Procesamiento del Lenguaje Natural, Revista nº 74, marzo de 2025, pp. 363-382 recibido 05-12-2024 revisado 04-02-2025 aceptado 14-02-2025

Interventional and Counterfactual Causal Reasoning for LLM-based AI Agents: A Dataset and Evaluation in Portuguese

Razonamiento Causal Intervencional y Contrafactual para Agentes de IA Basados en LLM: Un Conjunto de Datos y Evaluación en Portugués

Uriel Lasheras, Elioenai Alves, Vladia Pinheiro
Postgraduate Program in Applied Informatics, University of Fortaleza, Ceará, Brazil
uriel_andersonol@edu.unifor.br, l.oenaialves@edu.unifor.br, vladiacelia@unifor.br

Abstract: Large Language Models (LLMs) are increasingly central to advancements in generative AI across various domains. While some view these models as a potential step toward artificial general intelligence, their capacity to perform complex causal reasoning remains unverified. Causal reasoning, particularly at Pearl's interventional and counterfactual levels, is critical for achieving true general intelligence. In this study, we propose a causal reasoning framework that includes a three-axis taxonomy for causality, designed to capture the intent, action requirements, and the three rungs of causality as defined by Pearl: associational, interventional, and counterfactual; and a human-in-the-loop approach to generate golden collections of natural causal questions, annotated according to the proposal taxonomy. We evaluated the seed questions of a golden collection in Portuguese using the LLM GPT-4o and Llama3.1 with two prompt strategies. Our findings reveal that both LLMs face significant challenges in addressing interventional and counterfactual causal queries. These results suggest limitations in the indiscriminate use of these LLMs for extending annotation to additional natural questions or for developing LLM-based causal AI agents.
Keywords: Causal Reasoning, Interventional Reasoning, Counterfactual Reasoning, Large Language Models.

Resumen: Los Modelos de Lenguaje Extensos (LLMs, por sus siglas en inglés) están cada vez más al centro de los avances en la IA generativa en diversos dominios. Aunque algunos consideran que estos modelos representan un posible paso hacia la inteligencia general artificial, su capacidad para realizar razonamientos causales complejos sigue sin estar verificada. El razonamiento causal, especialmente en los niveles de intervención y contrafactual propuestos por Pearl, es fundamental para alcanzar una inteligencia general auténtica. En este estudio, proponemos un marco de razonamiento causal que incluye una taxonomía de tres ejes para la causalidad, diseñada para capturar la intención, los requisitos de acción y los tres niveles de causalidad definidos por Pearl: asociacional, intervencional y contrafactual; además de un enfoque de "humano en el circuito" para generar *golden collections* de preguntas causales naturales, anotadas de acuerdo con la taxonomía propuesta. Evaluamos las preguntas iniciales de una colección dorada en portugués utilizando los LLM GPT-4o y Llama3.1 con dos estrategias de prompt. Nuestros hallazgos revelan que ambos LLM enfrentan desafíos significativos al abordar preguntas causales intervencionales y contrafactuales. Estos resultados sugieren limitaciones en el uso indiscriminado de estos LLM para extender la anotación a preguntas naturales adicionales o para desarrollar agentes de IA causales basados en LLM.
Palabras clave: Razonamiento Causal, Razonamiento Intervencional, Razonamiento Contrafactual, Modelos de Lenguaje Extenso.

ISSN 1135-5948 DOI 10.26342/2025-74-25

1 Introduction

We are witnessing the massive use of Large Language Models (LLMs) in the development of generative AIs across a wide range of domains, including healthcare, legal decision-making, and customer service. Some researchers and commentators have speculated that these tools could represent a decisive step towards machines that demonstrate 'artificial general intelligence' (Kejriwal et al., 2024). However, on the path toward artificial general intelligence—which is purportedly being approached by modern LLMs like GPT-4 (OpenAI et al., 2024), Gemini (Anil et al., 2024), Claude [1], and open source LLMs like Llama (Touvron et al., 2023) and Gemma (Gemma Team et al., 2024) — the ability to understand cause-and-effect relationships and engage in causal reasoning is essential (Jin et al., 2023). In (Pearl and Mackenzie, 2018), the authors proposed the "Ladder of Causality" to categorize different levels of causal thinking. In the first rung, Associational, it is required the ability to detect correlations and patterns in observed data. LLMs already excel at this based on their pretraining data. But in the higher Pearl's rung - in which it is required to understand the effects of actions and interventions on a system (Interventional rung), and imagining and reasoning about hypotheticals and alternate realities (Counterfactual rung), in the best case, we need to evaluate how and whether LLMs have abilities to reason about these situations. (Jin et al., 2023) affirm that "these transformative developments raise the question of whether these machines are already capable of causal reasoning: *Do LLMs understand causality?*".

In this regard, we need to provide a set of natural causal questions to increase the capabilities of LLMs in interventional and counterfactual situations. However, there is a lack of a comprehensive collection of causal questions of this kind in previous works, even for high-resource languages, such as English. Existing causal datasets mainly focus on artificially crafted questions and have zero or limited coverage of natural human questions, not capturing pragmatic nuances and linguistic diversities (Ceraolo et al., 2024). The Portuguese language, despite being the 6th most spoken language in the world with around 270 million speakers, is considered a low-resource language (Blasi, Anastasopoulos, and Neubig, 2022) and this lack of datasets and golden standard collection for causal reasoning is even more critical. To date, there is no known benchmarking dataset that includes natural causal questions in Portuguese.

In this work, we propose a causal reasoning framework that includes a three-axis taxonomy for causality, designed to capture the intent and action requirements in causal reasoning chains and the three rungs of causality defined by (Pearl and Mackenzie, 2018), and a human-in-the-loop approach to generate a golden collection of natural causal questions, annotated according to the proposal taxonomy. We applied the taxonomy and the annotation methodology in a dataset comprising more than 7,000 causal natural questions in Portuguese, collected from public sources and produced by humans in interactions with other humans and software systems. We argue that this framework is promising for evaluating and fine-tuning LLM-based AI agents to: (1) determine when to apply causal reasoning versus non-causal knowledge, (2) identify the action class based on the interlocutor's intent, and (3) assess the required level of causal reasoning — associational, interventional, or counterfactual.

We evaluated the seed questions of the Golden Collection in Portuguese using the LLM GPT-4o and Llama3.1 (with two configurations - 70 and 8 billion parameters) with two prompt strategies (few-shot learning and Chain-of-Thought strategies). As expected, the LLM GPT-4o outperformed the LLM Llama3.1 in the majority of classifications. However, the findings indicated that both GPT-4o and Llama3.1 struggle to assess the type of reasoning required for causal questions (particularly interventional and counterfactual questions) and to recognize the need to identify cause-and-effect relationships between two variables or events (relation-seeking questions) and the effect of a cause (effect-seeking questions), yielding highly unsatisfactory results. These results did not support the indiscriminate use of these LLMs to extend annotation to additional natural questions and the use for the design of LLM-based causal AI agents. A more in-depth analysis of the error cases is essential, along with an evaluation of poten-

[1] Claude AI: http://claude.ai/

tial fine-tuning strategies to improve performance.

