CaLQuest.PT: Towards the Collection and Evaluation of Natural Causal Ladder Questions in Portuguese for AI Agents

Uriel Lasheras[1], Vládia Pinheiro[1]
[1]Postgraduate Program in Applied Informatics, University of Fortaleza, Ceará, Brazil
Correspondence: uriel_andersonol@edu.unifor.br, vladiacelia@unifor.br

Abstract

Large Language Models (LLMs) are increasingly central to the development of generative AI across diverse fields. While some anticipate these models may mark a step toward artificial general intelligence, their ability to handle complex causal reasoning remains unproven. Causal reasoning is essential for true general intelligence, particularly at Pearl's interventional and counterfactual levels. In this work, we introduce CaLQuest.PT, a dataset of over 8,000 natural causal questions in Portuguese, collected from real human interactions. Built upon a novel three-axis taxonomy, CaLQuest.PT categorizes questions by causal intent, action requirements, and the level of causal reasoning needed (associational, interventional, or counterfactual). Our findings from evaluating CaLQuest.PT's seed questions with GPT-4o reveal that this LLM faces challenges in handling interventional and relation-seeking causal queries. These results suggest limitations in using GPT-4o for extending causal question annotations and highlight the need for improved LLM strategies in causal reasoning. CaLQuest.PT provides a foundation for advancing LLM capabilities in causal understanding, particularly for the Portuguese-speaking world.

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 and et al., 2024), Gemini (et al., 2024), and Claude (Anthropic, 2023)—the ability to understand cause-and-effect relationships and engage in causal reasoning is essential (Jin et al. (2023)). In Pearl and Mackenzie (2018), Pearl proposed the "Ladder of Causality" to categorize different levels of causal thinking. The first rung, Associational, consists of detecting correlations and patterns in observed data. LLMs already excel at this from their pre-training data. But in the higher Pearl's hung - 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) affirms 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 language. 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 et al. (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 the development of CaLQuest.PT[1], a dataset comprising more than

[1]https://github.com/ GhosTheKaos3150/CalQuest_PT -

Proceedings of the First Workshop on Language Models for Low-Resource Languages (LoResLM 2025), pages 325–343
January 20, 2025.

8,000 natural causal questions in Portuguese, collected from public sources and produced by humans in interactions with other humans and software systems. CaLQuest.PT is constructed based on a three-axis taxonomy, also proposed in this work, designed to capture the intent and action requirements in causal reasoning chains and the three rungs of causality defined by Pearl (Pearl and Mackenzie (2018)). We argue that the proposal of CaLQuest.PT, addressed here, is promising, as it will allow the evaluation and training of AI agents to identify when to apply cause-and-effect knowledge or reasoning (Axis 1: "Causal/Non-Causal"); to identify the requested action according to the interlocutor's intent (Axis 2: "Action Class"), and finally, to identify the level of reasoning needed by an AI causal solver (Axis 3: "Causal Reasoning" - associational, interventional and counterfactual). An additional contribution of this work is the annotation methodology, which follows a human-in-the-loop approach.

We evaluating the seed questions of the CaLQuest.PT using the LLM GPT4o with two prompt strategies and the findings indicated that GPT-4o struggles to assess the type of reasoning required for causal questions (particularly interventional 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). These results did not support the indiscriminate use of GPT-4o to extend annotation to additional natural questions of CaLQuest.PT.

2 Related Works

For the English language, we have datasets with completely artificially generated causal questions, such as WIQA (Tandon et al., 2019), Head-Line 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 contain questions between humans and LLMs, due to having been proposed before the explosion in popularity of LLMs. Especially, Jin et al. (2023) proposes 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) propose 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 Lange 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 CaLQuest.PT Data Collection and Annotation

To guide the development of a causal question dataset in Portuguese, we defined a three-axis taxonomy for causality inspired in Ceraolo et al. (2024) and Bondarenko et al. (2022). We then gathered a total of 8,041 natural questions from databases and repositories containing human-generated queries, 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. 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*

CalQuest.PT Source Code and Datasets

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 requested actions 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 the 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 "If I add fruit to the cake, will it be sweet?";

- *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 what a hypothetical scenario would be like about what we observe 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?".

Figure 1 presents a diagram illustrating the axes of the taxonomy used in the CaLQuest.PT dataset.

