructed writers to employ the scenarios to develop the narratives, focusing on different perspectives of a character in the story. Finally, experts with advanced degrees and at least 5 years of experience in fields like Cognitive Science, Philosophy, Linguistics, and NLP research created question and answer pairs based on the facts and events in the scenarios.

Crucially, this dataset design ensures that the model has not seen any of the stories nor any of the questions related to the story. Since the scenarios and stories are completely unique and novel and have never been seen by any of the models, we are able to evaluate a model’s true causal reasoning capabilities, without concern that the model was trained on our data. Additionally, the prompt design is quite complex since each prompt embeds a single story and a single question that was generated based on the fact-based scenario (rather than the subjective story). The prompt design is also versatile as we can easily embed a robust set of questions, targeting the model’s ability to perform different causal reasoning tasks in the context of a single story. Hence, we can evaluate a model’s causal reasoning capabilities and consistency in demonstrating these capabilities when a series of events is described both subjectively and objectively.

Finally, many causal reasoning benchmarks only evaluate this capability in English, restricting full understanding of the robustness of these capabilities as demonstrated by the models. We address this gap by having all prompts translated by human translators with domain-specific expertise into the following high-resource languages: Spanish, Japanese, Korean, Standard Arabic, and Turkish.⁴ A monolingual English prompt is provided below alongside its Spanish counterpart.

⁴ Here, we follow Yong et al., (2023)’s definition of ‘high resource’ languages. This description includes languages with resource availability levels 4 and 5 following Joshi et al., (2020)’s description and classification.

Figure 3: Comparison of English and Spanish Causal Reasoning Prompt**Figure 3a: English Causal Reasoning Prompt**

You are a helpful assistant for causal relationship understanding. Review the following story and think about the cause-and-effect relationships. Then, answer the question that follows the story.

Story: I've been looking forward to my friend Fran's 60th birthday party for weeks. I got him the perfect gift--a gun. He's always talking about the Second Amendment. I find one at Sam's gun shop, pass the background check, and pick it up a few days later. I want to make sure it works well before giving it to Fran, so I stop by the gun range. While there, I see my ex-wife, Hiro, with her boyfriend. I am so flustered and jealous that I pack up the gun without putting the safety on or taking the bullets out.

On Saturday, I take the gun to Fran's party. Hiro and her boyfriend are there, as she was friends with Fran too. While Fran is opening his gifts, his dogs run in and jump on him, causing him to drop the gun, which then goes off. The bullet hits Hiro's new boyfriend, and he dies. Fran is arrested for manslaughter and felon in possession of a firearm. It turns out that Fran had gone to federal prison in 1998 for a year after being convicted of tax evasion. It was easy for Fran to keep it quiet and off the internet in the late 90s.

Question: What caused Fran's arrest after the birthday party?

Identify the best choice from the list of options A-J. Select the option with the greatest contribution to causing the outcome. Append the option to the json list following the structure below: {"correct_answer": []}.

Options:

- a. In 1998, Fran went to federal prison for a year for tax evasion.
- b. Yuki was distracted when he spotted his ex-wife Hiro with a new boyfriend.
- c. Yuki placed the present on the gift table.
- d. Yuki noticed that Hiro was also at the party with her new boyfriend.
- e. Fran dropped the gift.
- f. The gun's lone bullet shot Hiro's boyfriend dead.
- g. Fran kept quiet about going to federal prison for a year.
- h. Yuki forgot to take all the bullets out of the gun and put on the safety lock.
- i. There are no causal relationships.
- j. There is not enough information.

Figure 3b: Spanish Causal Reasoning Prompt

Serás un ayudante útil para la comprensión de relaciones de causalidad. Revisa la siguiente historia y piensa en las relaciones de causa y efecto. Luego, responde la pregunta tras leer la historia.

Story:

Llevaba semanas esperando con impaciencia la fiesta del 60 cumpleaños de mi amigo Fran. Le compré el regalo perfecto: una pistola, ya que siempre está hablando de la Segunda Enmienda. Encontré una en la armería de Sam, superé la comprobación de antecedentes y la recogí unos días después. Quería asegurarme de que funcionaba bien antes de dársela a Fran, así que me pasé por el campo de tiro. Mientras estaba allí, vi a mi exmujer, Hiro, con su novio. Me puse tan nervioso y celoso que guardé la pistola sin ponerle el seguro ni quitarle las balas.

El sábado, llevé el arma a la fiesta de Fran. Hiro y su novio estaban allí, ya que ella también era amiga de Fran. Mientras Fran abría sus regalos, sus perros entraron corriendo y saltaron sobre él, lo que le hizo soltar el arma, que se disparó. La bala alcanzó al nuevo novio de Hiro, y este murió. Fran fue detenido por homicidio involuntario y por ser un delincuente convicto en posesión de un arma de fuego. Resulta que Fran estuvo durante un año en una prisión federal en 1998 tras ser condenado por evasión fiscal. A finales de los 90, a Fran le resultó fácil mantenerlo en secreto y sin que se divulgara por Internet.

Question: ¿Qué causó el arresto de Fran tras la fiesta de cumpleaños?

Identifica la mejor opción de la lista de opciones A-J. Selecciona la opción que mayor contribución tenga a causar el resultado. Añade la opción a la lista de JSON según la estructura dada a continuación:
{“correct_answer”: []}.

Options:

- a. En 1998 Fran fue a una prisión federal durante un año por evasión de impuestos.
- b. Yuki estaba distraído cuando vio a su exmujer Hiro con un novio nuevo.
- c. Yuki puso el regalo en la mesa de los regalos.
- d. Yuki se dio cuenta de que Hiro también estaba en la fiesta con su nuevo novio.
- e. A Fran se le cayó el regalo.
- f. La única bala de la pistola mató al novio de Hiro.
- g. Fran no le dijo a nadie que había estado en una prisión federal durante un año.
- h. A Yuki se le olvidó sacar todas las balas de la pistola y poner el bloqueo de seguridad.
- i. No hay relaciones causales.
- j. No hay suficiente información

Dataset Design

Our data is designed to capture the type of causal reasoning found in typical human interactions. This is because all scenarios are domain-specific and include a series of facts and/or events that demonstrate logic-based causality. Logic-based causality refers to the use of logical reasoning and domain knowledge to reason about connections between facts or events to make determinations about cause-and-effect

relationships. We also narrow the focus of our scenarios to instances of specific causality. Specific causality refers to a distinct occurrence where a cause-effect relationship is clearly identified and can be observed or analyzed. Hence, we restrict the scope of this project to the analysis of cause-and-effect relationships in specific instances where a cause-and-effect relationship can be clearly identified through logical reasoning.⁵ Finally, we evaluate the model’s ability to i) identify causal relationships; ii) discern between a cause and a confounder; iii) determine the normality violation in a chain of causal events; and iv) perform these tasks in the context of language variation. The evaluation, therefore, is robust and comprehensive because it assesses the model’s capability to perform various causal reasoning tasks across subjective data in various linguistic contexts.

Our dataset includes 6 scenarios in each of the following domains: i) Legal & Criminal Justice; ii) Health, Medicine & Science; iii) Finance, Business, & Economics; and iv) General. The scenarios were further divided into 3 types: i) Standard Causation, ii) Normality Violation – Explicit, iii) Normality Violation – Implicit. Below, we provide a table with a definition for each of these scenario types:⁶

Table 8: Classification of Causal Scenarios

Standard Causation	Scenarios that depict a clear cause-and-effect relationship without involving any violations of explicit or implicit norms. These scenarios are centered on demonstrating how one event or action leads to another, highlighting the logical or natural sequence between facts.
Normality Violation – Explicit	A scenario where there is at least one violation of explicit norms. Explicit norms are policies, rules, laws, regulations, documented and enforced standards , etc.
Normality Violation - Implicit	A scenario where there is at least one violation of an implicit norm. Implicit norms are informal, unwritten rules that guide behavior, including social norms, cultural customs, workplace etiquette, and everyday manners . These implicit norms cover aspects like personal space, privacy, dress codes, polite language, and hierarchical respect.

An expert with an advanced degree and at least 5 years of career experience in the domain created the scenario. This scenario was then reviewed and adjusted by another expert with similar qualifications to ensure that it was i) domain-specific; ii) expressed logic-based causality in a specific causal instance; and iii) adhered to the scenario type. Each scenario was then distributed to two separate pipelines: story generation and question-answer generation.

In the story generation pipeline, multiple stories were created for each scenario, and each story was written from the perspective of a different participant in the scenario. For the sake of consistency and to

⁵ Logic-based causality contrasts with covariance-based causality, which identifies causal relationships through statistical covariation. Specific causal instances also differ from general causal instances, the latter describing patterns where a type of cause generally leads to a type of effect.

⁶ An implicit normality violation also includes cultural norms. We recognize that the dataset may include cultural norms that do not have a similar counterpart in other languages and cultures. We leave this aside for now as our initial evaluation aims to compare causal reasoning across languages with the linguistic and semantically equivalent prompts.

avoid any bias to overly simple or overly complex stories, the original English stories had an average of 232 words. To maintain consistency across the stories, we provided a stringent set of writing guidelines and requirements such as including all causal events in the story.

For each scenario, we also created 9-14 questions depending on the scenario type. For Causal Discovery, we test both binary and multiple-choice questions, but for Normality Violation and Language variation questions, we only evaluate multiple choice questions. Below is a general overview of the question categories and their types of questions. We provide a brief definition for each of these below and review them in greater detail in the Evaluation Section.

Table 9: Overview of Question Categories

Question Category	Definition	Question Type
Causal Discovery (Cause)	The process of finding out which variables directly cause a change in another variable.	Binary, Multiple Choice
Causal Discovery (Confounder)	The process of discovering whether a Confounder directly caused an outcome. We also refer to these as Causal Discovery – Non-causal questions since the two variables in question do not have a cause-and-effect relationship.	Binary
Language Variation	Same as the Causal Discovery – Multiple Choice Question (MCQ) but with a different verb phrase (i.e. <i>led to</i> , <i>triggered</i> , <i>brought about</i> , <i>resulted in</i> , etc.)	Multiple Choice
Normality Violation	Another MCQ about how likely a causal event aligns with the violation of a prescriptive norm (social, moral, or legal).	Multiple Choice

Causal Discovery refers to the LLM’s ability to successfully identify causal relationships in a piece of text. Although this LLM capability has been widely evaluated, less work has been done to evaluate an LLM’s ability to discern between causal and non-causal relationships. In addition to the complexity levels observed by Zhang et al., (2023), Kıcıman et al., (2023) also acknowledge that it can be difficult to formally identify and represent relationships between variables in a dataset since this requires background context and information to make causal judgments. They observe that the following causal judgments are difficult to make without human knowledge of various types of background context and information:

Table 10 (summarized from Kıcıman et al., 2023):

Causal Judgment	Definition
Causal Frame	The set of candidate causal events deemed relevant to a particular outcome, determined by domain knowledge and common sense after the outcome has occurred.
Necessary Causality	The concept that a candidate cause must have occurred for the outcome to take place. If the candidate cause did not happen, the outcome would not have occurred.
Sufficient Causality	The idea that the occurrence of a candidate cause would have led to the outcome event, even if other causal events had occurred differently. It suggests that the cause alone could produce the outcome under varying circumstances.
Normality	The extent to which causal events align with statistical or prescriptive norms (social, moral, or legal), with norm violations often judged as more causative.
Other Human Factors	Factors like bias towards action, intention, and the interpretation of bad outcomes, where humans tend to ascribe more causality to intentional actions and undesirable outcomes.

We put aside the categories Necessity, Sufficiency, and Other Human Factors at this time, and hence, our questions focus on Causal Frame and Normality in the context of linguistic variation.

Languages

We provide a brief overview of the languages included in our dataset as well as our empirical approach to evaluating multilingual causal reasoning. We collected data for six high-resource languages with varying population sizes and three different types of word orders. The result is a comprehensive dataset that enables us to assess an LLM’s multilingual causal reasoning capabilities.

Language Selection

Our causal reasoning assessment includes stories and questions in English, Spanish, Japanese, Korean, Standard Arabic, and Turkish. These 6 languages were chosen based on their resource level (Joshi et al., 2020; see footnote 3), approximate number of speakers (Ethnologue), and their word orders (Ethnologue). In Table 11 below, we provide the statistics for each of these per language.