2 Related Works

For the English language, we have datasets with completely artificially generated causal questions, such as WIQA (Tandon et al., 2019), HeadLine Cause (Gusev and Tikhonov, 2022), GLUCOSE (Mostafazadeh et al., 2020), CLadder (Jin et al., 2023) and Corr2Cause (Jin et al., 2024). The datasets e-Care (Du et al., 2022) e Webis-CausalQA-22 (Bondarenko et al., 2022) contain some natural questions Human-to-Human and Human-to-SearchEngine, however, these bases do not include questions between humans and LLMs, due to having been proposed before the explosion in popularity of LLMs. Especially, (Jin et al., 2023) propose the CLadder, a database developed artificially through a Causal Inference Engine, which processes queries, graphs, and other information available in questions classified in the ladder of causality of Pearl.

Recently, (Ceraolo et al., 2024) proposed the CAUSALQUEST database containing natural causal questions in their entirety, collected from interactions between humans (Human-to-Human), between humans and Search engines (Human-to-SE), and between humans and Large Language Models (Human-to-LLMs). This dataset seeks to meet the need for natural question bases of a causal nature and the need for question bases aimed at LLMs, which have very particular characteristics, such as the length of each question, which can exceed 100 words per question. The authors argue that the structure of the questions formulated, scenarios, conditions, and examples may be used to improve understanding of LLM and optimize its results in causal reasoning.

For the Portuguese language, no studies are addressing the construction of a dataset containing natural causal questions, as well as the various taxonomies for causality, at least to the best of our knowledge to date. This fact already corroborates the importance of this work, as it provides the Portuguese language computational processing community with a basis for evaluating LLMs in causal reasoning.

3 A Framework for a Natural Causal Golden Collection

To guide the development of a Golden Collection (GC) with natural causal questions, we defined a three-axis taxonomy for causality inspired by (Ceraolo et al., 2024), (Bondarenko et al., 2022), and (Pearl and Mackenzie, 2018), and a human-in-the-loop approach to the annotation of the causal questions. We then used this framework and gathered a total of 7,594 natural questions from databases and repositories containing human-generated queries in Portuguese, which we used to create our gold standard collection through a human-in-the-loop approach.

3.1 A Three-Axis Framework for Causal Taxonomy

Our proposed taxonomy aims to represent causal knowledge across three axes (Figure 1). It is important to acknowledge that causal relationships are often complex, with multiple plausible causes leading to a given effect and a single cause potentially resulting in different effects depending on the context. In this work, we focus on general causal knowledge about events rather than specific instances where the cause-effect relationship is uniquely determined by a constrained scenario.

On Axis 1: "Causal/Non-Causal" serves as the most fundamental distinction, categorizing questions as either causal or non-causal. This enables an AI agent to identify when to apply cause-and-effect knowledge or reasoning. Our definition of causal questions builds on and extends the definition by (Bondarenko et al., 2022), which identifies three possible natural mechanisms in questions that involve causality: (1) *Given the cause, predict the effect(s)* - when the question presents an action or cause, implicit or explicit, and asks what effect(s) result from it. Questions like "What is the impact of deforestation on global warming?" or "What happens if I mix bleach and vinegar?" are examples of this type; (2) *Given the effect, propose the cause(s)* - questions where the human interlocutor asks what the cause(s) of an observed or hypothetical effect are. For example, "What disease causes throat irritation?" and "What is the best algorithm to perform graph search?"; (3) *Given variables, judge their causal relation* - questions in which the human interlocutor asks

whether two variables have a causal relationship with each other. This is the case with questions such as "Does eating a lot of fruit cause diabetes?", "Does drinking coffee after lunch hinder the absorption of nutrients?" or "Does improving my public speaking increase my employability?".

On the second axis, we categorize causal questions with a focus on the speaker's intent and the required action to answer them. Understanding the most common action class can provide insight into the capabilities needed by an AI causal solver. Axis 2: "Action Class" in our taxonomy proposes five subclasses:

- *Cause-Seeking* - questions that seek the cause of an effect, where the interlocutor presents an observed event and questions what or what causes it. Example: "Why is the sky blue?".
- *Effect-Seeking* - questions that seek the effect of an action or cause, asking what the consequences of a certain action or scenario are. Example: "What is the impact of deforestation on global warming?";
- *Relation-Seeking* - questions that seek to identify the causal relationship between different events, where a set of variables are presented and the interlocutor questions the causal relationship between them. Example: "Does drinking coffee after lunch hinder the absorption of nutrients?";
- *Recommendation-Seeking* - questions that present a set of options, implicitly or explicitly, and ask which of these options will maximize the effect desired by the interlocutor. Example: "What language should I learn to work abroad?";
- *Steps-Seeking* - questions where the interlocutor requests instructions to achieve a desired objective or the creation of artifacts such as food recipes, diets, or algorithms that meet a certain need. Example: "What's the best recipe for making a fluffy chocolate cake?".

Finally, we incorporate the Ladder of Causality framework from (Pearl and Mackenzie, 2018) in Axis 3: "Causal Reasoning", which outlines three rungs of reasoning required for an AI agent to effectively answer causal questions:

- *Associational* - questions that can be answered through a statistical association, using a correlation between variables to understand the cause-and-effect relationship between them. These are questions like "What does a test grade say about the student?";
- *Interventional* - questions classified here require a more complex type of reasoning, modifying one of the variables involved in the question to understand whether it influences the outcome of the event. This can be understood as modifying an action to see what effect will result from it. An example of this type of question is "Should I move closer to work or stay where I am and face a two-and-a-half-hour public transport commute?";
- *Counterfactual*: questions that require even more complex reasoning, as they ask about alternative possibilities, events that did not happen, and purely hypothetical scenarios. It requires understanding how a hypothetical scenario would compare to what is observed in reality. Examples of this are "What would the world be like if dinosaurs hadn't gone extinct?" or "If I had studied more, would I have gotten a better grade?". While interventional causality predicts the consequences of actions, counterfactual causality compares reality to an alternative world where the action did not happen.

3.2 A Human-in-the-loop Approach to Annotation of Causal Questions

The pipeline of the human-in-the-loop approach to the annotation of causal questions is illustrated in Figure 2.

In Step 1, the first activity is the selection of sources of questions classified by the type of communication used (Human-to-Human, Human-to-Search Engine, and Human-to-LLMs). From these sources, we extracted texts representing questions, applying an initial automated filtering process followed by a manual review to remove noise and ensure non-questions that may have passed through the automated filter are excluded. Subsequently, we cataloged the questions by source, allowing us to trace their origins.

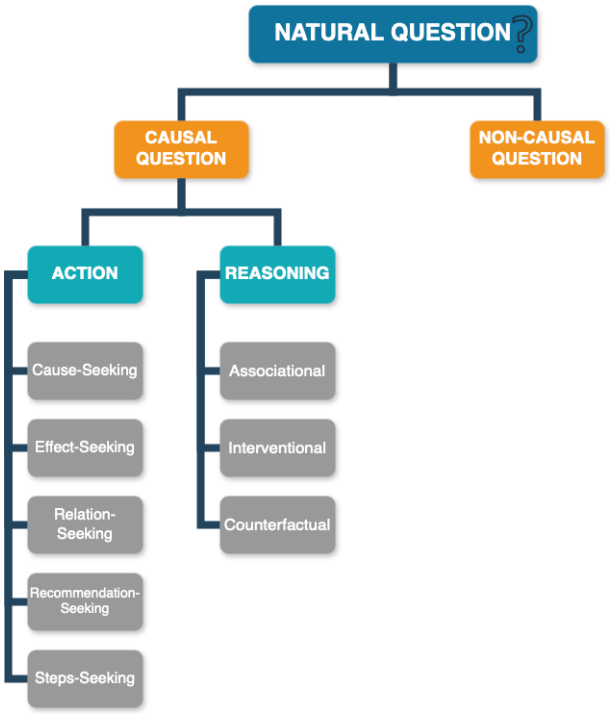

Figure 1: A Three-Axis Taxonomy for Causality Expression.