3.2 Causal Questions Collection and Annotation Process

To develop the CaLQuest.PT dataset, we aim to 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 first step was selecting public sources of human interactions. We chose three distinct sources, from which we collected three datasets totaling 8,041 questions (see the

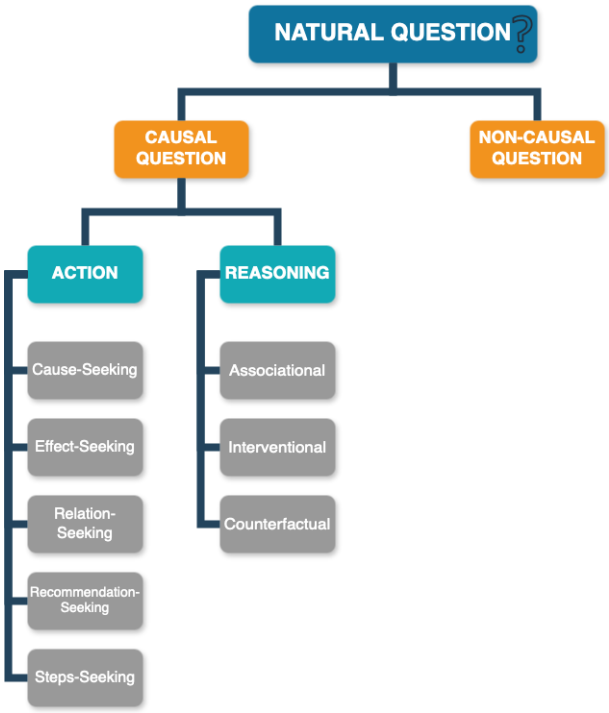

Figure 1: A Three-Axis Taxonomy of the CaLQuest.PT

Interaction Type	Datasets	#Samples
H-to-H	Reddit	3,541
H-to-LLM	ShareGPT	718
	WildChat	3,782
		8,041

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

datasets distribution in Table 1). The first set of natural questions was gathered from the question-and-answer forum Reddit[2], where interactions are Human-to-Human (H-to-H), using the Webcrawler from the Apify platform[3] and with the proper authorization from Reddit. The other two datasets are from sources where humans interact with LLMs (H-to-LLM): the dataset from WildChat (Zhang et al. (2023)), which contains data shared by ChatGPT users in the free service environment, and the ShareGPT[4] source, containing conversations with ChatGPT voluntarily shared by users.[5] The questions extracted from these datasets are predominantly formulated in the Brazilian dialect of Portuguese. However, a few isolated instances of questions in the European Portuguese dialect were identified, though they are not statistically significant. No questions written in other Portuguese dialects were found among the collected data.

3.2.1 Datasets Analysis

We analyze the datasets of the CaLQuest.PT in terms of its linguistic properties. Table 2 presents some linguistics features for each dataset. Overall, CaLQuest.PT has a good coverage of 8K 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) shows us the variety of words used for each question. On average, we have a high TTR value for the dataset, indicating 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. There is a prevalence of questions like "What" and "How, corresponding to 50.1% and 17.9% of the total questions, respectively. The type "Others" represents natural questions that do not follow the 5W-2H question pattern. Some examples are "Horror video reaction channels, no crime?" or "Urban life or rural life?". Analyzing the number of tokens per sample, we find that questions labeled as 'Others' are mostly below 100 tokens. This indicates that they do not represent the extensive LLM question group found in the dataset. Most of these questions are syntactically incorrect or ambiguous, which is why they could not fit into the 5W-2H question pattern.

3.2.2 CaLQuest.PT Annotation

The humam-in-the-loop approach to annotation of CaLQuest.PT followed the pipeline illustrated in Figure 2. A human-in-the-loop approach for linguistic corpus annotation combines the precision of human expertise with the efficiency of automated tools, enhancing annotation quality. This iterative process allows humans to correct model errors, ensuring higher reliability in ambiguous cases. Additionally, it supports continuous model improvement through feedback, leading to better performance in subsequent tasks.

In Step 1, 600 seed questions were selected equally from each dataset to initiate the annotation process for the entire CaLQuest.PT dataset using a human-in-the-loop approach and following the three-axis taxonomy (see Section 3.1). Details

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

[3]Apify Actor: https://apify.com/trudax/reddit-scraper-lite

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

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

Feature	Reddit	WildChat	ShareGPT	Total/Avg
Samples	37,82	3,541	718	8,041
Avg. Words/Sample	10.22	40.41	58.70	28.75
Vocab Size	6,110	20,693	10,210	32.393
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	1,649	1,929	445	4,023	50.1%
Who	145	44	11	200	2.5%
Why	269	107	12	388	4.8%
Where	137	161	24	322	4.0%
When	58	102	6	166	2.1%
How	685	640	118	1,443	17.9%
How much	113	49	7	169	2.1%
Others	485	750	95	1,330	16.5%
Total	3,541	3,782	718	8,041	100%

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

on linguistic features and analysis of 5W-2H question types are provided in Tables 4 and 5. In this selection, we preserve the general characteristics of the complete dataset.