Table 11: Overview of Language Characteristics and Resource Availability

Language	Word Order	Approximate Population Size	Resource Level
English	SVO	1,515,231,760	5
Spanish	SVO	559,520,830	5
Japanese	SOV	123,466,720	5
Korean	SOV	81,128,730	4
Standard Arabic	VSO	332,459,000	5
Turkish	SOV	90,095,180	4

Since it is widely observed that LLMs generally struggle with causal reasoning, we limited our language selection to high-resource languages (see footnote 4) with a large volume of speakers. It has also been widely observed that the dominant word orders across all languages is SVO and SOV with a smaller number of languages exhibiting primarily VSO word order (Dryer 2013).⁷ Hence, to further diversify the data, we selected languages with word orders: SVO, SOV, and VSO.

Translation

Both stories and questions were first written in English and then translated across the other five languages listed in Table 3. We recognize that this methodological approach may have limitations with respect to semantic equivalence and naturalness that result from both linguistic and cultural nuances. To minimize these risks, we engaged translators with domain-specific expertise, instructing them to avoid word-for-word translations and instead focus on preserving the original semantics and utilizing natural word choices and grammatical structure. By standardizing the data via translation, we can evaluate the LLM’s response to the same question about the same story in multiple languages.

Prompt Design

We then compiled prompts with generic task instructions, a story, question with answer options, and specific instructions for each question category, all in the same language. This dataset includes two sets of prompts: one utilizing Chain of Thought (CoT) prompting⁸ and one without it. The primary objective of our dataset was not to evaluate the effectiveness of CoT prompting; however, given inconsistent results with CoT prompting (Wei et al., 2022 and Yao et al., 2023 show how CoT prompting improves results whereas Zhang et al., 2023, Zečević et al., 2023, Paul et al., 2024, Gao et al., 2023 demonstrate that even

⁷ Word order refers to the order of the grammatical arguments subject (S) and object (O) and the finite verb (V).

⁸ CoT prompting is a technique used to enhance the reasoning capabilities of these models by guiding them to generate intermediate reasoning steps before arriving at a final answer. Instead of prompting the model to directly produce an answer to a complex question or problem, CoT prompting encourages the model to break down the problem into smaller, more manageable parts and to articulate the logical sequence of steps or thoughts that lead to the solution.

with CoT prompting, there are inconsistencies in responses), we included this as an attribute of our dataset to evaluate the effectiveness of CoT prompting in long, complex prompts. We discuss this in more detail in the Results section and explore opportunities for future research in this area.

Evaluation

In this section, we provide a general overview of our evaluation methodology and results. We first outline the scoring methodology and provide an overview of the metrics. Then, we review our results and discuss our findings. We evaluate both large and small LLMs from a variety of developers.⁹ We provide the full list of developers from whom we sampled models in the list below:

Table 12: Developers of Tested Models¹⁰

Anthropic	Cohere
Google	OpenAI
X	Mistral
AI21Labs	Alibaba Cloud
Meta	Amazon

We have provided the list of all developers, but we do not share the list of all models we evaluated. Each developer and their corresponding model are mapped to a pseudonym. We understand that there can be important distinctions made between model families and models of different sizes, and hence, we group all models from the same developer under the same family. Where possible, we provide individual scores to compare performance on various tasks. Finally, we discuss our findings and explore hypotheses for future research.

Evaluation Metrics

To evaluate the performance of various LLMs on our novel dataset, we adopted accuracy as the primary evaluation metric. This choice aligns with methodologies used in popular benchmarks like Massive

⁹ Models were categorized based on their relative size within their respective families. Specifically, a model is considered small if it is the smallest in its family; medium if it is of intermediate size within its family; and large if it is the largest model in its family. This relative classification helps in understanding the model's capacity in the context of its peers. If there was only a single model in the family, we considered the model as i) small if it has approximately 8 billion parameters or less; ii) medium if it has between 9 billion and ~100 billion parameters; large for all other models.

¹⁰ All of the models we tested are considered multilingual language models since they can all perform in multiple languages. Most of the developers provide a high level overview of the languages included in the training data, but the list of languages varies by model. All models have been trained on English data and the majority of models have also been trained on the Spanish, Korean, Japanese, and Arabic. Turkish has only been explicitly mentioned as being included in the training data for about 2 of the model families.

Multitask Language Understanding (MMLU, Hendrycks et al., 2020), enabling straightforward comparisons and ensuring consistency with established practices. Accuracy measures the proportion of correct responses among all evaluated questions, providing a clear and interpretable assessment of model capabilities in addressing complex causal reasoning tasks. This metric is particularly well-suited for tasks with a definitive set of correct answers, as is the case with our dataset. Further resources that underscore the relevance of accuracy in similar settings include the MMLU paper (Hendrycks et al., 2020) and discussions on evaluation paradigms in “Beyond Accuracy: Behavioral Testing of NLP Models” (Ribeiro et al., 2020).

Results

The initial results of our evaluation, which focus solely on monolingual prompts (i.e., both the stories and questions are in the same language), provide insights into various aspects of multilingual causality across parameters like languages, domains, causal tasks, and question types. The primary findings highlight: (i) accuracy differences by locale and domain, (ii) the challenge of causal discovery regarding the direct cause versus a normality violation, (iii) language variation, and (iv) the impact of Chain of Thought (CoT) prompting. We discuss the overall accuracy results followed by more specific analyses within languages, domain, and question types.

Overall Accuracy Overview

Although we evaluate a variety of parameters in this assessment, we first provide an overview of the overall accuracy of the models. Notably, we see a wide variety of overall accuracy scores across the models. We provide a visualization of the overall score distribution as well as a visualization with each model’s score. Figure 4 below shows that the median overall score is .56 with maximum values approaching .7 and minimum values around .35. There are also two outliers below the .35 minimum. Figure 5 below shows that these two outliers originate from the Family 3 models with scores of .33 and .15 respectively. Additionally, Figure 5 shows that Family 8: Model 1 has the top score of .68 – which is well above the average with a noticeable distance from models in the same family as well as those in other families.

Language-specific Accuracy Overview

As expected, we find that there is a high degree of variability in overall performance by language, and we hypothesize that some of this is related to a model’s resource availability per language as well as differences in the language’s standard word order. English and Spanish are top performers across all question types for all domains. We expect that these results are directly related to an existing training data bias in these models where these languages have a dominant presence (see Joshi et al., 2020). Interestingly, Figure 6 also shows that the maximum value in Spanish exceeds the maximum value in English with scores of .72 and .71 respectively. Although the difference between the two maximum overall accuracy scores is not very large, there is a non-negligible difference in minimum scores in these two languages (.48 in Spanish and .43 in English).

Figure 4: Accuracy – Overall Performance

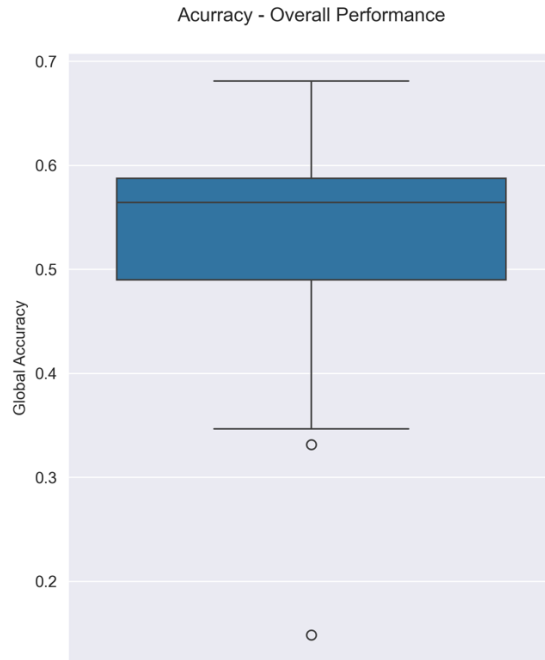


Figure 5: Global Accuracy – Overall Performance

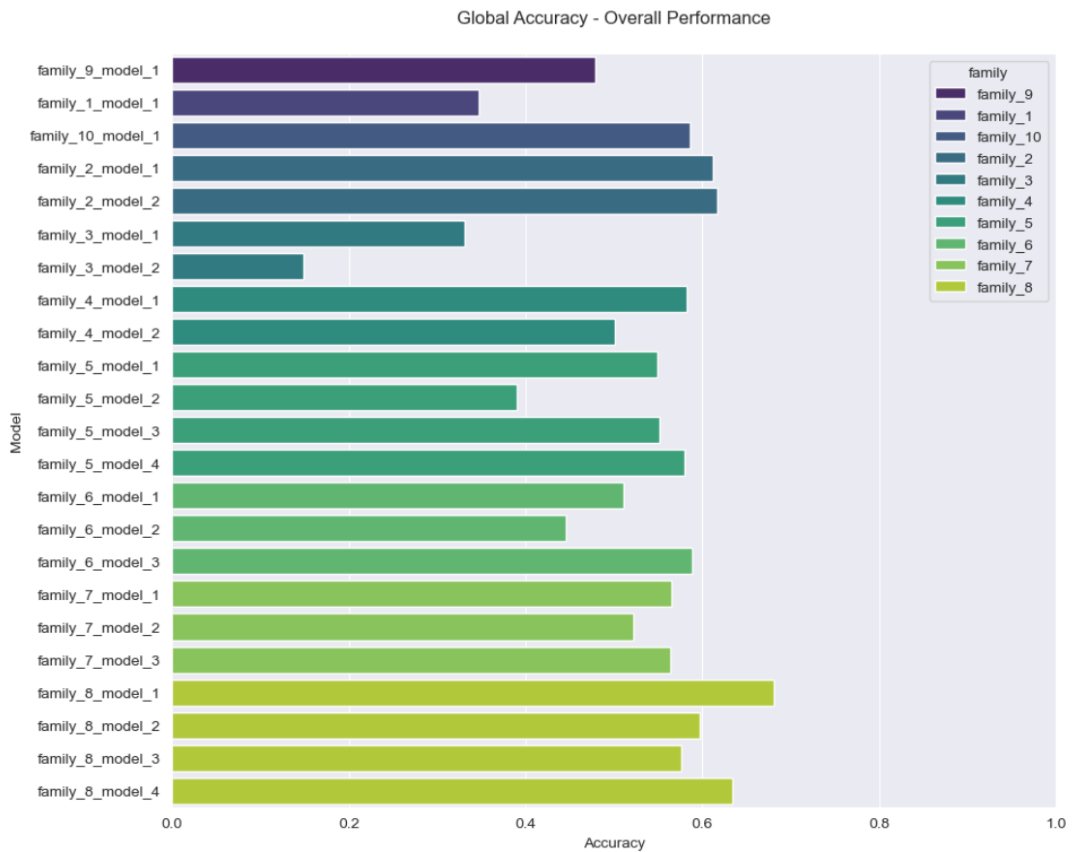
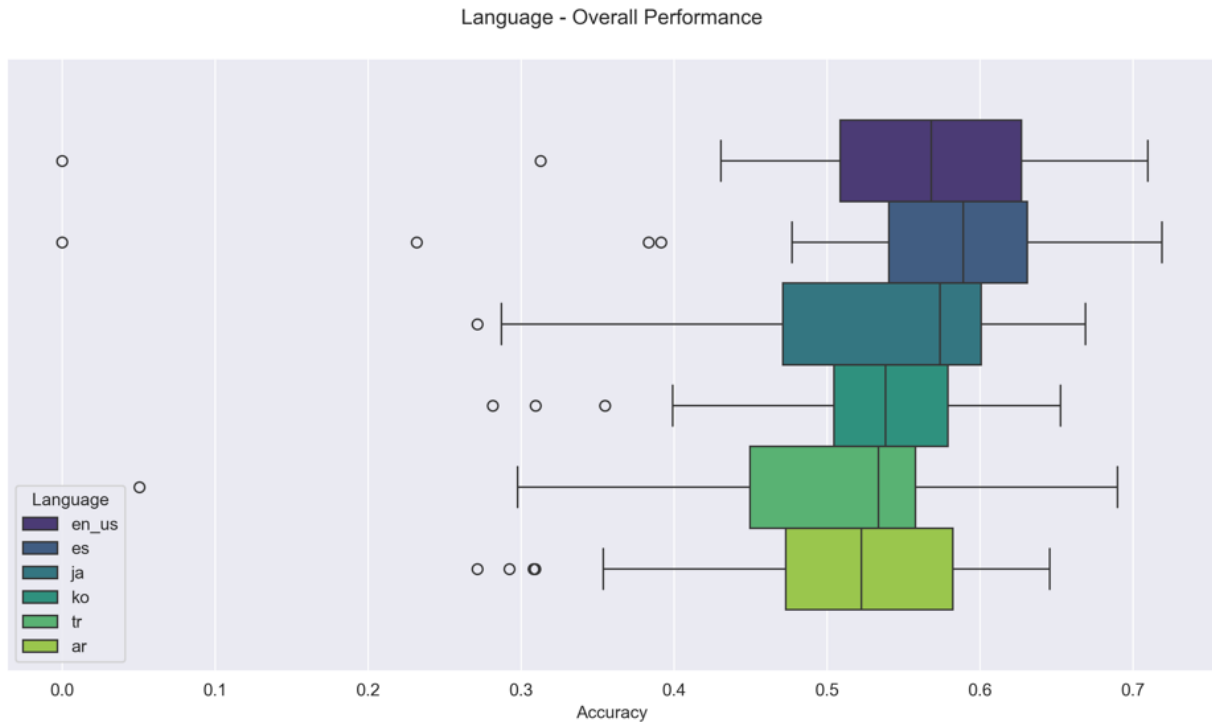


Figure 6: Language – Overall Performance

English and Spanish also contrast with Japanese and Turkish as these two languages have the highest degree of variability in their scores. Although the two languages do not necessarily perform the worst on the assessment, Japanese has an accuracy range of .38 and Turkish has an accuracy range of .39 – which is notably higher than most other languages. For example, Korean and Arabic have the lowest maximum values, and Arabic has the lowest median value. The average range of their scores, on the other hand, is approximately .25 and .2 respectively. We have calculated these values (excluding outliers) and we provide them in the table below.