This step is crucial for our analysis, as each source tends to exhibit distinct communication patterns, with some being more formal and others more informal. Notably, there is a significant difference in how questions are posed between humans and LLMs, with the latter typically being more detailed and complex.

In Step 2 occurs the selection of natural questions from the Natural Questions source dataset, generated in Step 1, which will be the seed to the entire golden collection annotation and guidelines refinement. An important requirement is that the seed questions must preserve the general characteristics of the complete source dataset.

In Step 3, two or more human annotators classify each of the seed questions in each of the three axes of the taxonomy - Axis 1: "Causal/Non-Causal"; Axis 2: "Action Class"; and Axis 3: "Causal Reasoning", using the guidelines of the taxonomy (see Section 3.1).

In Step 4, the agreement level among the annotators is evaluated to determine whether further refinement of the taxonomy guidelines and additional discussion and alignment sessions among the annotators are necessary. In this evaluation, the inter-annotator agreement metric Kappa (Cohen, 1960) can be used, with a minimum threshold of 0.81 (considered "almost perfect agreement" according to (Landis and Koch, 1977)). If the evaluation meets this threshold, the process advances to Step 6. Otherwise, the iterative process continues in Step 5.

In Step 5, a review of cases of disagreement is conducted, along with alignment sessions with the annotators, with the objective of refining and updating the taxonomy guidelines.

In Step 6, the annotators have reached an "almost perfect" agreement level, and a third-party reviewer conducts the adjudication process, resolving any discrepancies between the annotators' responses. The adjudication results, together with the seed questions that achieved agreement, form the Golden Collection of seed questions.

In Steps 7 and 8, we introduce evaluation cycles that combine LLM-driven annotation (with or without fine-tuning) and human review. In Step 7, we select the LLMs and define the prompting strategies, then perform inference on the Golden Collection of seed questions, which resulted from Step 6. The selection of LLMs and prompting strategies in Step 7 is critical to ensuring the success of the final Golden Collection – CaLQuest.PT.

In Step 8, we evaluate the model predictions against the reference classifications in the Golden Collection, using standard classification metrics such as precision, recall, and F1-Score. These steps assess whether the LLMs can accurately classify questions within the Three-Axis Taxonomy, thereby testing their ability to recognize causality. To evaluate this capability, we must establish a threshold for quality and consistency. For example, in the creation of CalQuest.PT (see Subsection 3.3), we tested GPT-4o and Llama3 with two prompting strategies: Few-shot learning and Chain-of-Thought (CoT). We set a minimum threshold of F1-Score ≥ 0.8 for each axis.

Step 9 marks the final stage, where the CaLQuest.PT Golden Collection is generated using the LLM and prompting strategy that achieve the best results. The dataset is formatted in a machine-readable format, such as JSON, and includes the question identifier, its source, the question itself, and its classification for each axis.

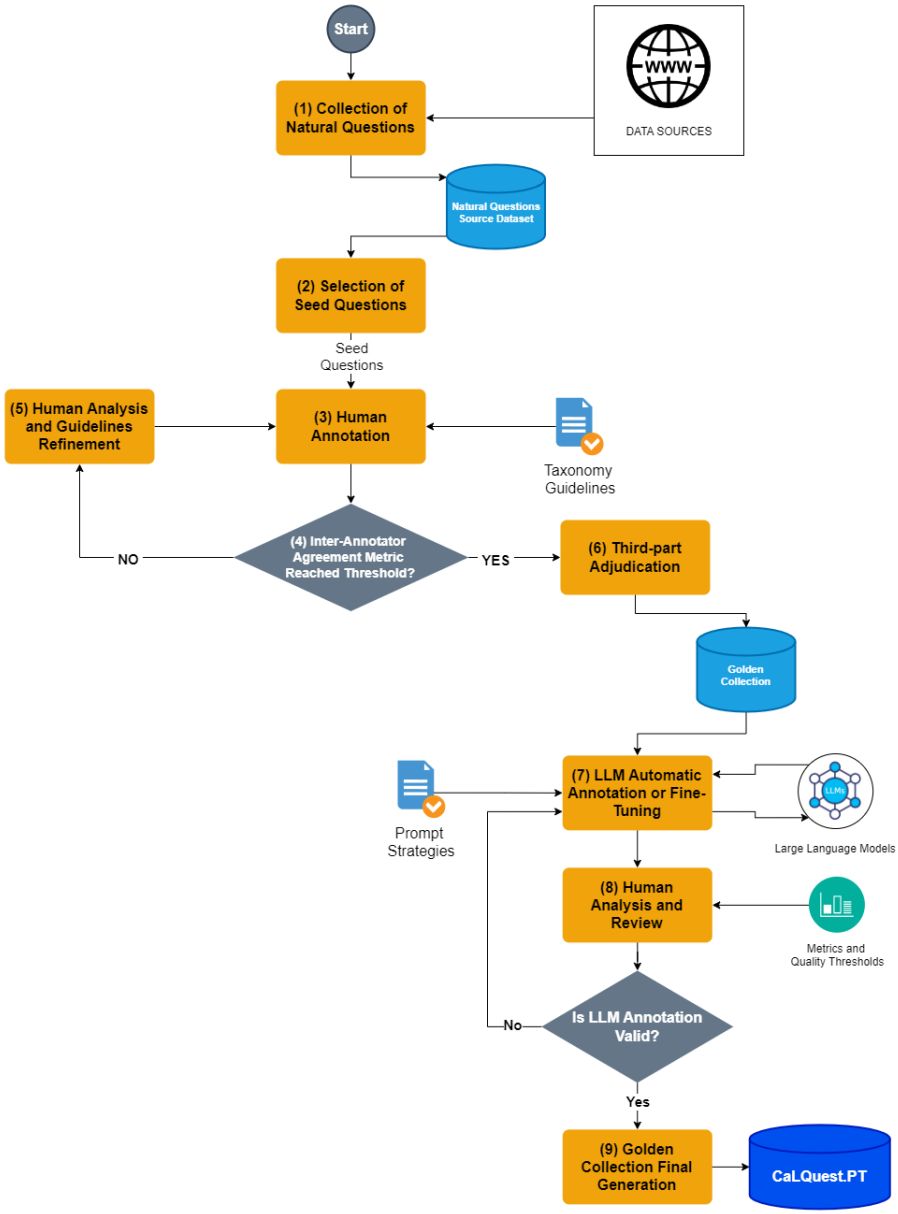

Figure 2: The Human-in-the-loop Approach to Annotation of Causal Datasets.