In Step 2, a human annotator classified each of the 600 questions in each of the three axes of the taxonomy - Axis 1: "Causal/Non-Causal"; Axis 2: "Action Class"; and Axis 3: "Causal Reasoning". Table 6 presents the distribution of each dataset across each axis of the taxonomy. On Axis 1 - "Causal/Non-Causal", we can see that 37.4% of the seed questions are causal questions (224) and 62.6% are non-causal questions (376). Due to the nature of public sources, some human-generated questions lacked clear meaning. Examples include questions in formats such as, “I’ve had a migraine for three days. Help?”, or incomplete sentences like, “Why?” or “How?”. These questions were classified as non-causal, as they do not allow for the identification of a clear causal relationship. 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 can see the nature of natural causal questions. We can see that humans ask questions to other humans about subjective matters, like "Recommendation-seeking" questions, since the dataset Reddit (Human-To-Human) has more questions in this class (55, corresponding to 48.6% of the 113 causal questions). WildChat and ShareGPT datasets, which contain interactions between humans and LLMS, the humans ask mainly for algorithms or food recipes ("Steps-Seeking" questions), corresponding to 48.7% and 61.4%, respectively, of the 41 and 70 causal questions. Finally, in Axis 3 - "Causal Reasoning", according to Pearl’s Ladder of Causality, the most common class of questions to LLMs are in the rung "associational" (77.2% of the causal questions), and Counterfactual questions have low representation. Appendix D presents an exemplary list of natural questions for each class across all axes.

In Steps 3 and 4, we conducted one annotation cycle involving LLM-driven annotation and human review. In this first cycle, we used GPT-4o (OpenAI, 2024) with the initial aim of assessing how well one 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.

Feature	Causal	Non-Causal	Total/Avg
Samples	224	376	600
Avg. Words/Sample	28.14	33.88	31.74
Vocab Size	2,966	5,153	6,940
Type-Token Ratio	0.89	0.96	0.87

Table 4: Linguistic features in the 600 seed questions.

Question Type	Causal	Non-Causal	Total	%
What	110	204	314	52.3%
Who	2	9	11	1.8%
Why	14	4	18	3.0%
Where	12	11	23	3.8%
When	2	6	8	1.4%
How	60	44	104	17.4%
How much	6	11	17	2.8%
Others	18	87	105	17.5%
Total	224	376	600	100%

Table 5: Analysis of the question types 5W-2H in the 600 seed questions.

Classification	Reddit	WildChat	ShareGPT	Total	%
AXIS 1 - "Causal / Non-Causal"					
Causal	113	41	70	224	37.4%
Non-Causal	87	159	130	376	62.6%
	.	.	.	600	100.0%
AXIS 2 - "Action Class"					
Cause-Seeking	9	10	6	25	11.2%
Effect-Seeking	2	1	2	5	2.2%
Steps-Seeking	32	20	43	95	42.4%
Recommendation-Seeking	55	8	18	81	36.2%
Relation-Seeking	15	2	1	18	8.0 %
	.	.	.	224	100.0%
AXIS 3 - "Causal Reasoning"					
Associational	72	37	64	173	77.2 %
Interventional	38	3	1	42	18.7 %
Counterfactual	3	1	5	9	4.1 %
	.	.	.	224	100.0%

Table 6: Distribution of the seed questions of the CaLQuest.PT across our Three-axis Taxonomy.

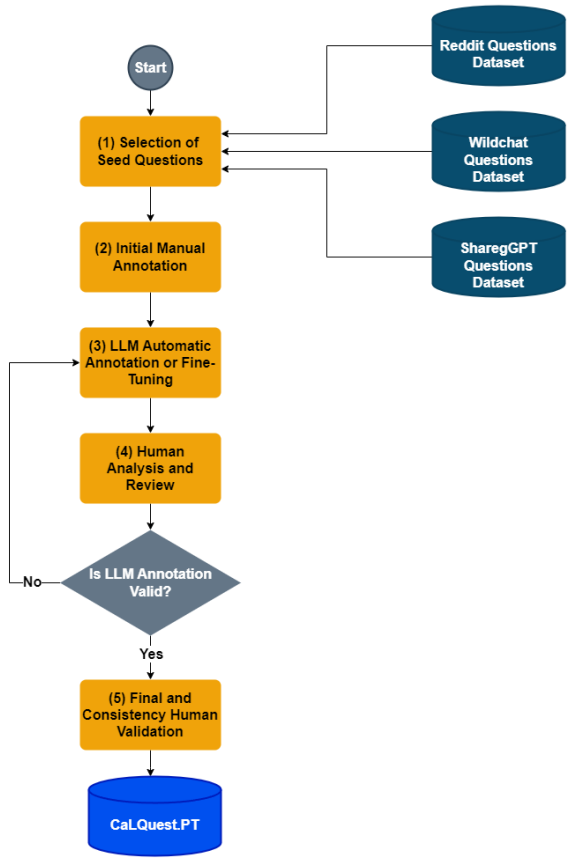

Figure 2: The humam-in-the-loop approach to CaLQuest.PT Annotation

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 - temperature (default value 1.0), top-t (default value 1.0), maximum number of tokens (no maximum value), among others; and the following prompt strategies - Few-shot Learning (Brown et al., 2020) and Chain-of-Thought (CoT) (Wei et al., 2022). The prompts in Portuguese, used in each axis of the taxonomy, are transcribed in Appendix A, B and C. Tables 7, 8 and 9 present the results in terms of Precision, Recall, and F1-Score of each prompt strategy for each classification axis.