Table 13: Language Performance Metrics: Accuracy Statistics

Language	Average Accuracy	Minimum Accuracy	Maximum Accuracy	Accuracy Range
Arabic (ar)	0.54	0.35	0.65	0.29
English (en_us)	0.58	0.43	0.71	0.28
Spanish (es)	0.60	0.48	0.72	0.24
Japanese (ja)	0.54	0.29	0.67	0.38
Korean (ko)	0.55	0.40	0.65	0.25
Turkish (tr)	0.51	0.30	0.69	0.39

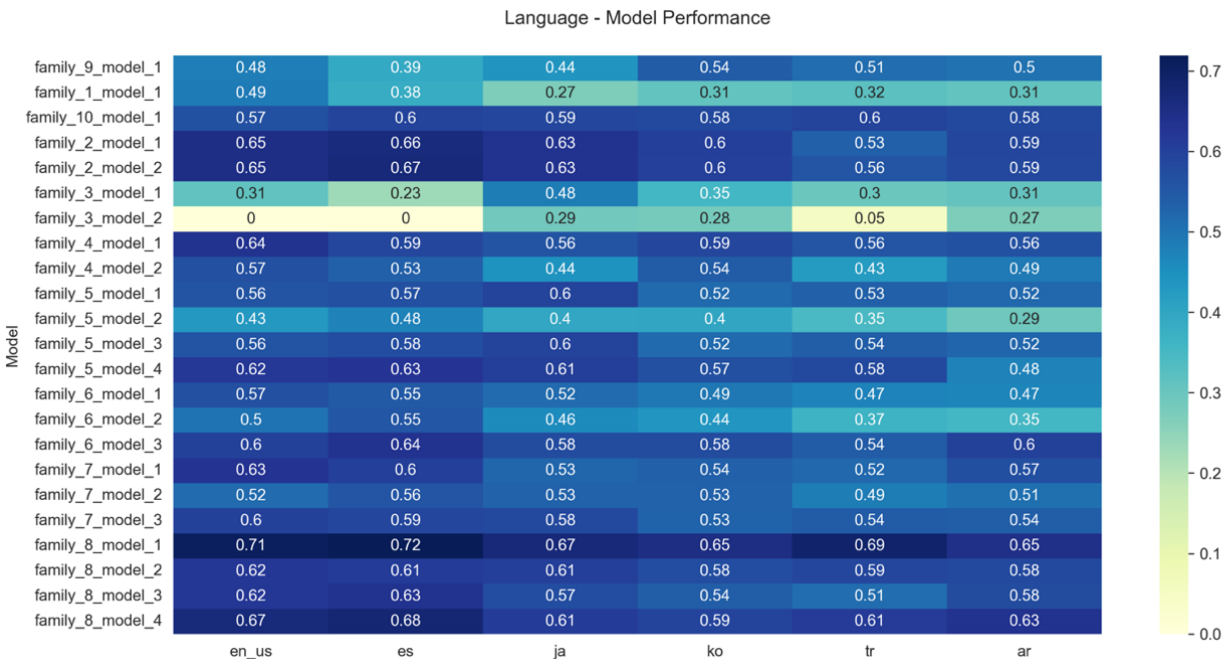
Model Deep-dive

We observe that languages with dominant resource availability and SVO word order (see footnotes 3 and 5 respectively for definitions) have the highest overall performance scores. In the next section, we explore the performance of each language by model and argue that although clear generalizations can be made with respect to word order and resource availability, performance by language is also largely model dependent. We also discuss the variability of scores within a particular model family and explore how model size may impact performance by language. We conclude with further discussion related to resource availability and word order and discuss their implications.

Across all Models

Although there are clear top, middle, and bottom overall scores by language, the generalization that English and Spanish are the top performers, and that Arabic and Korean are the bottom performers is not borne out across all models. Figure 7 below shows the distribution of scores for each language across all models.

Figure 7: Language – Model Performance



Unexpectedly, this visualization shows that Family 3: Model 2 and Family 10: Model 1's lowest scores are in English (at 0 and .57 respectively). These results also show that the Spanish score is often higher than the English score (Family 10: Model 1, Family 2: Model 1, Family 5: Models 1-4, Family 6: Model 3, Family 7: Models 1&2, Family 8: Models 1, 3, & 4). There is no correlation between these observations and model size. Since Spanish and English have the same word order and share some lexical items derived from Latin, we hypothesize that this performance increase is likely related to transfer learning – or learning and

applying representation from one language to another -- from English.¹¹ However, it remains unclear why both Family 3: Model 2 and Family 10: Model 1's lowest scores are in English.

Within Model Families

As described earlier, we anonymize all results by giving the model names pseudonyms. Since important distinctions can be made between model families and models of different size, we group all models from the same developer under the same family. In this section, we review the results of the individual models and compare them to other models within each model family.

In the previous section, we noted that Japanese and Turkish are the two languages with most variability in their scores. Similarly, when we compare individual model performance by language to the overall language performance, we see that Turkish is often the lowest score for some models whereas Japanese can also have very low or very high scores in comparison to the other locales. When we look at Figure 7, these observations tend to be borne out within model families. For example, Family 2: Model 1&2 score the lowest on Turkish with scores of .53 and .56. Family 3: Model 1& 2 both have even lower scores in Turkish (.3 and .005 respectively); however, the smaller model (Model 2) scores even lower in English and Spanish. In contrast, its larger model scores closer to the average (.33) in English and Turkish (.31 and .3 respectively). Family 5: Model 2&4 also score lower than Korean in Turkish but scores are still higher than Arabic. The lowest score for Family 4: Model 2 is also in Turkish followed closely by Japanese with scores of .43 and .44 respectively. Finally, with both Family 3 models, Japanese is actually the highest performing locale.

These observations suggest that, while general observations can be made with respect to language performance, there are some discrepancies within model families. Sometimes a particular model family struggles with one language outside of these generalizations (i.e., Turkish with Family 2 Models), and sometimes a particular model family excels with one of these languages (i.e. Japanese with Family 2 Models). We imagine that this variability is either again a result of training data where there is a general lack of resources for languages with less common word order or complex morphology (see Dataset Design section). Since English, the language with the most resource availability, is not highly inflectional (see footnote 12), models may have a higher likelihood of struggling with morphologically complex languages since training data for these types of language is relatively sparse (in comparison to English).

Linguistic Implications

Although the overall distribution of scores by languages suggests a relationship between resource availability and word order, we also observe that individual model scores vary widely across individual models and even within families. When we compare the performance by language within a model family, it quickly becomes obvious that while some models struggle most with Arabic, others struggle most with Turkish. Turkish has a resource level of 4 and an SOV word order (see Table 13 for comparison to other languages). In terms of overall score distribution, it patterns more similarly with Japanese which has a resource level of 5 and an SOV word order. Although both languages are considered synthetic, Turkish is

¹¹ See Ponti et al., (2019) for an in-depth discussion on the role of transfer learning in the context of natural language processing.

also highly agglutinative. We hypothesize that the training data for models with lower scores in Turkish do not include high volumes of morphologically rich and highly inflectional languages.¹² Furthermore, recent work has shown that there are disparities in the number of bytes needed to convey the same amount of information in different languages, known as the "byte premium" (Arnett et al., 2024, Arnett & Bergen 2024), due to orthographic encoding and linguistic reasons. Morphologically rich languages, often written with non-Latin scripts, require more bytes to be represented in common encoding standards like UTF-8. Additionally, these languages tend to have longer words, which may amplify the byte premium effect. Thus, byte premiums may exacerbate the data scarcity problem, meaning agglutinative languages like Turkish may be trained on effectively less data than it currently seems.

In contrast, we hypothesize that, in general, models with the lowest performance in Arabic result from the language's word order – VSO. According to Dryer (2013), only about 7% of the world's languages have VSO as the dominant order. This contrasts with SVO and SOV word orders which comprise 35% and 40% of the world's languages, respectively. As noted in the Dataset Design Section, English and Spanish are both SVO languages with dominant resource availability (i.e. both level 5). Moreover, the majority of languages with dominant resource availability also have SVO word order (i.e. 5 out of the 7 languages with top resource availability have SVO word order). We hypothesize that the coupling of resource availability and word order affects the performance of the non-SVO languages. It is hence unsurprising that performance in Arabic is so low in comparison to the SVO and SOV languages in our dataset.

Finally, we also hypothesize that for some models, the language's script helps to improve model performance in a particular language. For example, we observe that Family 3 Models all perform best in Japanese and Korean – and very low in English and Spanish. There is a relationship between languages with Latin Script and lower scores. Despite the extreme difference in word order described above, we see that this model family also exhibits similar performance in Arabic. If this model is highly sensitive to the uniqueness of the language's script, then the model might be performing better in languages with non-dominant word order and lower resource availability, because their scripts are not as widely used as Latin Script.

The results of this analysis imply many additional areas for future research. First and foremost, we need to test other high-resource languages with SVO word order to see if the models perform similarly to English and Spanish.¹³ We have also hypothesized here that in addition to resource availability, the dominance of the word order has an impact on model performance with languages demonstrating lower scores if the dominant worder is not SVO. However, it is important to note that resource availability and word order are potential confounding factors, making it challenging to isolate the specific impact of each on model performance. Furthermore, we also note that performance in morphologically complex languages may also be negatively impacted since the high-resource languages tend to have a lower

¹² Morphology, a branch of linguistics, studies word structure and formation, focusing on morphemes—the smallest meaningful language units. Inflectional morphemes modify a word to indicate grammatical information like tense or number (e.g., *-s* in *dogs* or *-ed* in *walked*) without changing its meaning. Synthetic languages use inflection to express grammatical relationships, combining multiple morphemes in words. Agglutinative languages, a subtype of synthetic languages, form words by sequentially combining morphemes, each with a distinct grammatical meaning, remaining unchanged when combined. In contrast, English is largely analytic, using word order and auxiliary words to express grammatical relationships more than inflectional morphemes.