3.3 A Golden Collection of Natural Causal Questions in Portuguese

We used the human-in-the-loop approach, proposed in this paper (see Figure 2) to develop a golden collection of natural causal questions in Portuguese - the CaLQuest.PT. According to Step 1, we collect both causal and non-causal questions, originally in the Portuguese language, that humans ask either other humans or software, such as search engines and chatbots. The starting point was selecting public sources of human interactions between other humans (H-to-H), LLMs (H-to-LLM), and Search Engines (H-to-SE). Unfortunately, we didn't find public sources with H-to-SE questions in Portuguese. So, we chose three distinct sources with H-to-H and H-to-LLM questions, which are well used in other works (Ceraolo et al., 2024), from which we collected all questions from these three datasets totaling 7,594 questions (see the distribution of the dataset in Table 1). The first set of natural questions was gathered from the question-and-answer forum Reddit[2], where interactions are H-to-H. The other two datasets are from sources where humans interact with LLMs

[2]Reddit: https://www.reddit.com (accessed on 11/12/2024)

Interaction Type	Datasets	#Samples
H-to-H	Reddit	3,251
H-to-LLM	ShareGPT	646
	WildChat	3,697
		7,594

Table 1: Overview of the datasets comprising the CaLQuest.PT collection.

(H-to-LLM): WildChat (Zhang et al., 2023), which contains data shared by ChatGPT users in the free service environment, and the ShareGPT[3] source, containing conversations with ChatGPT voluntarily shared by users.[4]

We analyze the datasets of the CalQuest.PT in terms of its linguistic properties (see Table 2). Overall, CaLQuest.PT has good coverage of 7K human questions in the Portuguese language, with 32K unique words in its vocabulary and 28.75 words per sample on average. The Type-Token Ratio (TTR) indicates that there are few repetitions of words in the natural questions. Table 3 shows the distribution of the datasets by question type according to the 5W-2H question categorization. We performed this classification to assess whether the distribution of question types in the Portuguese dataset aligns with previous findings in English-language studies, such as those by (Ceraolo et al., 2024) and (McClure et al., 2001). There is a prevalence of "What" and "How" questions, accounting for 50.7% and 18.0% of the total questions, respectively, which is consistent with the cited studies. The "Others" category includes natural questions that do not fit the 5W-2H pattern, often being syntactically incorrect or ambiguous (e.g., "Horror video reaction channels, no crime?"). Most of these questions have fewer than 100 tokens, suggesting they do not belong to the extensive LLM-generated question group in the dataset.

According to Step 2, we selected 553 seed questions equally from each dataset. Details on linguistic features and analysis of 5W-2H question types are provided in Appendix E.

In Step 3, two human annotators classified each of the 553 questions in each of the three axes of the taxonomy - Axis 1: "Causal/Non-Causal"; Axis 2: "Action Class"; and Axis 3: "Causal Reasoning", according to the Taxonomy Guidelines. We conducted two iterations, through steps 3, 4, and 5, to achieve a satisfactory level of agreement, using Cohen's Kappa (Cohen, 1960). In the first iteration, we observed a low level of agreement (Kappa = 21.5), so, we identified and reviewed the divergent annotations, and refined the Taxonomy Guidelines (Step 5). According to the new guidelines, the two annotators proceeded with the reclassification (Step 3). Finally, after the second iteration, we achieved an inter-annotator agreement Kappa = 83.8 (Step 4).

In Step 6, a third-party adjudication was performed in a few cases of divergence, resulting in the Golden Collection of seed questions. Table 4 presents the distribution of this Golden Collection across each axis of the taxonomy. On Axis 1 - "Causal/Non-Causal", we can see that 39.9% of the seed questions are causal questions (221) and 60.1% are non-causal questions (332). The dataset Reddit has more Causal seed questions, since, as it is an online forum, have more practical questions like "What can I do to get into the master's degree?" or "Is it worth taking the Administrative Assistant course?". On the other hand, Wildchat and ShareGPT datasets have more Non-Causal seed questions. Many of the questions on Human-to-LLM datasets are asking for information, as in "*Who is the professional who advises you to upgrade your computer?*", or asking for simple tasks like "*Put the following elements in ascending order of electronegativity: oxygen, nitrogen, sodium, silver, lead, polonium, bromine, iron, copper and calcium, please.*". On Axis 2 and Axis 3, we observe the nature of natural causal questions. In human-to-human (H-To-H) interactions (Reddit dataset), people often ask subjective questions, such as "Recommendation-seeking", which represent 34.9% of causal questions. In contrast, in H-to-LLM interactions (WildChat and ShareGPT), users primarily ask for algorithmic steps or food recipes ("Steps-Seeking" questions), accounting for 43.2% and 59.0%, respectively. Regarding Axis 3 ("Causal Reasoning"), following Pearl's Ladder of Causality, LLMs receive mostly associational questions (63.8%), while counterfactual questions are less represented.

[3]ShareGPT: https://huggingface.co/datasets/anon8231489123/ShareGPT_Vicuna_unfiltered (accessed on 11/12/2024)

[4]Data License: ShareGPT (Apache-2), WildChat (AI2 ImpACT - Low Risk), Reddit (Non-Commercial research only)

Feature	Reddit	WildChat	ShareGPT	Total/Avg
Samples	3,251	3,697	646	7,594
Avg. Words/Sample	10.21	40.50	56.07	28.86
Vocab Size	5,760	23,348	10,264	30,860
Type-Token Ratio	0.97	0.86	0.82	0.91

Table 2: Linguistic features in CaLQuest.PT datasets.

Question Type	Reddit	WildChat	ShareGPT	Total	%
What	1,530	1,906	415	3,851	50.71%
Who	136	42	10	188	2,48%
Why	264	107	12	383	5,04%
Where	117	157	19	293	3,86%
When	52	101	6	159	2,09%
How	625	636	112	1.373	18,08%
How much	111	49	7	167	2,20%
Others	416	699	65	1,180	15,54%
Total				7,594	100%

Table 3: Analysis of the question types 5W-2H in CaLQuest.PT datasets.

Appendix D provides examples of natural questions for each class across all axes.

In Steps 7 and 8, we completed one evaluation cycle for a set of 553 seed questions, combining LLM-driven annotation with subsequent human review. In this first cycle, we used GPT-4o (Team et al., 2024) and Llama3 (Patterson et al., 2022) with the initial aim of assessing how well two of the most robust LLMs currently available could recognize the nature of the seed questions. The evaluation of causal reasoning by LLMs and the results obtained will be presented and discussed in detail in Section 4.

4 *Evaluating Causal CommonSense Reasoning in LLMs*

Our main objective is to investigate how much more robust LLMs can recognize the nature of causal questions. In this evaluation cycle, we applied the LLM GPT-4o (through the API provided by OpenAI and with the default hyperparameters) and the LLM Llama 3.1 70B and 8B. The selection of two open-source LLMs aimed to investigate the performance of such models with varying parameter sizes (70B and a smaller one with 8B). We use two prompt strategies - Few-shot Learning (Brown et al., 2020) and Chain-of-Thought (CoT) (Wei et al., 2022). The prompts translated to English (because the prompts were used in Portuguese), used in each axis of the taxonomy, are transcribed in Appendix A, B and C. Tables 5, 6 and 7 present the results of GPT4o and Llama3.1-70B in terms of F1-Score of each prompt strategy for each classification axis.