LLM GPT-4o showed an interesting result in classifying causal and non-causal questions, achieving an F1-Score of 82.5% and 88.9%, respectively, when we used the Few-shot Learning prompt strategy. 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?"*. Contrary to our expectations, 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 the CoT version, GPT-4o incorrectly classified as non-causal, for example, the question *"How to make money without working?"* and incorrectly classified as causal the questions *"Am I being exploited, or is this the new normal?"*. This question is correctly classified in the Few-Shot Learning strategy. The first question is indeed causal, as it seeks a series of steps that would be the cause of a desired effect, namely "making money without working". The second question is indeed non-causal, as the human is not seeking causes/effects but rather opinions.

Evaluation Metrics	Causal	Non-Causal
Few-Shot Learning		
Precision	79.6%	91.1%
Recall	85.7%	86.9%
F1-Score	82.5%	88.9%
Chain-of-Thought		
Precision	81.4%	88.4%
Recall	80.3%	89.1&
F1-Score	80.9%	88.7%

Table 7: Classification Results of Seed Questions from CaLQuest.PT into Causal and Non-Causal Categories by GPT-4o Using Few-Shot Learning and Chain-of-Thought (CoT) Prompting Strategies.

In the second axis, LLM GPT4o also showed promising performance in classifying causal questions regarding action class, when we used the Few-Shot Learning prompt strategy. Its worst performance was in classifying questions in which the human sought to identify whether there is a cause-effect relationship between variables or events (Relation-Seeking), with F1-Score = 73.3%. The main reason for this was that LLM is confused with actions that search for causes or effects. For example, in a question like *"How important is a CV in a job interview?"*, although the question suggests a search action about a relationship between a good CV and a successful job interview, the LLM understands it as a search for a cause. Likewise, contrary

Evaluation Metrics	Cause-Seek.	Effect-Seek.	Steps-Seek.	Recomm.-Seek.	Rel.-Seek.
Few-Shot Learning					
Precision	95,6%	80,0%	90,9%	97,4%	91,7%
Recall	88,0%	80,0%	94,7%	92,6%	61,1%
F1-Score	91,6%	80,0%	92,8%	94,9%	73,3%
Chain of Thought					
Precision	78,5%	62,5%	88,9%	91,2%	100%
Recall	88,0%	100%	92,6%	90,1%	50,0%
F1-Score	83,0%	76,9%	90,7%	90,7%	66,7%

Table 8: Classification Results of Seed Questions from CaLQuest.PT into action classes by GPT-4o Using Few-Shot Learning and Chain-of-Thought (CoT) Prompting Strategies.

Evaluation Metrics	Associational	Interventional	Counterfactual
Few-Shot Learning			
Precision	93.7%	53.6%	80.0%
Recall	69.4%	80.4%	80.0%
F1-Score	79.7%	64.3%	80.0%
Chain of Thought			
Precision	94.6%	53.6%	100%
Recall	71.1%	80.4%	80.0%
F1-Score	81.1%	64.3%	88.9%

Table 9: Classification Results of Seed Questions from CaLQuest.PT into Pearl's Ladder of Causality by GPT-4o Using Few-Shot Learning and Chain-of-Thought (CoT) Prompting Strategies.

to what we predicted, the Chain of Thought (CoT) prompt strategy performed worse across all action classes. This is the case of the question *"How do you stay up to date with technology news?"*, incorrectly classified by the CoT prompt version as Recommendation-Seeking and, in fact, it is a Steps-Seeking question.

In axis 3 - Ladder of Causality, LLM GPT4o showed reasonable performance in recognizing the type of causal reasoning to be applied. The worst result was in the "Interventional" rung with F1-Score = 64.3% with very low precision = 53.6%, 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 have a correlation with the desired effect (entering the master's degree). The result in the "Counterfactual" rung is not conclusive due to the small number of seed examples (only 9 examples). Unlike the other axes, the CoT strategy showed a small improvement in results compared to the Few-Shot Learning prompt strategy.