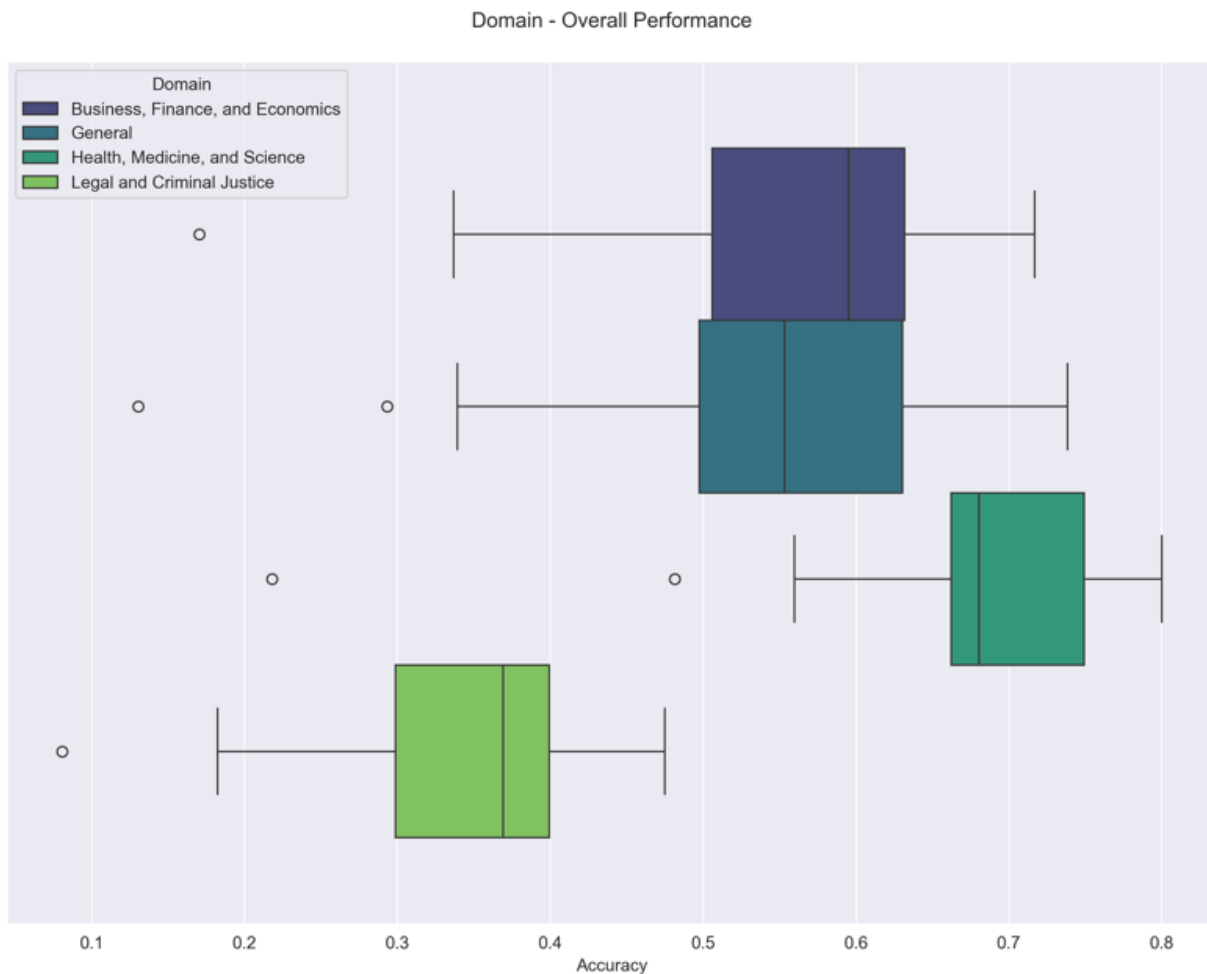
¹³ We have already started collecting data for two of these languages, and we expect to have initial results in these languages at the time of the publication of this paper.

degree of inflectional morphology. We leave these hypotheses to future research, which we aim to explore as we continue to expand our dataset.

Domains

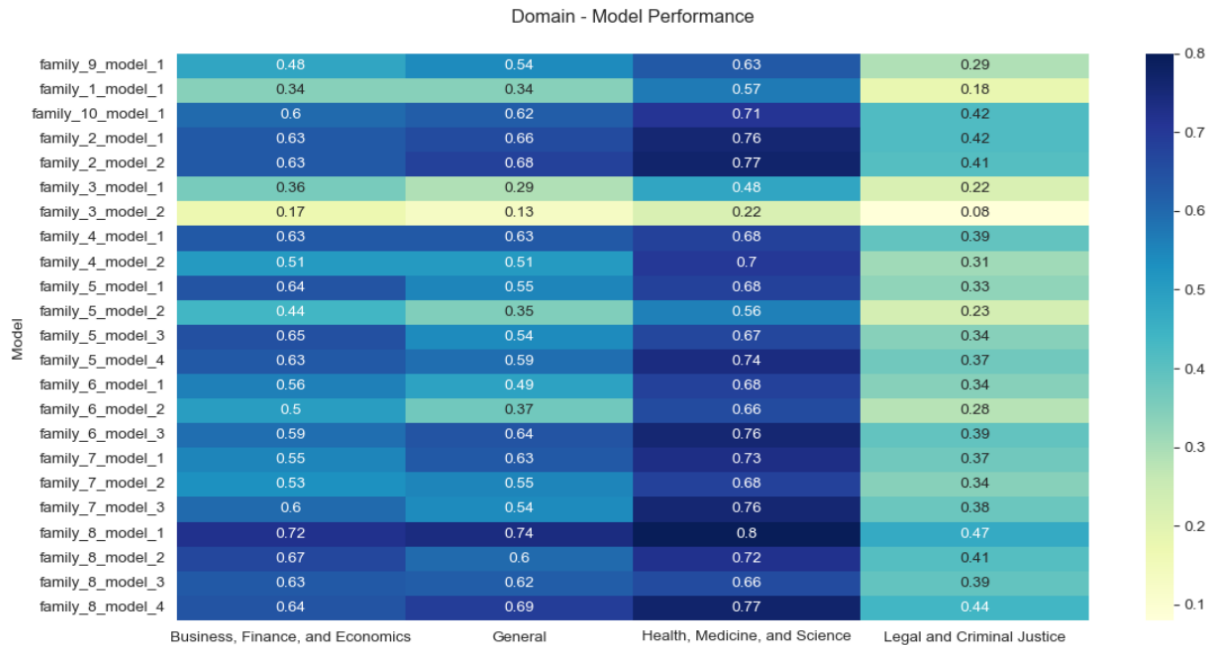
Like the overall performance in each language, we see a similar trend in performance with respect to the domain. Like language, there are clear top and bottom performers in the overall domain scores as well as median performers with extensive interquartile ranges. Figure 8 shows the distribution of scores across each domain.

Figure 8: Domain – Overall Performance



The highest overall accuracy score is in Health, Medicine, & Science (HMS), whereas the lowest score is in Legal & Criminal Justice (LCJ).¹⁴ As shown in Figure 9 below, these generalizations hold for all models; however, the distance between scores for these 2 domains is largely model-dependent.

Figure 9: Domain – Model Performance



Due to the high degree of specialized jargon, we expected both HMS and LCJ to be challenging domains whereas we expected General to be the easiest. This, of course, was not fully borne out as the performance in HMS domain is much higher than all other domains. We hypothesize that the HMS data has distinct jargon and lexical items with relatively straightforward scenarios while the LCJ domain may also have distinct jargon but with more nuanced and ambiguous scenarios. The variation in performance could also reflect differences in the legal systems across the world, which may introduce more complexity and inconsistency in the LCJ domain (see also footnote 14). In contrast, health systems may be more consistent globally, leading to a more uniform representation in the HMS domain. Thus, we expect that identifying causality requires more implicit domain and world knowledge for the LCJ domain than for the HMS domain.

Furthermore, Figure 8 also shows that the distance between the HMS and the General scores is generally very close. The individual model scores are provided in Figure 9 above, which shows that model performance between these domains have differences of $<.1$ (Family 9:Model 1, Family 10: Model 1, Family 2:Model 2, etc.). However, there is a more significant distance in the performance between these domains for other models. For example, Family 6: Model 2 has a difference of almost .3 between the HMS

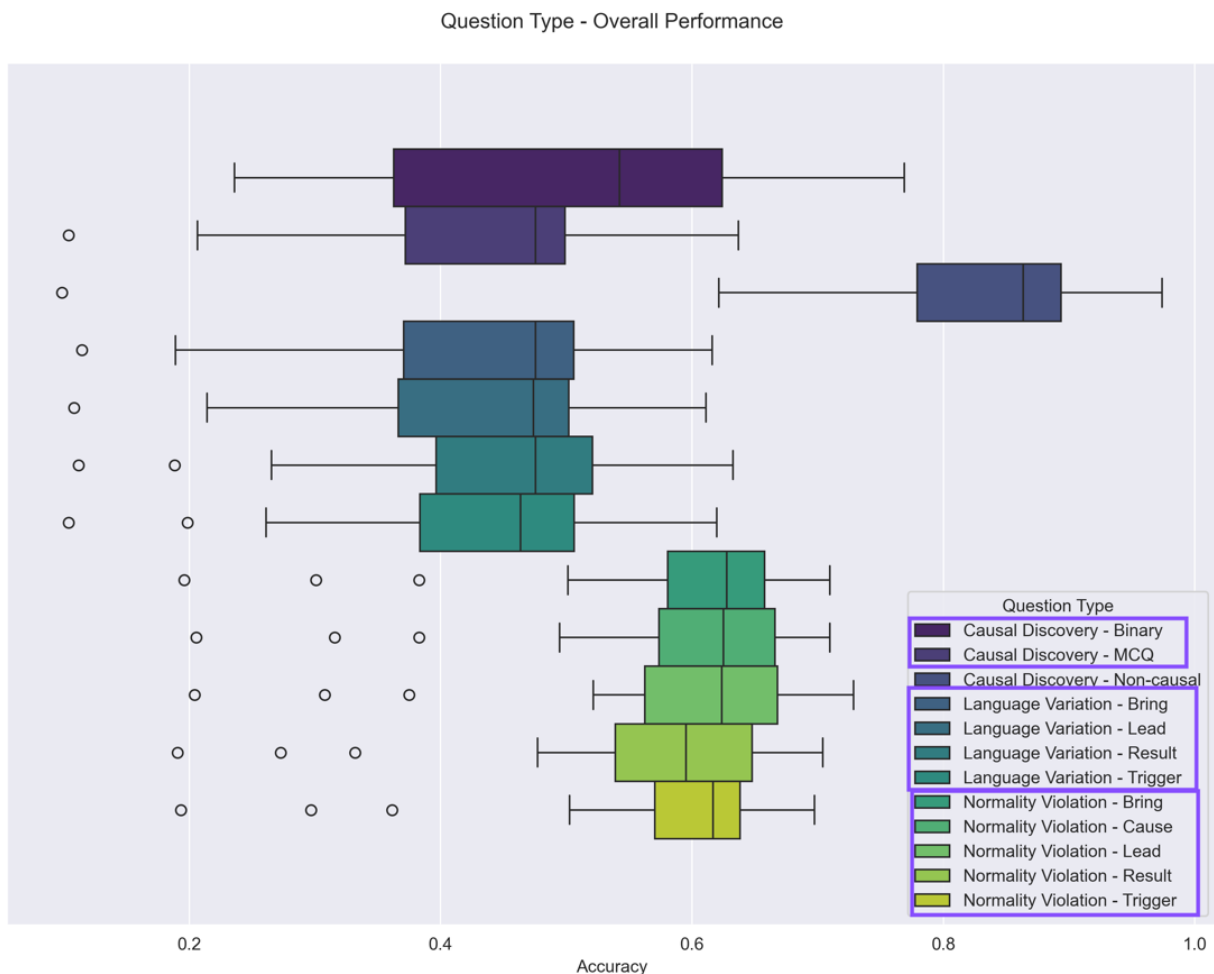
¹⁴ For this initial evaluation, we did not look at the performance of domain by language. Larry Carin (personal communication) correctly points out that legal and justice systems differ worldwide, and the disparities in performance may reflect differences in model responses by language. We intend to investigate this in more detail soon.

domain and the General domain, and this particular model actually has a higher score in Business, Finance, and Economics than in the General domain. Similarly, we also observe that the difference in overall accuracy between HMS and General sometimes appears to be related to size. In Families 4, 5, and 6, the greatest distance between the HMS and General is seen with model 2, which is the smallest model in both families. Given the results by domain, we hypothesize that there is an imbalance in training data from these four domains.

Question Types

We see the most interesting – and diverse – distribution of overall scores when we consider the Question Type. Figure 10 below shows the overall distribution of each of these Question Types. As described earlier, the dataset was designed to essentially have question classes, or a set of question types that leverage similar syntactic constructions or query the same information, from a set of question types. Although Figure 10 clearly demonstrates that Question Types with minor lexical variations exhibit minor differences in performance, other pairs are less evident from this visualization. Hence, we describe each of these in detail below.

