Journal of the Brazilian Computer Society, 2025, 31:1 doi: 10.5753/jbcs.2025.5825

Open LLMs Meet Causality in Portuguese: A Corpus-Based Fine-Tuning Approach

Uriel Lasheras [**University of Fortaleza** | ***uriel_andersonol@edu.unifor.br***]
Elioenai Alves [**University of Fortaleza** | ***l.oenaialves@edu.unifor.br***]
Caio Ponte [**University of Fortaleza** | ***caioponte@unifor.br***]
Carlos Caminha [**Kunumi Lab - Federal University of Ceará** | ***caminha@ufc.br***]
Vládia Pinheiro [**University of Fortaleza**| ***vladiacelia@unifor.br***]

University of Fortaleza, Av. Washington Soares, 1321, Edson Queiroz, Fortaleza, CE, 60811-905, Brazil

Received: 01 April 2025 • **Accepted:** 20 May 2025 • **Published:** 10 October 2025

Abstract Causal reasoning is a key component in the development of more robust, fair, and explainable language models. However, the ability of open-source Large Language Models (LLMs) to perform causal reasoning, especially in languages other than English, remains an open challenge. In this paper, we introduce an expanded version of CaLQuest.PT, a corpus of 2,500 natural questions in Portuguese designed to support multi-level causal evaluation. This dataset enables three layers of classification: (1) causal vs. non-causal questions, (2) causal action types such as cause-seeking, effect-seeking, and recommendation-seeking, and (3) reasoning types based on Pearl's Ladder of Causality—associational, interventional, and counterfactual. We also present an enhanced Few-Shot Learning prompting strategy and evaluate the performance of open-source models fine-tuned on this corpus. Our results show that, with targeted training and prompt design, smaller open-source LLMs can approach and even surpass the performance of larger models in several causal classification tasks. This study highlights the viability of corpus-based fine-tuning as a low-resource alternative for enhancing causal reasoning in open LLMs and advancing natural language understanding in Portuguese.

Keywords: Causal Reasoning, Open Source LLMs, Corpus-Based Fine-Tuning, Portuguese NLP

1 Introduction

Large Language Models (LLMs) have transformed not only the way we interact with and process language, but also the execution of tasks that demand extensive knowledge utilization. The inference abilities of very LLMs such as GPT-4o [OpenAI and et al., 2024], Llama 3 [Meta, 2024] and DeepSeek [DeepSeek-AI *et al*., 2025] form the foundation of their remarkable proficiency in understanding, processing, and responding to a wide range of inquiries. These abilities also underpin their adaptability across high-impact domains, including healthcare, legal decision-making, and customer service [Kejriwal *et al*., 2024]. Consequently, extensive research efforts have been dedicated to measuring and enhancing these capabilities, ranging from assessing the reasoning abilities of LLMs to scrutinizing their decision-making processes and addressing challenges such as concept alignment across different modalities and mitigating hallucination [Liu *et al*., 2024]. In many instances, language tasks require not only predicting or generating text based on patterns in the data but also understanding the underlying causal mechanisms driving these patterns. Causal inference has shown great potential in improving predictive accuracy, fairness, robustness, and explainability of Natural Language Processing (NLP) models [Feder *et al*., 2022]. Conversely, other authors affim that on the path toward artificial general intelligence, understanding cause-and-effect relationships and engaging in causal reasoning is essential [Jin *et al*., 2023]. As Jin *et al*. [2023] affirm, "these transformative developments raise the question of whether these machines are already capable of causal reasoning: *Do LLMs understand causality?*".

Pearl's "Ladder of Causality" [Pearl and Mackenzie, 2018] provides a structured framework to evaluate causal reasoning capabilities across three levels: (1) Associational, involving the detection of correlations and patterns within observed data; (2) Interventional, requiring an understanding of the effects of interventions on a system; and (3) Counterfactual, involving reasoning about hypothetical or unrealized scenarios. One of our hypotheses is that, although LLMs excel at the associational level due to their training on vast amounts of text data, their ability to navigate the higher rungs of causality—specifically interventional and counterfactual reasoning—remains a critical and unresolved challenge. Evaluating whether LLMs can effectively engage in these higher-order reasoning tasks is essential for advancing their capabilities toward genuine causal understanding.

The ability to classify the type of question—such as whether it is causal or non-causal, whether it seeks an effect, a cause, or a recommendation, and whether it involves causal reasoning of the associational, interventional, or counterfactual type—can significantly enhance an agent's ability to respond to causal queries more effectively. By understanding the underlying intent and reasoning structure of a question, the agent can select more appropriate knowledge sources, reasoning strategies, and response formats. This classification enables a more precise alignment between the user's informational needs and the system's capabilities, ultimately leading to more accurate and contextually relevant answers.

Despite the growing interest in evaluating LLMs' causal reasoning capabilities, significant gaps remain. Existing benchmarks predominantly focus on artificially constructed datasets, which do not capture the pragmatic nuances and linguistic diversity found in natural human queries [Ceraolo *et al.*, 2024]. Furthermore, while there has been progress in developing such benchmarks for English, the availability of datasets for evaluating causal reasoning in Portuguese remains critically limited, despite it being the sixth most spoken language in the world with approximately 270 million speakers.