5 Conclusion

This work presents an unprecedented proposal for a collection of causal questions, produced by humans in Portuguese - CaLQuest.PT, which aims to serve as a basis for evaluating and training AI agents to identify when to apply cause-and-effect knowledge or reasoning, to identify the requested action according to the interlocutor's intention, and finally, to identify the level of reasoning needed by an AI causal solver (rungs associational, interventional and counterfactual). We then proposed a three-axis Taxonomy and an annotation methodology, which follows a human-in-the-loop approach. CalQuest.PT will, therefore, serve to promote studies of AI agents with the capacity for causal commonsense reasoning in Portuguese, considered a low-resource language. We evaluated the LLM GPT4o in the classification of seed questions from CalQuest.PT, according to our three-axis taxonomy, and the findings indicated that GPT-4o struggles to assess the type of reasoning interventional and cause-and-effect relationships. These results did not support the indiscriminate use of GPT-4o to extend annotation to additional natural questions of CaLQuest.PT. In future works, we plan to ex-

plore other LLMs, like Open Source LLMs - Llama, Gemma e Phi, and fine-tuning processes to enhance results. The variation of examples in the Few-Shot Learning prompt strategy will also be a focus of future investigations, alongside efforts to measure the consistency, repeatability, and reproducibility of LLM responses.

5.1 Limitations and Challenges

The main obstacle in developing this work was obtaining questions in Portuguese with sufficient scope and representativeness, considering the various human-machine interaction scenarios. For example, it has not yet been possible to collect questions in Portuguese that humans ask in search engines, such as Bing and Google, due to the lack of public data in Portuguese on these platforms. As a strong premise of this work was to use sources and questions originally in Portuguese, to capture the pragmatics of the language and cultural nuances, we chose not to use translations of natural questions in English. Besides this, counterfactual questions do not seem to occur very frequently in the scenarios and environments used. Another challenge is the subjective and dubious nature of the questions and the consequent difficulty in including 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.

Another limitation of this study was the annotation process by a single annotator, which may introduce biases into the dataset and hinder a more detailed analysis of the ambiguity of the questions. The involvement of multiple annotators would allow for the evaluation of potential interpretation differences regarding the classification of a question as causal or not, enriching the analysis and contributing to greater robustness of the results. A multi-annotator approach is planned as a future enhancement of this linguistic resource.

References

Anthropic. 2023. Introducing claude.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Alexander Bondarenko, Magdalena Wolska, Stefan Heindorf, Lukas Blübaum, Axel-Cyrille Ngonga Ngomo, Benno Stein, Pavel Braslavski, Matthias Hagen, and Martin Potthast. 2022. CausalQA: A benchmark for causal question answering. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3296–3308, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Tom B. Brown, 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, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Roberto Ceraolo, Dmitrii Kharlapenko, Amélie Reymond, Rada Mihalcea, Mrinmaya Sachan, Bernhard Schölkopf, and Zhijing Jin. 2024. Causalquest: Collecting natural causal questions for ai agents. *Preprint*, arXiv:2405.20318.

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

Gemini Team et al. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Ilya Gusev and Alexey Tikhonov. 2022. HeadlineCause: A dataset of news headlines for detecting causalities. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6153–6161, Marseille, France. European Language Resources Association.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng LYU, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. 2023. Cladder: Assessing causal reasoning in language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 31038–31065. Curran Associates, Inc.

Zhijing Jin, Jiarui Liu, Zhiheng LYU, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona T. Diab, and Bernhard Schölkopf. 2024. Can large language models infer causation from correlation? In *The Twelfth International Conference on Learning Representations*.

Mayank Kejriwal, Henrique Santos, Alice M. Mulvehill, Ke Shen, Deborah L. McGuinness, and Henry Lieberman. 2024. Can ai have common sense? finding out will be key to achieving machine intelligence. *Nature*, 634:291–294.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners. *Preprint*, arXiv:2205.11916.

Nasrin Mostafazadeh, Aditya Kalyanpur, Lori Moon, David Buchanan, Lauren Berkowitz, Or Biran, and Jennifer Chu-Carroll. 2020. GLUCOSE: GeneraLized and COntextualized story explanations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4569–4586, Online. Association for Computational Linguistics.

OpenAI. 2024. Hello gpt-4o.

OpenAI and Josh Achiam et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

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

Niket Tandon, Bhavana Dalvi, Keisuke Sakaguchi, Peter Clark, and Antoine Bosselut. 2019. WIQA: A dataset for "what if..." reasoning over procedural text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6076–6085, Hong Kong, China. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *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*.

Li Zhang, Hainiu Xu, Yue Yang, Shuyan Zhou, Weiqiu You, Manni Arora, and Chris Callison-Burch. 2023. Causal reasoning of entities and events in procedural texts. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 415–431, Dubrovnik, Croatia. Association for Computational Linguistics.