Figure 10: Question Type – Overall Performance



In this section, we review the results of the following question pairs and question sets:

Table 14: Comparison of Various Question Types

Question Type 1	Question Type 2	Comparison Type
Causal Discovery – Binary	Causal Discovery – Confounder	Pair
Causal Discovery – MCQ	Causal Discovery – Binary	Pair
Causal Discovery – MCQ	Normality Violation – Cause	Pair
Causal Discovery – MCQ	Language Variation	Set
Language Variation	Normality Violation	Set

We analyze how models perform with each question type and provide generalizations about performance by language according to question type. We reserve further analysis of individual model performance by language for future research.

Causal Discovery - Binary vs. Causal Discovery - Confounder

In this section, we discuss and compare the performance of the two Binary questions: Causal Discovery – Binary and Causal Discovery – Confounder. As defined in Table 9, the binary question evaluates the model’s ability to identify the cause in a binary question format. The Confounder, or Non-Causal, question uses the same format to ask if the confounder is the cause of the outcome.

Evaluating performance on these two question types was motivated by previous research on the model’s ability to identify causes in the presence of a confounder (Kıcıman et al. 2023) and to identify non-causal relationships (Gao et al. 2023). Kıcıman et al., (2023) observe that LLMs struggle with discerning between confounders in causal identification tasks. In Table 15 below, they provide two example outputs from GPT-4.

Table 15: Probing causal reasoning Kıcıman et al., (2023)

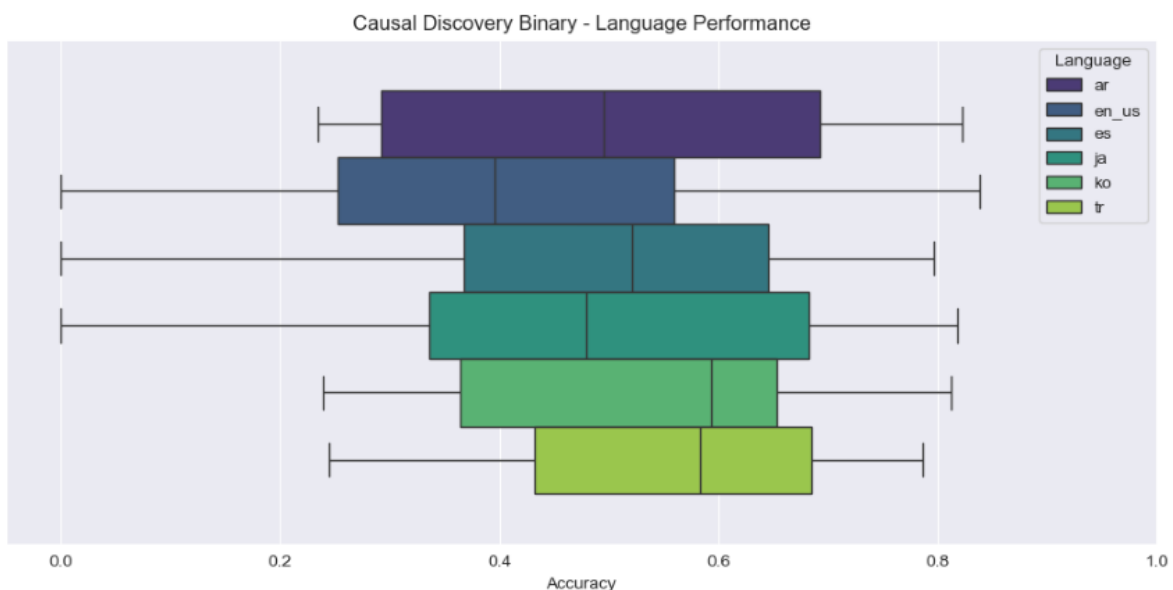
A. A GPT-4 powered advisor gives sound causal advice		B. Containing the conversation, GPT-4 gives a wrong causal answer	
SYSTEM:	You are a business advisor and give clear, well-reasoned Init brief (5-15 line) advice in response to questions about what people should do.	USER:	Thanks. In February, the toy shop owner split their audience into two groups to show them two different ads. They showed the old ad to people on the newspaper; and showed the new ad to people online. When people came into the store, the owner would ask them which ad they saw. In February, they found out that they sold \$6000 of toys to people who saw the new online ad, and \$4500 to people who saw the old newspaper ad. What should the toymaker do in March?
USER:	<p>A toy shop owner in the western USA wants to decide whether the ad they bought in early December is really better than their previous ads. Here is their sales data: October: \$10,200; November: \$10,000; December: \$13,000; January: \$10,100</p> <p>They now want to decide which ad to show in February. Can you help them decide whether the increase in sales in December was due to the ads? Note that the new ad costs \$1,000 more to produce, so the toy maker is interested in maximizing their profit.</p> <p>Based on the available data, it is difficult to definitively attribute the increase in December sales to the new ad, as holiday season and other external factors may have influenced sales. To make a more informed decision, I recommend conducting an A/B test in February by dividing the targeted audience into two groups, showing one group the old ad and the other group the new ad. Compare the sales performance of both groups and calculate the net profit (considering the \$1,000 additional cost for the new ad) to determine which ad is more effective in driving sales and maximizing profit. This will provide a clearer understanding of the ad's impact on sale.</p>	GPT-4:	<p>GPT-4: Based on the February results, the new online ad generated \$1,500 more in sales than the old newspaper ad.</p> <p>Considering the new ad costs \$1,000 more to produce, the net profit increase is 500. In this case, it seems that the new online ad is more effective in driving sales. For March, the toy shop owner should continue using the new online ad, Init also consider monitoring the long-term performance of the ad, as well as any potential changes in the target audience's preferences or external factors that may influence sales. Additionally, the owner may want to explore optimizing the new ad further to maximize its cost-effectiveness and reach.</p>

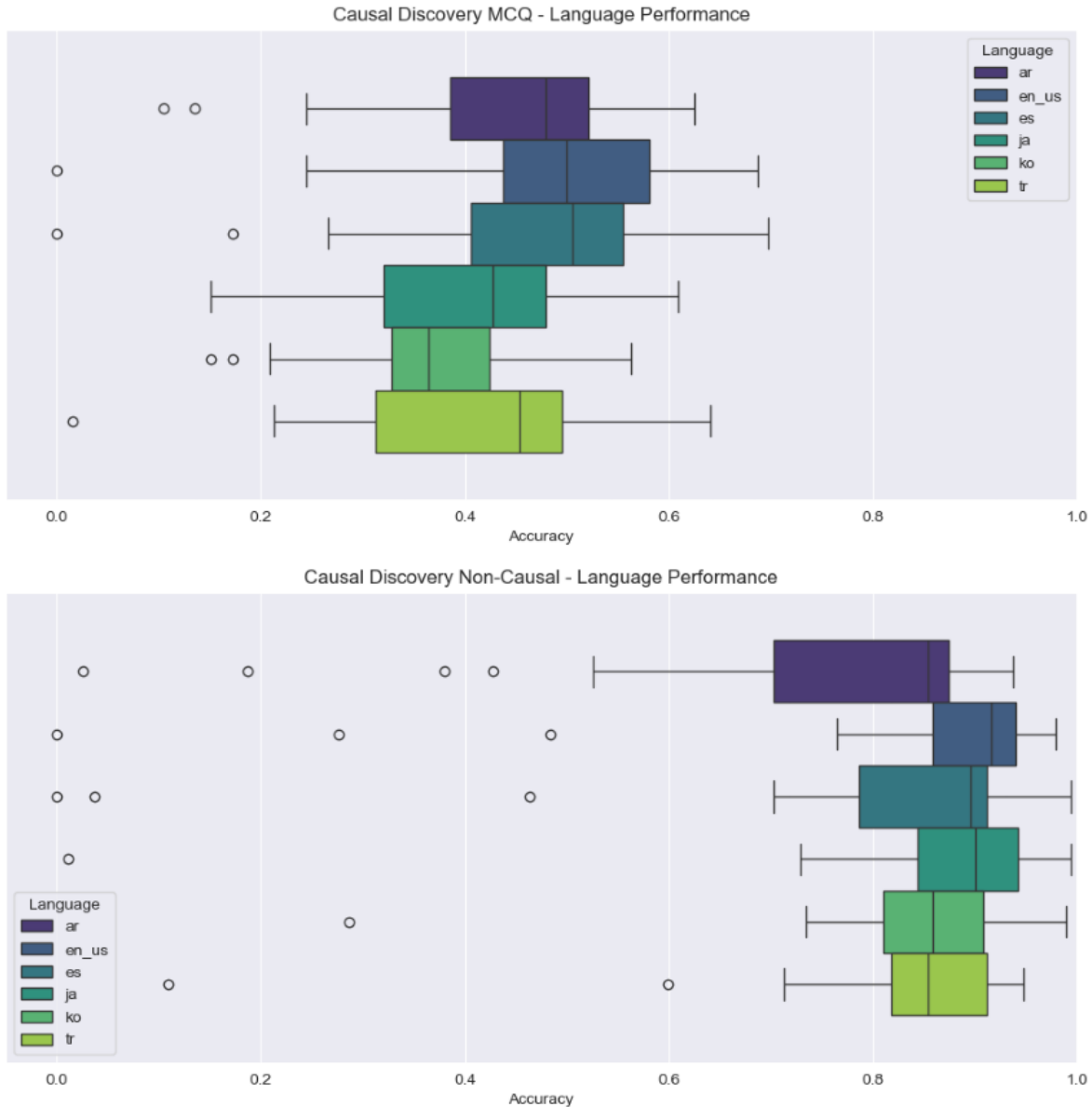
In the first example, the LLM demonstrates an understanding of causal issues by identifying a potential confounder and recommending an A/B experiment to accurately characterize effects and inform decision-making. In the second dialog, the LLM is presented with a similar scenario requiring awareness of potential confounders. Despite the need for the same level of causal reasoning, the LLM does not address these confounders and ultimately provides an incorrect answer. Given these results and the complex nature of our dataset, our domain experts also included confounders in every scenario so we could create binary questions which asks if a causal relationship exists between the confounder and the target outcome for every scenario.

Gao et al., (2023) compared ChatGPT’s ability to identify both causal and non-causal relationships and found that with simple datasets like E-CARE (Du et al. 2022) and COPA, ChatGPT performed better at identifying non-causal relationships. That is, the model was successful in determining that a causal relationship did *not* exist between a pair of events. Interestingly, however, Gao et al., (2023) also found that in more complex causal discovery tasks like MAVEN-ERE (Wang et al., 2022) and Causal-Timebank (Mirza et al., 2014), ChatGPT performed significantly worse when determining if a causal relationship exists between two events that are not causally related.

Our results are most consistent with Gao et al., (2023)’s initial analysis as the Causal Discovery – Confounder question types have the highest scores and the smallest interquartile ranges per language (Figure 11). Performance in English (.82) is consistent with the average across all locales (.81). In Gao et al.’s initial analysis on E-CARE and COPA, the average score across models was approximately .79 and .69 respectively. Although most languages perform around our average of .81, performance in Arabic is significantly worse with an average of .72. This demonstrates a need for additional training data in Arabic to help make the models competitive cross-linguistically.

Figure 11: Question Pair Boxplots





Similar to Gao et al.'s results, we find that the models demonstrate lower performance on Causal Discovery – Binary questions across all languages as well. Notably, the average overall score on this question type is approximately .5 on our dataset whereas in Gao et al.'s analysis, the average score was .58 on E-CARE and .65 on COPA. Interestingly, English has the lowest average score of .4 and Arabic scored .5 – consistent with the average.

Causal Discovery-Binary vs. Causal Discovery – MCQ

Considering the model performance on the Causal Discovery – Binary questions, we observe that most models often perform better in all languages on these question types in comparison to the Causal Discovery – MCQ question types. These results contrast with the Gao et al., (2023)'s multiple choice

evaluations using E-CARE and COPA. In Gao et al.’s study, performance was generally higher on multiple choice questions than on binary questions.