As expected, overall, the LLM GPT-4o outperformed the LLM Llama3.1-70B in the vast majority of classifications. In the following classes, Llama3.1-70B outperformed GPT-4o: non-causal (Axis 1), effect-seeking, and relation-seeking (Axis 2). In Axis-1, LLM GPT-4o showed an interesting result in classifying causal and non-causal questions, achieving, respectively, an F1-Score of 84.5%, using the Few-Shot Learning prompt strategy, and 81.9% using the CoT prompt strategy (see Table 5). The main errors in detecting causality occurred in questions with unconventional formulations, such as *"Courses to gift for the TJ SP public contest for clerk?"* and *"How did you get started with alcohol?"*.

In Axis-2, LLM GPT-4o also showed promising performance in classifying causal questions regarding action class, when we used the Few-Shot Learning prompt strategy (see Table 6). Its worst performance was in classifying questions in which the human sought to identify the effects of an action or intervention (Effect-Seeking), with F1-Score = 57.1% (Few-Shot strategy) and the relation between variables (Relation-Seeking), with F1-Score = 64.7% (Few-Shot strategy). In these classes, Llama3.1-70B outperformed GPT-4o, with F1-Score = 57.7% and 74.3%, respectively. Contrary to our expectations,

Classification	Reddit	WildChat	ShareGPT	Total	%
AXIS 1 - Causal / Non-Causal					
Causal	123	37	61	221	39.9%
Non-Causal	73	154	105	332	60.1%
	.	.	.	553	100.0%
AXIS 2 - Action Class					
Cause-Seeking	11	8	5	24	10.9%
Effect-Seeking	23	7	2	32	14.5%
Steps-Seeking	29	16	36	81	36.6%
Recommendation-Seeking	43	4	16	63	28.5%
Relation-Seeking	17	2	2	21	9.5 %
	.	.	.	221	100.0%
AXIS 3 - Causal Reasoning					
Associational	60	27	54	141	63.8 %
Interventional	35	4	5	44	19.9 %
Counterfactual	28	6	2	36	16.3 %
	.	.	.	221	100.0%

Table 4: Distribution of the seed questions of the CaLQuest.PT across our Three-axis Taxonomy.

LLM (Prompt)	Causal	Non-Causal
GPT-4o (FS)	84.5%	82.9%
GPT-4o (CoT)	81.9%	88.9%
LLAMA3.1 70B (FS)	77.6%	85.1%
LLAMA3.1 70B (CoT)	74.6%	83.9%
LLAMA 8B (FS)	69.0%	70.9%
LLAMA 8B (CoT)	68.5%	72.0%

Table 5: GPT-4o and Llama3.1 classification results of seed questions into Causal and Non-Causal Categories (Axis-1) using Few-Shot Learning (FS) and Chain-of-Thought (CoT) Prompting Strategies.

LLM (Prompt)	Cause-Seek.	Effect-Seek.	Steps-Seek.	Rec-Seek.	Rel-Seek.
GPT-4o (FS)	89.8%	57.1%	92.8%	84.7%	64.7%
GPT-4o (CoT)	82.3%	54.9%	91.7%	82.0%	66.7%
LLAMA3.1 70B (FS)	75.6%	57.7%	84.9%	76.5%	74.3%
LLAMA3.1 70B (CoT)	77.5%	58.2%	88.6%	76.7%	55.2%
LLAMA 8B (FS)	75.0%	39.4%	76.4%	57.4%	34.5%
LLAMA 8B (CoT)	56.6%	47.1%	77.7%	61.4%	13.8%

Table 6: GPT-4o and Llama3.1 classification results of seed questions into Action Classes (Axis-2) using Few-Shot Learning (FS) and Chain-of-Thought (CoT) Prompting Strategies.

LLM (Prompt)	Associational	Interventional	Counterfactual
GPT-4o (FS)	79.3%	63.5%	52.0%
GPT-4o (CoT)	80.6%	64.6%	46.8%
LLAMA3.1 70B (FS)	70.9%	14.3%	25.0%
LLAMA3.1 70B (CoT)	71.2%	51.3%	39.2%
LLAMA 8B (FS)	66.4%	33.6%	46.4%
LLAMA 8B (CoT)	58.3%	32.8%	33.3%

Table 7: GPT-4o and Llama3.1 classification results of seed questions into Pearl's Ladder of Causality (Axis-3) using Few-Shot Learning (FS) and Chain-of-Thought (CoT) Prompting Strategies.

the Chain of Thought (CoT) prompt strategy performed worse. Reviewing studies such as (Kojima et al., 2023), we observe that CoT prompts tend to underperform in multiple-choice and simple classification tasks due to minor logical construction errors that are typically only noticeable by humans.

In Axis 3 - Ladder of Causality, LLM GPT4o showed reasonable performance in recognizing the type Associational with F1-Score = 80.6% (CoT strategy). But, on other levels, the performance fell below expectations for such a robust LLM (see Table 7). In the "Interventional" rung achieved F1-Score = 64.6% (CoT strategy) with very low precision = 49.4%, indicating many false-positives, as in the case of the question " *What can I do to get into the master's degree?* ", that was classified as "Interventional" but it has an associative nature since it is seeking methods that correlate with the desired effect (entering the master's degree). The "Counterfactual" rung result achieved the worst result with F1-Score = 52.0% (Few-Shot strategy). LLM Llama3.1-70B also underperformed in interventional and counterfactual causal reasoning with 51.3% and 39.2% of F1-Score (CoT strategy), respectively. Unlike the other axes, the CoT strategy by Llama3.170B showed a small improvement in results compared to the Few-Shot Learning prompt strategy.

One possible explanation for the observed difference is that GPT-4o already possesses strong implicit inference capabilities, allowing it to arrive at the correct answer without requiring explicitly structured reasoning through CoT. In this scenario, adding intermediate steps in CoT could be redundant or even detrimental, introducing unnecessary variations in the final response. In contrast, Llama-70B may rely more on CoT to better structure its decision-making process, as it might not have the same level of implicit inference as GPT-4o. To test this hypothesis, one could analyze the distribution of tokens generated by each model with and without CoT, checking whether GPT-4o maintains greater response stability even when intermediate steps are omitted. Another approach would be to measure the entropy of the outputs to see if CoT increases variability in GPT-4o (indicating a possible negative impact) while improving Llama-70B's consistency. Additionally, ablation experiments could be conducted by selectively removing parts of the CoT to assess whether Llama-70B's performance degrades more significantly compared to GPT-4o.

5 Conclusion

This work presents a proposal of a causal reasoning framework with a causal taxonomy and an annotation methodology. We argue that this framework is promising for evaluating and fine-tuning LLM-based AI agents to: (1) determine when to apply causal reasoning versus non-causal knowledge, (2) identify the action class based on the interlocutor's intent, and (3) assess the required level of causal reasoning — associational, interventional, or counterfactual. We evaluated the LLM GPT-4o and Llama3.1-70B and 8B in the classification of seed questions of a Golden collection of natural causal questions in the Portuguese Language. Our findings indicated that both GPT-4o and Llama3.1-70B struggle to assess the type of reasoning interventional and counterfactual and cause-and-effect relationships. These results did not support the indiscriminate use of these LLMs in the development of AI Causal Agents. In future works, we plan to explore other LLMs and fine-tuning processes and a more in-depth analysis of the error cases.