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

A pergunta que se segue foi feita por um humano e você deve classificar esta pergunta em uma das categorias a seguir:

Categoria 1: Causal. Essa categoria inclui perguntas que implicam relações de causa e efeito em geral, sendo necessário uso de conhecimento de causa e efeito e raciocínio para obter uma resposta. Uma pergunta causal pode ter três tipos de objetivo ou comportamento. Podem ser (1) Dado a causa, prever o efeito: busca por entender o impacto ou desfecho de uma causa específica, que pode envolver em prever o futuro ou cenários hipotéticos (Exemplo: O que acontece se eu apostar na loteria? Eu deveria aprender uma nova linguagem de programação? Energias renováveis serão o futuro de nossa matriz energética? O que aconteceria se não existissem redes sociais?); (2) Dado um efeito, prever uma causa: Pergunta o "porquê" de algo ter ocorrido (Exemplo: Por que maçãs caem?), questionar a causa de um certo efeito, perguntar sobre a razão por trás de algo ou as ações que são necessárias para se obter um objetivo específico, como fazer algo, de forma implícita ou explícita (Exemplo: Por que movimentos extremistas estão aumentando ultimamente? Como ganhar um milhão de reais? Como posso aprender uma nova linguagem em 30 dias?). Isso também inclui casos onde o efeito não é explícito: qualquer pedido com um propósito, buscando por meios de cumpri-lo. Isso torna necessário achar a ação (causa) que melhor realizaria um certo objetivo (efeito), sendo este último podendo também ser implícito. Se alguém pede por recomendação de um restaurante, o que ele ou ela busca é a melhor causa (restaurante) para obter um certo efeito (Exemplo: comer saudável). Se busca por uma receita vegana, ele ou ela está buscando uma receita que seja a causa da melhor refeição possível. Perguntas requisitando "a melhor forma" de fazer alguma coisa se encaixam nessa categoria; (3) Dado um conjunto de variáveis, julgar a relação causal entre elas: Questiona a relação causal entre um conjunto de entidades (Exemplo: Fumar causa câncer? Eu fui rejeitado na entrevista de emprego porquê não tenho experiência?).

Categoria 2: Não-Causal. São perguntas que não implicam nenhuma das relações causais citadas anteriormente. Por exemplo, uma pergunta não causal pode ser um pedido de tradução, correção, para parafrasear um texto, para criar uma história, jogar um jogo, encontrar uma solução para um problema matemático, ou um enigma que requer um raciocínio matemático, prover alguma informação sobre algo (softwares, sites, endereços, eventos, locais em geral) ou usar tal informação para fazer uma comparação, sem muito raciocínio envolvido. Estas perguntas seriam não-causal, pois não o usuário está apenas buscando por uma informação.

Exemplos: \\
Pergunta: Qual o pior jeito de ganhar dinheiro? Categoria: <Causal> \\
Pergunta: Por que casas de apostas como tigrinho ou blazer não são derrubadas? Categoria: <Causal> \\
Pergunta: Tem alguma coisa que você faz por obrigação? Categoria: <Não-Causal> \\
Pergunta: Qual sua lembrança mais feliz? Categoria: <Não-Causal> \\
Pergunta: Qual é a temperatura adequada para um rack de pabx e de um rack de rede switch? Categoria: <Causal> \\
Pergunta: Qual será a reação química se "(NH2)2CO" for adicionado a "NaCl"? Categoria: <Causal> \\
Pergunta: Em que países atualmente vigoram a monarquia eletiva? Categoria: <Não-Causal> \\
Pergunta: Qual é a ultima versão do pytorch lançada? Categoria: <Não-Causal> \\
Pergunta: Qual é a melhor forma de retirar o fundo de uma fotografia no Photoshop? Categoria: <Causal> \\
Pergunta: Qual o trajeto de carro eu posso fazer entre São Paulo e Brasília? Categoria: <Causal> \\
Pergunta: Você consegue achar no nosso histórico de conversas por um assunto especifico? Categoria: <Não-Causal> \\
Pergunta: O que é um disco de vinil? Categoria: <Não-Causal> \\

Segue a pergunta: {PERGUNTA} e solicito que você classifique em uma das duas categorias acima detalhadas: Causal e Não-Causal; retornando a categoria e o raciocínio que justifica sua classificação, no seguinte formato:
CATEGORIA:
RACIOCÍNIO:

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").

> [...] solicito que você classifique em uma das duas categorias acima detalhadas: Causal e Não-Causal. Faça uma linha de raciocínio passo-a-passo. Ao final, responda no seguinte formato [...]

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

B Prompts to Axis 2 - "Action Class" classification

A pergunta que se segue foi feita por um humano. Essa pergunta é uma pergunta causal. Você deve classificar a pergunta em uma das seguintes categorias de ação:

Busca-Causa: Explica a causa que é origem de um determinado fenômeno. O indivíduo busca descobrir a causa ou justificativa para algo ser como é. Pode ser uma pergunta contendo um "Por quê" (Exemplo: Por quê as folhas caem no outono?; Por quê o céu é azul?). Ele também pode buscar entender uma explicação ou importância de uma sentença, ideia ou trabalho criativo, tal como o significado de letras musicais, poemas ou da narrativa de uma estória (Exemplo: Qual o significado da musica "Tempo Perdido"?, Quais são as principais causas de Alzheimer?). Resumindo, esse tipo de pergunta busca descobrir a causa ou justificativa para algo ser como é ou ter acontecido como aconteceu. Fórmula
Dado: efeito, Pede por: causa(s).