Our evaluation shows that model performance is highest in English, Spanish, and Arabic on Causal Discovery – MCQ questions with average scores of .48, .43, and .4 respectively. For this question type, we also observe there is a wide range of scores across all languages, indicating a high degree of variability in model performance for this question type within each locale. Model performance on Causal Discovery – MCQ is also lowest in Japanese, Korean, and Turkish (.41, .36, .4 respectively) and scores are significantly lower than on the Causal Discovery – Binary questions (.48, .53, .55 respectively). We anticipate that models perform worse on Causal Discovery – MCQ because i) our questions are significantly more complex than the ones in E-CARE and COPA and ii) there is a greater likelihood of selecting an incorrect answer than with a Binary question.¹⁵

Standard Causation vs. Normality Violation

Relatedly, our results indicate that models are better at identifying the Normality Violation that contributed to the outcome than they are at identifying the direct cause of the same outcome. As discussed earlier, causality judgments in the context of event abnormality are evaluated as part of the BigBench dataset (Srivastava et al. 2022). The dataset was designed to elicit model responses to see if the models align with human causality judgments when causal events are abnormal and violate social, moral, or legal norms. Kıcıman et al., (2023). perform a similar evaluation, leveraging short stories from the BigBench-Hard (Suzgun et al. 2022). In this evaluation, Kıcıman et al. asked the LLM to identify the abnormal event in the causal chains (see Table 16 below). They demonstrate that the models infer normality violations in the short stories only about 70% of the time and contend that models successfully infer normality with a higher frequency than they can infer causal judgment.

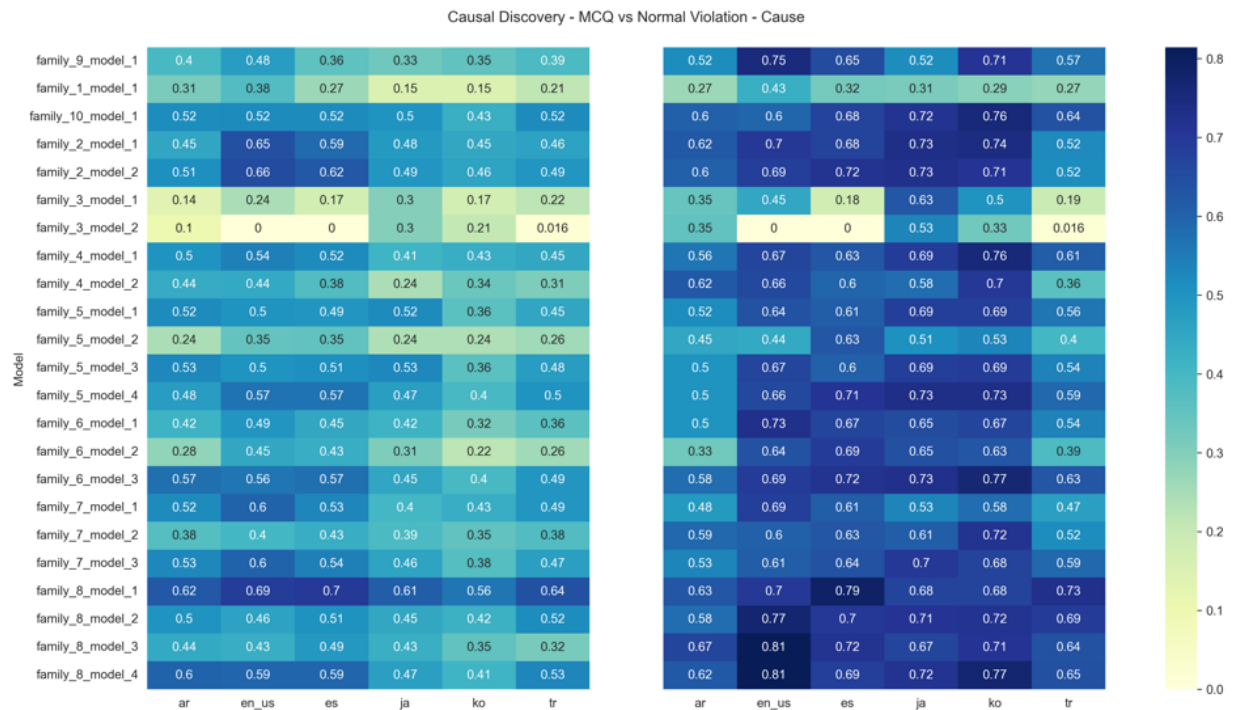
¹⁵ One limitation of this dataset is its heavy bias to Multiple Choice Questions as this also bias top scores in overall accuracy to models which perform better on multiple choice questions than on binary questions – globally and by language. For the sake of simplicity, we do not discuss individual model performance on these questions, but we plan to investigate this in detail in the near future.

Table 16: Comparative Assessments of normality between text-davinci-003 and GPT-4-32k.

Passage and Questions	Correct Answer	Text-davinci-003 rationale	GPT-4-32K rationale
The receptionist in the philosophy department keeps her desk stocked with pens. The administrative assistants are allowed to take the pens, but faculty members are supposed to buy their own. The administrative assistants typically do take the pens. Unfortunately, so do the faculty members. The receptionist has repeatedly emailed them reminders that only administrative assistants are allowed to take the pens. On Monday morning, one of the administrative assistants encounters Professor Smith walking past the receptionist's desk. Both take pens. Later that day the receptionist needs to take an important message... but she has a problem. There are no pens left on her desk. Did the administrative assistant cause the problem?	The administrative assistant taking the pen is a normal event because she followed policy.	Taking pens is expected behavior among both administrative assistants and faculty members according to the rules set by the receptionist. Therefore, the causal event of both the administrative assistant and Professor Smith taking pens is normal. (correct)	Taking a pen from the receptionist's desk is abnormal for the administrative assistant in this situation because they contributed to the problem of the receptionist having no pens left, even though it is typically allowed for them to take pens. However, it is important to note that the faculty member's action of taking a pen is even more abnormal and against the established rules. (incorrect)
Joe was about to go pick up his child from a school in the suburbs. In the parking lot, he stopped to talk with friends. Because of this delay in picking up his son, Joe's neighbor, who waited 15 min for Joe to show up, took the child home in his car. On the way home, the neighbor's car was struck by a drunk driver. Joe's son, the only seriously hurt victim, received severe leg injuries. Did Joe cause his son's injury?	Joe delaying in picking up his son to chat with friends is abnormal because it is a social norm violation.	Stopping to talk with friends in a parking lot is a common and expected behavior, making this causal event quite normal. (incorrect)	Joe stopping to talk with friends in the parking lot is a normal event as people often engage in casual conversations when they encounter acquaintances. It is not unexpected, unlikely, or a violation of social norms. (incorrect)

Our results are relatively consistent with Kiciman et al., (2023), but our average score in English is .63. For our evaluation, the average score on Normality Violation – Cause questions is approximately .59, suggesting a significant decline in performance in languages other than English. Unsurprisingly, the lowest scores are in Arabic and Turkish with .52 and .51 respectively. Figure 12 below shows the individual scores across all locales for all models. As we have seen previously, individual model scores vary greatly, but the trend is consistent across all models and all locales – scores are higher for Normality Violation – Cause than for Causal Discovery – MCQ. Like Kiciman et al., we hypothesize that this task may have higher accuracy due to the saliency of the abnormal event.¹⁶

Figure 12: Causal Discovery and Normality Violation Heatmaps



Although our results are consistent with Kiciman et al., (2023)’s findings, we also demonstrate that the generalization can be extended to other languages. As observed in the previous section on the overall language results, the lowest scores – which are only slightly higher than chance – are in the two languages that various models struggle with the most: Arabic and Turkish. As discussed previously, we hypothesize that the low scores in Arabic and Turkish are due to a variety of factors, including non-dominant word orders, morphological complexity, and scripts that are highly prevalent in the training data. We argue that there is an urgent need for more linguistically diverse training data in languages with these linguistic and orthographic characteristics to help fine tune models in widely spoken languages with features not

¹⁶ We also observe that in the Normality Violation scenarios, some models would overgeneralize the abnormal event to be the direct cause of the outcome. We leave quantification of this error type to future research where we will investigate which models prefer to label the normality violation as the direct cause – and how frequently they do so.

exhibited by the most representative linguistic features in the training data. Additional training data and fine tuning in languages like Arabic and Turkish would help increase model accuracy on these questions¹⁷ and would likely improve user engagement, making the model more usable in general and specialized applications.

Language Variation

As discussed in the Dataset Design section, we also tested the model’s causal discovery capabilities in the context of language variation. Like Gao et al, we test model accuracy in multiple choice questions with semantically weaker – yet pragmatically equivalent -- variations of the same verb. We specifically chose the verbs *bring*, *lead*, *result*, and *trigger* for these question types since all four of these verbs are lexically distinct but semantically equivalent in context across all languages in our dataset.

As shown earlier in Figure 5, there is not a significant range of scores for the Language Variation questions that test the model’s ability to identify the direct cause. Figure 13 below shows individual model performance across the Causal Discovery – MCQ and Language Variation question types. As expected – and consistent with other analyses described in this paper – the score variability is largely dependent on the model. Moreover, for this group of questions, scores are lowest in Korean (as shown in Figure 14 below).

We also evaluate these verbs in the context of the Normality Violation questions. Since we were evaluating the use of the same verbs in a specific type of causal discovery question (i.e. identifying the abnormal event that contributed to the outcome), we anticipated that performance between Normality Violation – Cause and the Normality Violation Language Variation questions would be parallel – or at least consistent – with the trends we observed between the Causal Discovery - MCQ and Language Variation questions. This is generally borne out, but there are some language-specific differences.

As described in the previous two sections, performance on Normality Violation questions is generally much higher than on Causal Discovery questions, and performance on Causal Discovery – MCQ and Language Variation questions is relatively comparable. In general, these results are consistent with the Normality Violation and Language Variation questions. Figure 14 below shows that Normality Violation Language Variation questions have higher scores, on average, than their causal discovery counterpart.¹⁸ In contrast to the Language Variation scores, though, the average scores in Korean are quite high – with overall *highest* performance on Normality Violation – Result and Normality Violation -Trigger questions.

¹⁷In the Complexity in Causal Reasoning section, we discussed the Causal Judgment task from the BigBench dataset (Srivastava et al. 2022) and its aim to compare model and human judgments in context-reliant scenarios. Nie et al., (2023) later used a modified version of this dataset – MoCA -- to evaluate model alignment with human responses, focusing on sensitivity to abnormality and preference for labeling statistical or prescriptive norm violations as causes of events. While humans typically cite prescriptive norm violations, models vary in their preferences, aligning with our observations in footnote 16.

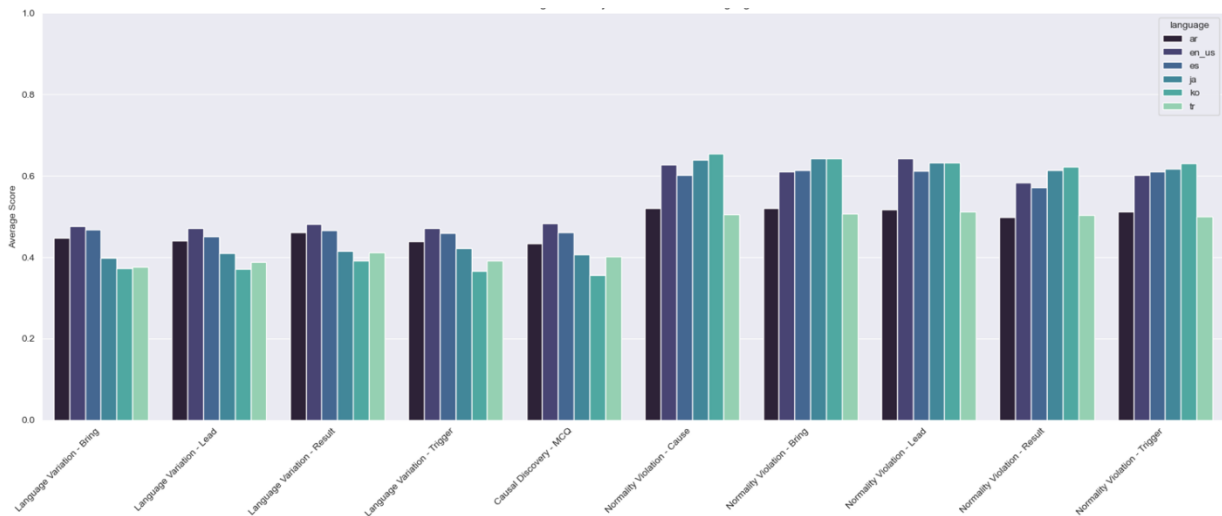
¹⁸ Interestingly, Figure 10 shows that the interquartile ranges for the Causal Discovery and Normality Violation Language Variation questions are similar, except for the Result questions. The interquartile range scores are higher for Language Variation – Result and lower for Normality Violation – Result. Further investigation is needed, but we hypothesize that models are more sensitive to temporality with *Result*, despite context in the prompt instructions.

Consistent with Normality Violation – Cause, overall performance on Normality Violation Language Variation questions is lowest in Arabic and Turkish.