5.1 Limitations and Challenges

A key premise of this work was to use original Portuguese questions to preserve pragmatic and cultural nuances, avoiding translations from English. The main challenge was obtaining a diverse and representative set of natural questions across different human-machine interaction scenarios. Notably, we could not collect Portuguese search engine queries (e.g., Bing, Google) due to the lack of publicly available data. Additionally, counterfactual questions were less frequent in the explored environments. Another challenge is the subjective and dubious nature of the questions and the consequent difficulty in classifying some questions in a taxonomy, whatever it may be. The dynamicity and expressiveness of natural languages allow us to ask a question in different ways and, often, the intention is quite implicit.

References

Anil, R., S. Borgeaud, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai,

A. Hauth, K. Millican, D. Silver, M. Johnson, I. Antonoglou, J. Schrittwieser, A. Glaese, J. Chen, and E. Pitler. 2024. Gemini: A family of highly capable multimodal models.

Blasi, D., A. Anastasopoulos, and G. Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland, May. Association for Computational Linguistics.

Bondarenko, A., M. Wolska, S. Heindorf, L. Blübaum, A.-C. Ngonga Ngomo, B. Stein, P. Braslavski, M. Hagen, and M. Potthast. 2022. CausalQA: A benchmark for causal question answering. In N. Calzolari, C.-R. Huang, H. Kim, J. Pustejovsky, L. Wanner, K.-S. Choi, P.-M. Ryu, H.-H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, and S.-H. Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3296–3308, Gyeongju, Republic of Korea, October. International Committee on Computational Linguistics.

Brown, T. B., B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. 2020. Language models are few-shot learners.

Ceraolo, R., D. Kharlapenko, A. Reymond, R. Mihalcea, M. Sachan, B. Schölkopf, and Z. Jin. 2024. Causalquest: Collecting natural causal questions for ai agents.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Du, L., X. Ding, K. Xiong, T. Liu, and B. Qin. 2022. e-CARE: a new dataset for exploring explainable causal reasoning. In S. Muresan, P. Nakov, and A. Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 432–446, Dublin, Ireland, May. Association for Computational Linguistics.

Gemma Team, T. M., C. Hardin, R. Dadashi, S. Bhupatiraju, L. Sifre, M. Rivière, M. S. Kale, J. Love, P. Tafti, L. Hussenot, and et al. 2024. Gemma.

Gusev, I. and A. Tikhonov. 2022. HeadlineCause: A dataset of news headlines for detecting causalities. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, and S. Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6153–6161, Marseille, France, June. European Language Resources Association.

Jin, Z., Y. Chen, F. Leeb, L. Gresele, O. Kamal, Z. LYU, K. Blin, F. Gonzalez Adauto, M. Kleiman-Weiner, M. Sachan, and B. Schölkopf. 2023. Cladder: Assessing causal reasoning in language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 31038–31065. Curran Associates, Inc.

Jin, Z., J. Liu, Z. LYU, S. Poff, M. Sachan, R. Mihalcea, M. T. Diab, and B. Schölkopf. 2024. Can large language models infer causation from correlation? In *The Twelfth International Conference on Learning Representations*.

Kejriwal, M., H. Santos, A. M. Mulvehill, K. Shen, D. L. McGuinness, and H. Lieberman. 2024. Can ai have common sense? finding out will be key to achieving machine intelligence. *Nature*, 634:291–294.

Kojima, T., S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. 2023. Large language models are zero-shot reasoners.

Landis, J. R. and G. G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

McClure, J., D. J. Hilton, J. Cowan, L. Ishida, and M. Wilson. 2001. When people explain difficult actions, is the causal question how or why? *Journal of Language and Social Psychology*, 20(3):339–357.

Mostafazadeh, N., A. Kalyanpur, L. Moon, D. Buchanan, L. Berkowitz, O. Biran, and J. Chu-Carroll. 2020. GLUCOSE: GeneraLized and COntextualized story explanations. In B. Webber, T. Cohn, Y. He, and Y. Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4569–4586, Online, November. Association for Computational Linguistics.

OpenAI, J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, R. Avila, I. Babuschkin, S. Balaji, V. Balcom, P. Baltescu, H. Bao, M. Bavarian, and J. B. et al. 2024. Gpt-4 technical report.

Patterson, D., J. Gonzalez, U. Hölzle, Q. Le, C. Liang, L.-M. Munguia, D. Rothchild, D. So, M. Texier, and J. Dean. 2022. The carbon footprint of machine learning training will plateau, then shrink.

Pearl, J. and D. Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. Basic Books, Inc., USA, 1st edition.

Tandon, N., B. Dalvi, K. Sakaguchi, P. Clark, and A. Bosselut. 2019. WIQA: A dataset for "what if..." reasoning over procedural text. In K. Inui, J. Jiang, V. Ng, and X. Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6076–6085, Hong Kong, China, November. Association for Computational Linguistics.

Team, O., A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford, A. Madry, A. Baker-Whitcomb, and A. B. et al. 2024. Gpt-4o system card.

Touvron, H., T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample. 2023. Llama: Open and efficient foundation language models.

Wei, J., X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Zhang, L., H. Xu, Y. Yang, S. Zhou, W. You, M. Arora, and C. Callison-Burch. 2023. Causal reasoning of entities and events in procedural texts. In A. Vlachos and I. Augenstein, editors, *Findings of the Association for Computational Linguistics: EACL 2023*, pages 415–431, Dubrovnik, Croatia, May. Association for Computational Linguistics.

A Prompts to Axis 1 - "Causal/Non-Causal" classification

The following question was asked by a human, and you must classify it into one of the following categories:

Category 1: Causal. This category includes questions that imply cause-and-effect relationships, requiring knowledge of causality and reasoning to derive an answer. A causal question may have three types of objectives or behaviors:

Given a cause, predict the effect: Seeks to understand the impact or outcome of a specific cause, which may involve predicting future scenarios or hypothetical situations (e.g., What happens if I play the lottery? Should I learn a new programming language? Will renewable energy dominate our energy mix in the future? What would happen if social networks didn't exist?).

Given an effect, predict a cause: Asks why something occurred (e.g., Why do apples fall?), questions the cause of a specific effect, inquires about the reason behind something, or seeks actions necessary to achieve a specific goal, either implicitly or explicitly (e.g., Why are extremist movements increasing lately? How can I earn a million dollars? How can I learn a new language in 30 days?). This also includes cases where the effect is not explicit: any request with a purpose, seeking the action (cause) that best fulfills a specific goal (effect), even if the goal is implicit. If someone asks for a restaurant recommendation, they are looking for the best cause (restaurant) to achieve a certain effect (e.g., eating healthily). Similarly, asking for a vegan recipe implies seeking a recipe that is the cause of the best possible meal. Questions requesting "the best way" to do something fall into this category.

Given a set of variables, judge the causal relationship between them: Questions the causal relationship between a set of entities (e.g., Does smoking cause cancer? Was I rejected from the job interview because I lack experience?).

Category 2: Non-Causal. These are questions that do not imply any of the causal relationships mentioned above. For instance, a non-causal question might involve a request for translation, correction, paraphrasing a text, creating a story, playing a game, solving a mathematical problem or puzzle, providing information about something (e.g., software, websites, addresses, events, general locations), or using such information for comparison, without much reasoning involved. These questions are non-causal because the user is simply seeking information.