Busca-Efeito: Procura prever os efeitos de uma ação, ou prever o futuro dado circunstâncias do passado, ou ainda prever um cenário hipotético dado uma condição contrafactual (Exemplo: Energias renováveis serão nossa principal fonte de energia no futuro? Como o mundo seria se a internet não tivesse sido inventada?). Fórmula
Dado: causa(s), Pede por: efeito(s).

Busca-Relação: Questiona a relação de causa-e-efeito entre entidades distintas. O indivíduo busca entender se há uma relação de causa e efeito entre as entidades apresentadas na pergunta (Exemplo: Fumar causa câncer de pulmão? Poluição do ar pode aumentar o risco de doenças respiratórias?). Essa classe difere das classes "Busca-Causa" e "Busca-Efeito" pois quem questiona apresenta uma hipótese de causa e efeito, e se questiona se há relação causal na situação. Fórmula
Dado: conjunto de causas e efeitos, Pede por: relação causal.

Busca-Recomendação: Dado um objetivo implícito ou explícito e um conjunto de opções, pede para apontar a melhor opção para cumprir o objetivo (Exemplo: “Eu deveria tentar passar em um concurso para ter melhores chances de trabalho?”; “Qual a melhor pizzaria de Fortaleza?”). O indivíduo possui um objetivo e um conjunto de opções a sua escolha, e ele deseja escolher a opção que maximize os resultados do seu objetivo. Esta categoria se difere da “Busca-Passos” pois o individuo possui um conjunto de opções, e necessariamente deseja escolher a melhor delas. Fórmula
Dado: (efeito/propósito humano teleológico), Pede por: Guia que maximize os resultados(Satisfaz o propósito)

Busca-Passos: Propõe a solução de um problema por meio de um conjunto de passos ou algrítmo (Exemplo: “Como posso aprender inglês em 6 meses?”; “Crie uma receita vegana com batata-doce e feijão.”; “Otimize este código para que ele fique mais rápido.”). O indivíduo tem um propósito a ser cumprido, e deseja obter uma solução em forma de um conjunto de passos que possam ser seguidos. A resposta para essa pergunta pode ser tanto uma lista de passos, como um programa de computador, como uma receita. As perguntas podem tanto ter apenas uma forma de serem respondidas, como também ter mais de uma forma de atingir seu objetivo. Não implica a necessidade de ponderar as possibilidades e escolher a melhor entre elas. Fórmula
Dado: (efeito/propósito humano teleológico), Pede por: Causas em formato de guia passo a passo, código ou receita.

Lembre-se, em resumo: "Busca-Causa" busca descobrir a causa ou justificativa para algo ser como é ou ter acontecido como aconteceu; "Busca-Efeito" busca prever os efeitos de uma ação, ou prever o futuro dado circunstâncias do passado, ou ainda prever um cenário hipotético dado uma condição contrafactual; “Busca-Relação” busca entender se há uma relação de causa e efeito entre as entidades apresentadas na pergunta; “Busca-Recomendação” possui um objetivo e um conjunto de opções a sua escolha, e ele deseja escolher a opção que maximize os resultados do seu objetivo; “Busca-Passos” possui um propósito a ser cumprido, e deseja obter uma solução em forma de um conjunto de passos que possam ser seguidos;

Exemplos:

Pergunta: Porque as lojas dão desconto pra pagamento via pix, mas pra boleto não? Categoria: Busca-Causa
Pergunta: Quais são os sinais de que um relacionamento é feliz e saudável na opinião de vocês? Categoria: Busca-Efeito
Pergunta: Existe uma idade mínima ou ideal para aprender sobre política e economia? Categoria: Busca-Relação
Pergunta: Qual midia da mais liberdade criativa pro criador?livro,filme,serie ou quadrinho? Categoria: Busca-Recomendação
Pergunta: Como mudo meu nome no reddit? Categoria: Busca-Passos
Pergunta: Como acontecem invasões em sites que foram criados usando WordPress? Categoria: Busca-Causa
Pergunta: Os refrigerantes com 50\% de fruta são saudáveis? Categoria: Busca-Relação
Pergunta: Qual a melhor forma de usar o ChatGTP para criar conteúdo? Categoria: Busca-Recomendação
Pergunta: Como implementar uma instancia de leitura e uma de escrita do banco de dados no Laravel 7.4 Categoria: Busca-Passos

Segue a pergunta: {PERGUNTA} e solicito que você classifique em uma das cinco categorias acima detalhadas: Busca-Causa, Busca-Efeito, Busca-Relação, Busca-Recomendação e Busca-Passos; retornando a categoria e o raciocínio que justifica sua classificação, no seguinte formato:
CATEGORIA:
RACIOCÍNIO:

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").