Figure 13: Causal Discovery and Language Variation Heatmap



Figure 14: Causal Discovery, Language Variation, and Normality Violation by Language



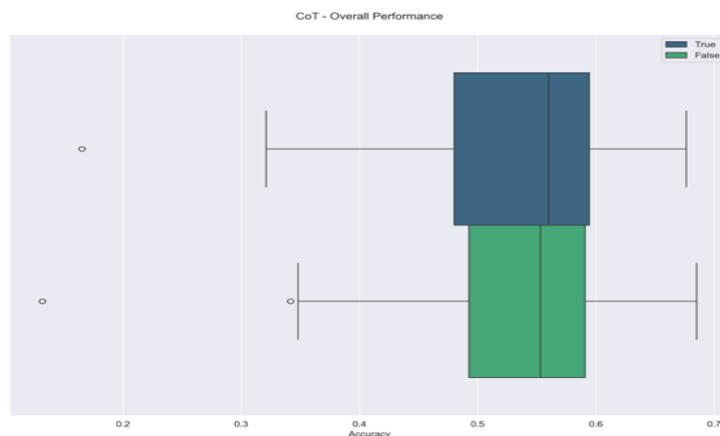
Summary

Our results reveal significant insights into the performance variability of models across different languages and question formats. Our findings indicate that while models generally exhibit strong performance on Causal Discovery – Confounder questions, they struggle more with Causal Discovery – Binary and Causal Discovery – Multiple Choice questions, particularly in languages such as Arabic and Turkish. This underscores the need for more linguistically diverse training data to address the unique challenges posed by morphological complexity and non-dominant word orders in these languages. Additionally, our results align with Kıcıman et al., (2023)’s observations on the importance of normality violations in causal inference, showing that models perform better when identifying abnormal events contributing to outcomes. However, performance still varies significantly by language, with higher scores in English and lower scores in languages like Arabic and Turkish. Finally, the evaluation of language variation questions demonstrates that while models can handle semantic equivalence to some extent, there remains substantial room for improvement, especially in less commonly represented languages. These insights collectively highlight the necessity for ongoing model refinement and targeted training to enhance cross-linguistic performance and ensure more accurate and reliable causal reasoning across diverse linguistic contexts.

Chain of Thought (CoT) Prompting

Finally, since our prompt design easily allowed us to embed additional instructions into each prompt, we also evaluate the model’s performance with and without Chain of Thought (CoT) prompting (see footnote 8 for definition). The goal is to understand if CoT prompting improves performance with long, complex prompts. While studies such as those by Wei et al., (2022) and Yao et al., (2023) suggest that CoT reasoning techniques can enhance model performance, others like Zhang et al., (2023) have shown that the reasoning can sometimes be incorrect.

Figure 15: CoT Prompting – Overall Performance Boxplot



Most surprisingly, our results show that CoT prompting did not significantly improve overall performance across all models. The visualization in Figure 15 above shows a very small median performance increase with the inclusion of CoT prompting (indicated as *True* in visualization). However, comparison of the two plots also shows that prompts *without* CoT prompting have both higher minimum and maximum values. There is also a greater interquartile range for prompts that include the CoT instructions.

Although there is not a significant difference in performance with or without CoT prompting for most models, Figure 16 below shows that there is disparate performance in this category. For example, some models show significant performance improvements with CoT prompting (Table 17 below), but others show significant decreases in performance (Table 18 below).

Figure 16: CoT Prompting – Overall Performance by Model

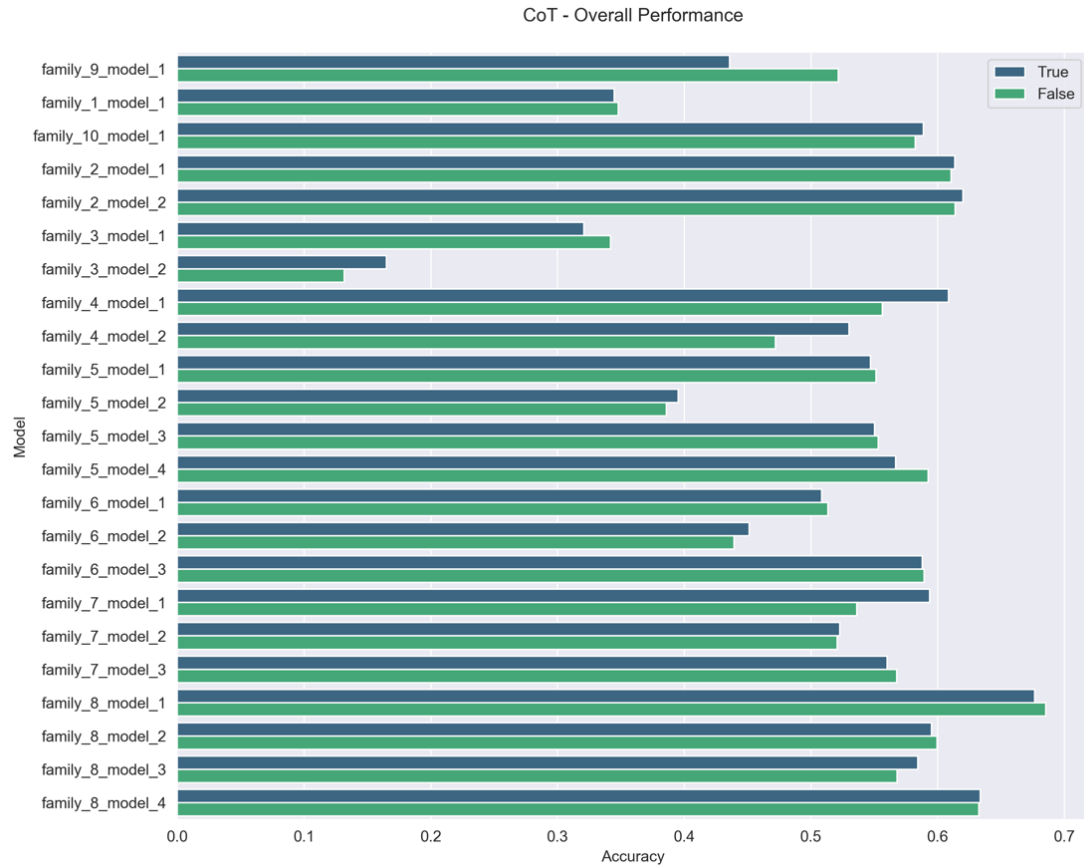


Table 17: Performance Increase

Family	Model	Size
3	2	Small
4	1	Large
4	2	Medium

Table 18: Performance Decrease

Family	Model	Size
3	1	Large
5	4	Large
6	2	Small
9	1	Medium

Table 17 and Table 18 (previous page) show no obvious or immediate generalizations can be made with respect to model family or model size. It can be observed that Family 3's small model (2) had a performance increase whereas its large model (1) had a performance decrease, and it can be observed that both of Family 4's models had a performance increase. However, further research is needed to better understand the results to determine if CoT prompting is further impacted by language, domain, or even question type.

Note that all prompts include instructions to provide the final answer in a JSON dictionary. With CoT prompts, many models also provide text describing the rationale or reasoning for their provided answer. Paul et al. (2024) examine how LLMs utilize intermediate reasoning steps when generating a final answer, which is not considered in our evaluation since our scoring methodology only considers the final answer. In their study, Paul et al., (2024) show that the model's final response is not always aligned with their provided reasoning description. Similarly, we observe instances where the text describing the step-by-step reasoning does not align with the final answer. Interestingly, we have also observed responses where the CoT prompt final answer matches the non-CoT prompt final answer although the step-by-step reasoning description should lead the LLM to a different final answer. We provide some examples of this in Appendix A.

For this particular project, we do not calculate how frequently the step-by-step reasoning and the final answer are contradictory. We also leave this to future research as we are also interested in determining how frequently the model provides the same final answer as the non-CoT prompt in their reasoning process despite a different reasoning process. We are interested to learn more about contexts which bias the model to a final answer that does not match the reasoning process.

Results Summary

The results of our evaluation reveal significant insights into multilingual causality across various parameters such as languages, domains, causal tasks, and question types. Our primary findings indicate substantial accuracy differences by locale and domain, challenges in causal discovery, language variations, and the effects of Chain of Thought (CoT) prompting. Performance varies significantly by language, with English and Spanish leading, while Japanese and Turkish exhibit high variability. Moreover, some models particularly struggle with Turkish and Arabic, likely due to word order and morphological complexity. Domain analysis shows the highest accuracy in Health, Medicine, & Science (HMS) and the lowest in Legal & Criminal Justice (LCJ), suggesting the impact of domain-specific jargon and scenario complexity. Additionally, question type analysis highlights that binary causal discovery questions perform better than multiple-choice questions, with significant variability across languages and models. Finally, CoT prompting does not universally improve performance, with mixed results across models and no clear correlation with model size or family. These findings underscore the need for further research into language-specific training data and the potential effects of linguistic and domain-specific complexities on model performance in evaluating the causal reasoning capabilities of LLMs.

Evaluation Implications

The implications of our evaluation demonstrate a clear and urgent need for more comprehensive causal reasoning benchmarks for LLMs. We demonstrate here that although LLMs are successful in some causal tasks in English, model performance declines with additional complexity of prompts either because of domain complexity or linguistic complexity. Crucially, our results show that models struggle with this capability in languages which do not have the dominant word order (i.e. Subject-Object-Verb (SVO) languages, see footnote 7) of the dominant resource languages. These results suggest that there is an urgent need for more nuanced and complex training data in the domain of causal reasoning to fine tune models on this capability. Ultimately, greater volumes of cross-linguistic training data are needed to fully support these models in these capabilities, which will in turn, improve user engagement, generating more training data, and bringing us closer to the advent of AGI.

Conclusion

We have addressed various gaps and challenges in benchmarking LLM causal reasoning capabilities, proposed solutions for a robust and comprehensive multilingual causal reasoning evaluation, and reviewed results from our initial evaluation. We have described existing benchmarking datasets and addressed their various shortcomings with respect to complexity and linguistic diversity and shown that existing evaluations do not address multilingual causal reasoning capabilities. As part of our dataset design, we leveraged human experts to craft novel, fact-based scenarios, subjective stories based on the scenarios, and question and answer pairs. We then compiled the stories and questions into prompts and evaluated model accuracy and consistency of responses in English, Spanish, Japanese, Korean, Turkish, and Standard Arabic. Our results show these models struggle with consistent causal reasoning in English with novel, complex prompts. Furthermore, the results demonstrate an urgent need to evaluate causal reasoning in other languages since multilingual model performance typically declines with linguistic complexity. We argue that additional multilingual causal reasoning training data is an emerging need – and is necessary for fine-tuning models on causal reasoning capabilities.

References

- Arnett, Catherine and Benjamin Bergen. “Why do language models perform worse for morphologically complex languages?” arXiv.org, November 21, 2024. <https://arxiv.org/abs/2411.14198>
- Arnett, Catherine, Tyler A. Chang, and Benjamin Bergen. 2024. A bit of a problem: Measurement disparities in dataset sizes across languages. In *Proceedings of the 3rd Annual Meeting of the Special Interest Group on Under-resourced Languages @ LREC-COLING 2024*, (pp 1–9), Torino, Italia. ELRA and ICCL.
- Bender, Emily, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. “On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?” March 1, 2021. <https://dl.acm.org/doi/pdf/10.1145/3442188.3445922>.
- Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, Dario Amodei. “Language Models are Few-Shot Learners.” arXiv.org, July 22, 2020. <https://arxiv.org/pdf/2005.14165>
- Conneau, Alex, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, (pp. 8440–8451), Association for Computational Linguistics. <https://aclanthology.org/2020.acl-main.747/>
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, (pp. 4171–4186), Association for Computational Linguistics. <https://aclanthology.org/N19-1423/>
- Matthew S. Dryer. 2013. Order of Subject, Object and Verb. In: Dryer, Matthew S. & Haspelmath, Martin (eds.) *WALS Online* (v2020.4) <http://wals.info/chapter/81>.
- Li Du, Xiao Ding, Kai Xiong, Ting Liu, Bing Qin. 2022. e-CARE: A new dataset for exploring explainable causal reasoning. In S. Muresan, P. Nakov, & A. Villavicencio (Eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 432–446). Association for Computational Linguistics. <https://aclanthology.org/2022.acl-long.3333>
- Gao, Jinglong, Xiao Ding, Bing Qin, and Ting Liu. “Is ChatGPT a good causal reasoner? A comprehensive evaluation.” arXiv.org, May 12, 2023. <https://arxiv.org/abs/2305.07375>
- Hendrycks, Dan, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. “Measuring Massive Multitask Language Understanding.” arXiv.org, January 12, 2021. <https://arxiv.org/abs/2009.03300>.