Examples:

Question: What's the worst way to make money? Category:

Question: Why are betting companies like Tigrinho or Blazer not shut down? Category:

Question: Is there something you do out of obligation? Category:

Question: What's your happiest memory? Category:

Question: What's the ideal temperature for a PABX rack and a network switch rack? Category:

Question: What will be the chemical reaction if "(NH2)2CO" is added to "NaCl"? Category:

Question: Which countries currently have elective monarchies? Category:

Question: What is the latest version of PyTorch? Category:

Question: What's the best way to remove a background from a photo in Photoshop? Category:

Question: What car route can I take between São Paulo and Brasília? Category:

Question: Can you find a specific topic in our chat history? Category:

Question: What is a vinyl record? Category:

Here is the question: {QUESTION}, and I ask you to classify it into one of the two categories detailed above: Causal or Non-Causal, returning the category and the reasoning that justifies your classification in the following format:
CATEGORY:
REASONING:

Figure 3: Few-Shot Learning Prompt to Axis 1 - "Causal/Non-Causal" classification.

For Chain of Thought prompting, we modified the last paragraph to include the instruction "Faça uma linha de raciocínio passo-a-passo" ("Make a reasoning step-by-step").

[...] I request that you classify it into one of the five categories detailed above: Cause-Seeking, Effect-Seeking, Relation-Seeking, Recommendation-Seeking, or Step-Seeking. Provide a step-by-step reasoning process. Finally, respond in the following format [...]

Figure 4: Chain-of-Thought Prompt to Axis 1 - "Causal/Non-Causal" classification.

B Prompts to Axis 2 - "Action Class" classification

The following question was asked by a human. This is a causal question, and you must classify it into one of the following action categories:

Cause-Seeking: Explains the cause behind a certain phenomenon. The individual seeks to discover the reason or justification for something being the way it is. It could be a question containing "Why" (e.g., Why do leaves fall in autumn? Why is the sky blue?). It may also involve understanding an explanation or the significance of a statement, idea, or creative work, such as the meaning of song lyrics, poems, or story narratives (e.g., What is the meaning of the song "Tempo Perdido"? What are the main causes of Alzheimer's?). In summary, this type of question aims to uncover the cause or justification for something being the way it is or happening the way it did.
Formula:
Given: Effect, Asks for: Cause(s).

Effect-Seeking: Aims to predict the effects of an action, foresee the future given past circumstances, or predict a hypothetical scenario based on a counterfactual condition (e.g., Will renewable energy become our main source of energy in the future? What would the world be like if the internet had never been invented?).
Formula:
Given: Cause(s), Asks for: Effect(s).

Relation-Seeking: Questions the cause-and-effect relationship between distinct entities. The individual seeks to understand whether there is a causal relationship between the entities presented in the question (e.g., Does smoking cause lung cancer? Can air pollution increase the risk of respiratory diseases?). This class differs from "Cause-Seeking" and "Effect-Seeking" because the questioner presents a hypothesis about causality and asks if there is a causal relationship in the situation.
Formula:
Given: Set of causes and effects, Asks for: Causal relationship.

Recommendation-Seeking: Given an explicit or implicit goal and a set of options, it asks for the best option to achieve the goal (e.g., Should I try to pass a public exam to have better job opportunities? What's the best pizzeria in Fortaleza?). The individual has a goal and a set of options to choose from and wants to select the option that maximizes the results of their goal. This category differs from "Step-Seeking" because the individual has a set of options and necessarily wants to choose the best among them.
Formula:
Given: (Effect/human teleological purpose), Asks for: Guide that maximizes results (satisfies the purpose).

Step-Seeking: Proposes solving a problem through a series of steps or an algorithm (e.g., How can I learn English in 6 months? Create a vegan recipe with sweet potatoes and beans. Optimize this code to make it faster.). The individual has a purpose to be fulfilled and wants a solution in the form of a series of steps that can be followed. The answer can be a list of steps, a computer program, or a recipe. Questions can have a single way to be answered or multiple ways to achieve the goal. It does not involve weighing possibilities and choosing the best one among them.
Formula:
Given: (Effect/human teleological purpose), Asks for: Causes in the form of a step-by-step guide, code, or recipe.

In summary:

"Cause-Seeking" aims to uncover the reason or justification for something being the way it is or happening the way it did.

"Effect-Seeking" predicts the effects of an action, foresees the future, or hypothesizes based on counterfactuals.

"Relation-Seeking" seeks to understand the causal relationship between entities.

"Recommendation-Seeking" identifies the best option to achieve a goal from a set of choices.

"Step-Seeking" seeks a step-by-step solution to achieve a specific purpose.

Examples:

Question: Why do stores offer discounts for payments via Pix but not for bank slips? Category: Cause-Seeking

Question: What are the signs of a happy and healthy relationship, in your opinion? Category: Effect-Seeking

Question: Is there a minimum or ideal age to learn about politics and economics? Category: Relation-Seeking

Question: Which medium offers more creative freedom for creators: books, films, series, or comics? Category: Recommendation-Seeking

Question: How do I change my name on Reddit? Category: Step-Seeking

Question: How do attacks happen on websites created using WordPress? Category: Cause-Seeking

Question: Are sodas with 50% fruit content healthy? Category: Relation-Seeking

Question: What's the best way to use ChatGPT to create content? Category: Recommendation-Seeking

Question: How can I implement separate read and write instances for a database in Laravel 7.4? Category: Step-Seeking

Here is the question: {QUESTION}, and I ask you to classify it into one of the five categories detailed above: Cause-Seeking, Effect-Seeking, Relation-Seeking, Recommendation-Seeking, or Step-Seeking; returning the category and the reasoning that justifies your classification in the following format:
CATEGORY:
REASONING:

Figure 5: Few-Shot Learning Prompt to Axis 2 - "Action Class" classification.

For Chain of Thought prompting, we modified the last paragraph to include the instruction "Faça uma linha de raciocínio passo-a-passo" ("Make a reasoning step-by-step").

[...] I request that you classify it into one of the two categories detailed above: Causal or Non-Causal. Provide a step-by-step reasoning process. Finally, respond in the following format [...]

Figure 6: Chain of Thought Prompt to Axis 2 - "Action Class" classification.

C Prompts to Axis 3 - ”Causal Reasoning Ladder” classification

The following question was asked by a human. This question is a causal question, and you must classify it into one of the following categories according to Pearl's Causal Hierarchy:

Associational: This category pertains to questions that raise a relationship of statistical association and correlation between two variables, questioning the likelihood of event Y occurring given an initial event X. Examples include: *"What does rejection from a job application say about the candidate?"* or *"What is the best programming language for data science?"* This association can be explicit, as in the examples above, or implicit, such as in *"I have eye pain and joint pain, what disease could it be?"* where the speaker seeks to understand the condition most correlated with their symptoms. It also applies to recommendation questions, such as *"What are the best fixed-income investments for a student?"* where the speaker seeks advice on investments most suited to their financial profile. This type of question encompasses various formats, such as searching for methods correlated with a specific goal (*"How can I work two jobs?"*), a place or object that meets the speaker's needs (*"Where can I go to relax?"*), or a reason associated with an event (*"Why are the leaves turning yellow?"*).