[...] solicito que você classifique em uma das cinco categorias acima detalhadas: Busca-Causa, Busca-Efeito, Busca-Relação, Busca-Recomendação e Busca-Passos; Faça uma linha de raciocínio passo-a-passo. Ao final, responda no seguinte formato [...]

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

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

A pergunta que se segue foi feita por um humano. Essa pergunta é uma pergunta causal. Você deve classificar a pergunta em uma das seguintes categorias, de acordo com a Cadeia de Causalidade de Pearl:

Associacional: Esta categoria se refere a perguntas que levantam uma relação de associação estatística e correlação entre duas variáveis, questionando sobre a possibilidade de ocorrência de evento Y dado um evento inicial X. Exemplo disto são perguntas como "O que a rejeição na vaga nos diz sobre o candidato?" ou "Qual a melhor linguagem de programação para ciência de dados?". Essa associação pode ser explicita, como nos exemplos anteriores, como pode implicita como em "Estou com dor nos olhos e nas juntas, que doença poderia ser?", onde o interlocutor busca saber qual enfermidade que possuiria a maior correlação com os sintomas que ele ou ela sente. Também podemos ver isso em perguntas de recomendação, como "Quais os melhores investimentos de renda fixa para um estudante?", onde o interlocutor busca uma recomendação de investimento que tenha uma melhor correlação com seu perfil financeiro. Esse tipo de pergunta abrange vários formatos, como buscar métodos que tenham uma correlação com determinado fim (Exemplo: Como trabalhar em dois empregos?), buscar um local ou objeto que tenha uma correlação com uma necessidade do interlocutor (Exemplo: Onde posso ir para relaxar?) ou buscar um motivo que tenha correlação com um evento (Porquê as folhas estão ficando amareladas?).

Intervencional: Esta categoria contém perguntas que buscam entender para além de uma correlação entre dois eventos. Para isso, o indivíduo pergunta de forma a intervir no sistema, modificando ou adicionando uma ação para entender o efeito final dela. Exemplo disto são perguntas como "Se ela ganhar mais experiência de trabalho, ela será contratada?" ou "Se eu adicionar frutas ao bolo, ele ficará doce?". Esse tipo de pergunta pode também ser uma comparação entre opções, onde o interlocutor deseja saber qual das duas trará o melhor resultado, como em "Devo acordar mais cedo todos os dias e ter mais tempo ou acordar mais tarde e ficar mais descansado?". Ela também pode ser implicita, como em "Eu deveria comprar equipamento novo para meu trabalho?", onde o interlocutor deseja saber qual o impacto que realizar essa ação/intervenção terá em seu futuro.

Contrafactual: Esta categoria contém perguntas sobre realidades alternativas, modificando variáveis de um evento que já ocorreu para entender como ele ocorreu e que possíveis futuros poderiam ter ocorrido se alguma das variáveis envolvidas tivesse sido diferente. As perguntas causais contrafactuais geram hipóteses de outras possíveis causas. Exemplos deste tipo de pergunta são "Eu fui rejeitado por que não tinha experiência?" ou "Eu desenvolvi condromalácia por estar acima do peso?".

Exemplos de perguntas classificadas em uma das três categorias – Associacional, Intervencional, Contrafactual:
Pergunta: Por que entrevista de emprego virou tortura? Categoria: Associacional
Pergunta: Consigo fazer mestrado me graduando em EAD? Categoria: Intervencional
Pergunta: Eu teria ótimas oportunidades de emprego com estes cursos no currículo + minha experiência? Categoria: Contrafactual
Pergunta: Essa nova geração é realmente pior que a passada? Categoria: Associacional
Pergunta: Vocês acham que perderiam suas amizades se descobrissem tudo o que você pensa? Categoria: Intervencional
Pergunta: O que teria acontecido se nunca tivesse existido exploração no mundo? Categoria: Contrafactual

Segue a pergunta: {PERGUNTA} e solicito que você classifique em uma das três categorias acima detalhadas: Associacional, Intervencional ou Contrafactual; Caso não consiga classificar em uma delas, classifique como "None". Ao final, responda no seguinte formato:
CATEGORIA:
RACIOCÍNIO:

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").

> [...] Caso não consiga classificar em uma delas, classifique como "None". Faça uma linha de raciocínio passo-a-passo. Ao final, responda no seguinte formato [...]

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

D Examples of Seed Questions of the CaLQuest.PT

Below we have some examples of seed questions of the CaLQuest.PT, 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 mutch of your salary what do you spend on rent?	Non-Causal

Table 10: Examples of Seed Causal / Non-Causal Questions of the CaLQuest.PT, 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 11: 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 12: Examples of Seed Questions of the CaLQuest.PT, classified according to the Axis-3 of the taxonomy.