Joshi, P., Santy, S., Budhiraja, A., Bali, K., & Choudhury, M. (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 6282–6293). Association for Computational Linguistics. <https://aclanthology.org/2020.acl-main.560/>

Kıcıman, Emre, Robert Ness, Amit Sharma, and Chenhao Tan. “Causal Reasoning and Large Language Models: Opening a New Frontier for Causality.” arXiv, August 20, 2024. <https://arxiv.org/abs/2305.00050>.

Knobe, Joshua. 2003. Intentional action and side effects in ordinary language. *Analysis* **63**(279):190–194. doi.org/10.1111/1467-8284.00419.

Paramita Mirza, Rachele Sprugnoli, Sara Tonelli, and Manuela Speranza. 2014. “Annotating causality in the TempEval-3 corpus.” In *Proc. of EACL*.

Mooij, Joris, Jonas Peters, Dominik Janzing, Jakob Zscheischler, and Bernhard Schölkopf. “Distinguishing Cause from Effect Using Observational Data: Methods and Benchmarks.” *Journal of Machine Learning Research* 17, no. 32 (2016). <https://jmlr.org/papers/v17/14-518.html>.

Mumuni, Alhassan & Mumuni, Fuseini. 2025. Large language models for artificial general intelligence (AGI): A survey of foundational principles and approaches. arXiv, January 6, 2025. <https://arxiv.org/html/2501.03151v1>

Nie, Allen, Yuhui Zhang, Atharva Amdekar, Chris Piech, Tatsunori Hashimoto, and Tobias Gerstenberg. “MOCA: Measuring Human-Language Model Alignment on Causal and Moral Judgment Tasks.” arXiv.org, October 31, 2023. <https://arxiv.org/abs/2310.19677>.

Paul, Debjit, Robert West, Antoine Bosselut, and Boi Faltings. “Making Reasoning Matter: Measuring and Improving Faithfulness of Chain-of-Thought Reasoning.” arXiv.org, October 6, 2024. <https://arxiv.org/html/2402.13950>.

Ponti, Edoardo, Helen O’Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, Anna Korhonen. “Modeling Language Variation and Universals: A Survey on Typological Linguistics for Natural Language Processing.” ACL Anthology, September 2019. <https://aclanthology.org/J19-3005/>

Ponti, Edoardo, Goran Glavas, Olga Majewska, Qianchu Liu, Ivan Vulic, and Anna Korhonen. “XCOPA: A Multilingual Dataset for Causal Commonsense Reasoning”. May 1, 2020. <https://ducdauge.github.io/files/xcopa.pdf>

Ribeiro, Marco Tulio, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. “Beyond Accuracy: Behavioral Testing of NLP Models with Checklist.” ACL Anthology, July 2020. <https://aclanthology.org/2020.acl-main.442/>.

Roemmele, Melissa, Cosmin Adrian Bejan, and Andrew S. Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *Proceedings of the 2011 AAAI Spring Symposium Series*.

Romanou, Angelika, Syrielle Montariol, Debjit Paul, Leo Laugier, Karl Aberer, and Antoine Bosselut. “Crab: Assessing the Strength of Causal Relationships between Real-World Events.” arXiv.org, November 7, 2023. <https://arxiv.org/abs/2311.04284>.

Srivastava, Aarohi, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown et al. “Beyond the imitation game: Quantifying and extrapolating the capabilities of language models.” *arXiv preprint arXiv:2206.04615* (2022).

Suzgun, Mirac, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, Jason Wei. “Challenging BIG-Bench tasks and whether chain-of-thought can solve them.” arXiv.org, October 17, 2022. <https://arxiv.org/abs/2210.09261>

Tu, Ruibo, Kun Zhang, Bo Christer Bertilson, Hedvig Kjellström, and Cheng Zhang. “Neuropathic Pain Diagnosis Simulator for Causal Discovery Algorithm Evaluation.” arXiv.org, October 28, 2019. <https://arxiv.org/abs/1906.01732>.

Xiaozhi Wang, Yulin Chen, Ning Ding, Hao Peng, Zimu Wang, Yankai Lin, Xu Han, Lei Hou, Juanzi Li, Zhiyuan Liu, Peng Li, and Jie Zhou. 2022. Maven-ere: A unified large-scale dataset for event coreference, temporal, causal, and subevent relation extraction. In *Proceedings of EMNLP*.

Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.” arXiv.org, January 28, 2022 <https://arxiv.org/abs/2201.11903>

Yang, Yinfei, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, Ray Kurzweil. “Multilingual universal sentence encoder for semantic retrieval. arXiv.org, July 9, 2019. <https://arxiv.org/abs/1907.04307>

Linying Yang, Vik Shirvaikar, Oscar Clivio, and Fabian Falck 2024. “A Critical Review of Causal Reasoning Benchmarks for Large Language Models”. arXiv.org, July 10, 2024. <https://arxiv.org/abs/2407.08029>

Yao, Yao, Zuchao Li, and Hai Zhao. “Beyond Chain-of-Thought, Effective Graph-of-Thought Reasoning in Language Models”. arXiv.org, May 26, 2023. <https://arxiv.org/abs/2305.16582>

Zečević, Matej, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. “Causal Parrots: Large Language Models May Talk Causality but Are Not Causal.” arXiv.org, August 24, 2023. <https://arxiv.org/abs/2308.13067>

Yong, Zhing Xin, Cristina Menghini, Stephen H. Bach. “Low-Resource Languages Jailbreak GPT-4.” arXiv.org, October 23, 2023. <https://arxiv.org/abs/2310.02446>

Zhang, Cheng, Stefan Bauer, Paul Bennett, Jiangfeng Gao, Wenbo Gong, Agrin Hilmkil, Joel Jennings, Chao Ma, Tom Minka, Nick Pawlowski, and James Vaughan. “Understanding Causality with Large Language Models: Feasibility and Opportunities.” arXiv.org, April 11, 2023. <https://arxiv.org/abs/2304.05524>

Appendix A: Chain of Thought (CoT) Inconsistencies

The following example responses are from the same model and are responses to the same question and story pair. The examples are to demonstrate how CoT prompting affects the model response. As described earlier, CoT Prompting can yield the incorrect response (Table 19); the correct response with incorrect reasoning (Table 20 below); and the correct response (Table 21 below). Based on our evaluation results we argue that CoT prompting is generally not effective for long, complex causal reasoning prompts as it does not significantly increase accuracy.

For this evaluation, we tested CoT prompting on an exploratory basis. In the future, we plan to analyze the discrepancies between prompts with CoT and those without in more detail. At this time, we only observe that the effects of CoT prompting are heavily model-dependent. In future research, we aim to do more granular analysis of the results and/or model testing to better understand patterns of both degradation and improvement.

Table 19: CoT Prompting – Incorrect Response

No CoT Prompting: <i>Correct</i>	With CoT Prompting: <i>Incorrect</i>
<p>Based on the story, the normality violation that triggered Yuki to want to sue Fran and NextGen Hospital was that Fran omitted inquiring Yuki which foot should be amputated. This is evident from the following details in the story:</p> <ul style="list-style-type: none"> - Yuki was scheduled for a right foot amputation, but the surgery was performed on his left foot instead. - NextGen Hospital requires Fran and Sam to ask the patient which body part is subject to the operation, but Fran only asked if Yuki had any questions. - Fran marked an "X" on Yuki's left foot, indicating that the left foot was the intended surgical site, which contradicts the planned right foot amputation. <p>Therefore, the correct answer is:</p> <p><code>{"correct_answer": ["h"]}</code></p>	<p>To determine the normality violations that triggered Yuki to want to sue Fran and NextGen Hospital, we need to analyze the cause-and-effect relationships in the story.</p> <p>The key events are:</p> <ol style="list-style-type: none"> 1. Yuki was scheduled for a right foot amputation (b). 2. Fran marked an "X" on Yuki's left foot (e). 3. Yuki underwent a left foot amputation (g). 4. Yuki woke up with her left foot in pain and bandaged, while her right foot was still intact (g). <p>The normality violation that directly led to the outcome was Fran performing the amputation on the wrong foot (g). This is a clear violation of the standard medical practice of performing the correct surgical procedure on the intended body part.</p> <p>Additionally, Fran's omission of inquiring which foot should be amputated (h) also contributed to the normality violation, as this is a standard safety check that should have been performed.</p> <p>Therefore, the correct answer is:</p> <p><code>{"correct_answer": ["g", "h"]}</code></p>

Table 20: CoT Prompting – Correct Response with Incorrect Reasoning

No CoT Prompting: <i>Correct</i>	With CoT Prompting: <i>Correct but Wrong Reasoning</i>
<p>Based on the story, the normality violation that led Yuki to want to sue Fran and NextGen Hospital was that Fran omitted inquiring Yuki which foot should be amputated. This is a clear violation of the standard medical procedure of verifying the correct body part to operate on with the patient.</p> <p>{"correct_answer": ["h"]}</p>	<p>To determine the normality violations that led Yuki to want to sue Fran and NextGen Hospital, I will go through the story step-by-step and identify the key events and their causal relationships.</p> <ol style="list-style-type: none"> 1. Yuki was scheduled for a right foot amputation (option b). 2. Fran, the surgeon, marked an "X" on Yuki's left foot (option e). 3. Fran omitted inquiring Yuki which foot should be amputated (option h). 4. Yuki underwent a left foot amputation (option g). 5. Yuki woke up with her left foot in pain and bandaged, while her right foot was still intact (from the story). 6. Yuki was struck with panic and disbelief, and immediately called her lawyer to prepare a medical malpractice lawsuit against Fran and NextGen Hospital (from the story). <p><u>The key normality violation that led to Yuki wanting to sue was that Fran amputated the wrong foot (option g), which was a clear medical error. This was caused by Fran's omission of confirming which foot should be amputated (option h).</u></p> <p>Therefore, the correct answer is:</p> <p>{"correct_answer": ["h"]}</p>

Table 21: CoT Prompting – Correct Response

No CoT Prompting: <i>Correct</i>	With CoT Prompting: <i>Correct</i>
<p>Based on the story, the normality violation that most significantly contributed to Yuki wanting to sue Fran and NextGen Hospital was:</p> <pre>{"correct_answer": ["h. Fran omitted inquiring Yuki which foot should be amputated."]}</pre> <p>The story clearly states that Fran did not follow the hospital's standard operating procedure, which required him to confirm with Yuki which foot he would amputate. Instead, he only explained the surgical procedure and incorrectly marked her left foot for amputation, leading to the wrong foot being removed. This omission by Fran was a critical normality violation that directly caused the adverse outcome for Yuki.</p>	<p>Based on the story, the normality violation that most likely led Yuki to want to sue Fran and NexGen Hospital is:</p> <pre>{"correct_answer": ["h. Fran omitted inquiring Yuki which foot should be amputated."]}</pre> <p>Reasoning process step-by-step:</p> <ol style="list-style-type: none"> 1. The story indicates that the hospital had a standard operating procedure that required the surgeon, Fran, to confirm with the patient, Yuki, which foot would be amputated. 2. However, Fran did not follow this procedure and instead only explained the surgical procedure to Yuki and asked if she had any questions, while incorrectly marking her left foot for amputation. 3. This omission by Fran to confirm with Yuki which foot should be amputated is a clear violation of the hospital's standard operating procedure and a normality violation that directly led to the wrong foot being amputated. 4. The other options, while potentially contributing factors, do not represent the most significant normality violation that directly caused the outcome.

Acknowledgments

Thank you, Larry Carin, for advising on every aspect of this project – from strategy to design to execution.

Thank you to the amazingly talented folks at Welo Data, including Deana Jackson, David Harper, Paul Carr, Siobhan Hanna, Sarha Mavrakis, Canute Haroldson, Carlos Hernández García, Randy Knebel, Waldo Fattore, Carme Tronchoni, Aaron Schliem, Emilia Mateiciuc, Amanda Martín-Alcaide García, Fabian Enrique Rojas Gonzalez, Kevin Tong, and Olesia Khrapunova.

In memory of Esteban Pereyra – thank you for all of your contributions and your continuous positivity through your time on this project and at Welo Data.