Interventional: This category includes questions that seek to understand beyond the correlation between two events. To achieve this, the individual asks questions framed to intervene in the system by modifying or adding an action to comprehend its ultimate effect. Examples include: *"If she gains more work experience, will she be hired?"* or *"If I add fruit to the cake, will it become sweet?"* This type of question can also compare options, where the speaker wants to know which will yield the best result, such as *"Should I wake up earlier every day for more time or later for more rest?"* It can also be implicit, as in *"Should I buy new equipment for work?"* where the speaker wants to know the impact of taking this action on their future.

Counterfactual: This category encompasses questions about alternative realities, altering variables from an event that has already occurred to understand how it unfolded and what potential outcomes might have arisen if some variables had been different. Counterfactual causal questions generate hypotheses about other possible causes. Examples include: *"Was I rejected because I lacked experience?"* or *"Did I develop chondromalacia because I was overweight?"*

Examples of questions classified into one of the three categories – Associational, Interventional, Counterfactual:

Question: Why has a job interview become torture? **Category:** Associational

Question: Can I pursue a master's degree if I graduate from an online program? **Category:** Interventional

Question: Would I have great job opportunities with these courses on my resume + my experience? **Category:** Counterfactual

Question: Is this new generation really worse than the last? **Category:** Associational

Question: Do you think you would lose your friendships if people knew everything you thought? **Category:** Interventional

Question: What would have happened if there had never been exploitation in the world? **Category:** Counterfactual

Here is the question: {QUESTION}, and I ask you to classify it into one of the three categories detailed above: **Associational, Interventional,** or **Counterfactual**. If the question does not fit any of these categories, classify it as "None." Finally, respond in the following format:
CATEGORY:
REASONING:

Figure 7: Few-Shot Learning Prompt to Axis 3 - ”Causal Reasoning Ladder” classification.

For Chain of Thought prompting, we modified the last paragraph to include the instruction ”Faça uma linha de raciocínio passo-a-passo” (”Make a reasoning step-by-step”).

[...] I request that you classify it into one of the three categories detailed above: Associational, Interventional, or Counterfactual. If it cannot be classified into any of these, classify it as "None." Provide a step-by-step reasoning process. Finally, respond in the following format [...]

Figure 8: Chain of Thought Prompt to Axis 3 - ”Causal Reasoning Ladder” classification.

D Examples of Causal Seed Questions

Below we have some examples of causal seed questions, separated by each class of the three-axis taxonomy.

Causality		
Question(BR)	**Question(EN)**	**Class**
Vale a pena fazer o curso de Assistente Administrativo?	Is it worth taking the course of Administrative Assistant?	Causal
Como ganhar dinheiro sem trabalho?	How to make money without working?	Causal
Desabafo: por quê o povo é tão iludido ??	Outburst: why the people are so deluded??	Causal
Consigo fazer mestrado me graduando em EAD?	Can I take a Master's degree being graduated on distance learning?	Non-Causal
Você sente cansaço quando você está programando em projetos chatos?	Do you feel tired when you are programming boring projects?	Non-Causal
Quanto do seu salário você gasta com aluguel?	How much of your salary do you spend on rent?	Non-Causal

Table 8: Examples of Seed Causal / Non-Causal Questions, classified according to the Axis-1 of the taxonomy.

Class of Action		
Question(PT)	**Question(EN)**	**Class**
Por que sempre tem tanta vaga de QA?	Why are there always so many QA vacancies?	Cause-seek.
Qual é o perfil do usuário médio do Reddit?	What is the average Reddit user?	Cause-seek.
Gente, o que pode ser isso? Na orelha esquerda da minha gata?	What is thas? On the left ear of my cat?	Cause-seek.
Quais são os sinais de que um relacionamento é feliz e saudável?	What are the signs of a happy and healthy relationship?	Effect-Seek.
alguém aqui já deu a vacina v10 em cachorro filhote? Percebeu algum sintoma mesmo depois dos dias de efeitos colaterais?	Has anyone here ever given the v10 vaccine to a puppy? Did you notice any symptoms even after days of side effects?	Effect-Seek.
Quão importante é o currículo para seleção de mestrado?	How inportant is a CV for master's degree selection?	Relation-Seek.
Faz sentido clean architecture em frameworks como Rails e Laravel?	It makes any sense using clean architecture on frameworks like Rails and Laravel?	Relation-Seek.
É muito errado armazenar um token JWT no local/session storage?	Is it bad to storage a JWT token on local/session storage?	Relation-Seek.
Onde posso aprimorar meu conhecimento?	Where can I improve my knoledge?	Recomm.-Seek.
Quantas horas por semana eu deveria ocupar com aulas na minha grade?	How many hours per week should I be using for classes on my schedule?	Recomm.-Seek.
Focar em Django para a construção de sistemas web vale a pena?	Is focusing on Django for building Web Systems woth it?	Recomm.-Seek.
Como posso iniciar trabalhando com suporte tecnico?	How can I start working on technical support?	Steps-Seek.
Como estudar e trabalhar?	How to study and work?	Steps-Seek.
Como viver feliz tendo tão pouco?	How to live happy having less resources?	Steps-Seek.

Table 9: Examples of Seed Questions of the CaLQuest.PT, classified according to the Axis-2 of the taxonomy.

Pearl's Ladder of Causality		
Question(BR)	**Question(EN)**	**Class**
Como otimizar buscas por chamadas em aberto para publicação em revista?	How to optimize search for open calls for publications in magazines?	Associat.
Como vocês fazem pra não morder os lábios?	What do you do to not bite your lips?	Associat.
Como vermifugar meus gatos?	How to deworm my cats?	Associat.
Fazer mestrado ou não fazer mestrado?	Taking a master's degree or not?	Interven.
Minha primeira graduação: Ciência de Dados e I.A., ou Ciências Econômicas?	My first graduation: Data Science and A.I. or Economy Science?	Interven.
Largar o curso de medicina para ganhar 10k ou mais?	Give up my medicine school to earn 10k or more?	Interven.
Que conselho você daria para o seu eu do passado quando começou a aprender programação?	What advice would you give to your past self when you started learning programming?	Counterf.
Eu teria ótimas oportunidades de emprego com estes cursos no currículo + minha experiência?	Would I have great job opportunities with these courses on my resume + my experience?	Counterf.
Valeu a pena recusar a oportunidade ou cometi um erro?	Was it worth refusing the opportunity? Or did I make a mistake?	Counterf.

Table 10: Examples of Seed Questions of the CaLQuest.PT, classified according to the Axis-3 of the taxonomy.

E Linguistic Features and 5W-2H Question Analysis for the Golden Collection of Seed Questions

Feature	**Causal**	**Non-Causal**	**Total/Avg**
Samples	221	332	553
Avg. Words/Sample	23.25	36.03	31.04
Vocab Size	2,376	5.017	6,379
Type-Token Ratio	0.89	0.96	0.87

Table 11: Linguistic features in the Golden Collection of seed questions.

Question Type	**Causal**	**Non-Causal**	**Total**	**%**
What	120	188	308	55.7%
Who	2	8	10	1.8%
Why	13	2	15	2.7%
Where	10	9	19	3.4%
When	2	5	7	1.3%
How	57	42	99	17.9%
How much	5	11	16	2.9%
Others	12	67	79	14.3%
Total	221	332	553	100%

Table 12: Analysis of the question types 5W-2H in the Golden Collection of seed questions.