Recognizing these limitations, our previous work introduced the CaLQuest.PT dataset [Lasheras *et al.*, 2025] [Lasheras and Pinheiro, 2025], which provided an initial golden collection of natural causal questions in Portuguese. However, the initial size of this dataset, only with 221 natural causal questions and the lack of tailored training procedures, limited the performance of open-source models, particularly smaller architectures such as Llama3.1-8B.

In this paper, we present an evolution of the CaLQuest.PT - the Golden Collection of Natural Causal Questions in Portuguese, with 2,500 natural questions, containing 1,085 causal questions. In this process, we enhance the prompting strategy using an improved Few-Shot Learning methodology. By providing more robust instructions and contextual examples, we aim to better guide the model's reasoning process across all three rungs of Pearl's Ladder of Causality.

With the expansion of CaLQuest.PT, we explored the potential of fine-tuning open-source Large Language Models (LLMs) using subsets of the newly augmented dataset. In particular, we focused on the Llama 3.1-8B model to investigate our central research question: *Can smaller open-source LLMs achieve competitive performance when properly fine-tuned?*. To address this question, we conducted two fine-tuning experiments using training sets composed of 800 and 1,728 causal and non-causal questions. The results revealed that the fine-tuned Llama 3.1-8B model significantly outperformed its original pre-trained version. Notably, it achieved classification scores that were comparable to those of much larger models, including Llama 3.1-70B, DeepSeek, and GPT-4o. In summary, The fine-tuning results demonstrate that small models, such as LLaMA 8B, can achieve competitive performance in causal classification tasks when trained on moderately sized supervised datasets (800 to 1,728 examples). Regarding the ability to classify questions as causal or non-causal, the fine-tuned model reached up to 84.11% F1-score on causal questions, significantly narrowing the gap with larger models like GPT-4o, LLaMA 70B, and DeepSeek. In the task involving a more fine-grained classification of causal questions by action type, substantial improvements were observed in categories such as Causal-Seeking and Effect-Seeking, with the fine-tuned model outperforming larger models in some cases. In the task of recognizing more complex types of causal reasoning based on Judea Pearl's Ladder of Causation (including the Associational, Interventional, and Counterfactual levels), the model fine-tuned with 1,728 examples achieved 80.59% in the Associational class and 64.36% in Interventional, surpassing all evaluated large models. However, performance in the Counterfactual class remained limited, likely due to the small number of training examples available for this category. All models were evaluated on the same test set (100 questions), ensuring fair comparisons and highlighting the potential of smaller models when paired with high-quality data and effective fine-tuning strategies.

These findings suggest that fine-tuning smaller, open-source LLMs with targeted datasets can bridge the performance gap with significantly larger models, both proprietary and open-source, with tens or even hundreds of billions of parameters. By employing efficient fine-tuning techniques on smaller models, we demonstrate that causal reasoning capabilities can be significantly enhanced using fewer computational resources. This presents a promising approach for deploying high-performance LLMs in resource-constrained environments, expanding accessibility and practical applicability in causal reasoning tasks.

The remainder of this paper is organized as follows. Section 2 reviews related work on causal reasoning in LLMs, covering both English and Portuguese datasets and benchmarks, as well as studies that explore how LLMs can contribute to the generation of causal knowledge. Section 3 revisits the Framework for a Natural Causal Golden Collection proposed in our previous work [Lasheras *et al.*, 2025], which introduced the initial version of CaLQuest.PT—a golden collection of natural causal questions in Portuguese. Section 4 presents the expanded version of the CaLQuest.PT dataset, detailing the annotation cycles supported by LLMs and the design of the prompting strategy. Section 5 describes the fine-tuning experiments conducted with a smaller open-source model, LLaMA 3.1-8B, including training procedures, evaluation metrics, and experimental results across the three axes of causal classification. This section also discusses the implications of the findings, emphasizing both the potential and limitations of smaller LLMs in causal reasoning tasks. Finally, Section 6 concludes the paper and outlines future research directions.

2 Related Works

This section organizes the related work into two main categories. The first category focuses on studies that propose datasets containing causal questions; these datasets are specifically designed to evaluate the ability of Large Language Models (LLMs) to identify different types of causality. The second category delves into studies that assess the capability of LLMs to generate causal arguments and knowledge.

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 but do not contain questions between humans and LLMs, due to having been proposed before the explosion in popularity of LLMs. Notably, CLadder [Jin *et al.*, 2023] was artificially built using a Causal Inference Engine that processes queries, graphs, and information based on Pearl's causality ladder.

Recently, the CAUSALQUEST dataset [Ceraolo *et al.*, 2024] was introduced, featuring entirely natural causal questions collected from Human-to-Human, Human-to-Search Engine, and Human-to-LLM interactions. The authors formalized the notion of causality using the definitions established in Pearl's Causal Hierarchy [Pearl and Mackenzie, 2018], and extended this framework by incorporating the concepts of Causal Chain [Schank, 1995; Bondarenko *et al.*, 2022]. According to their study, 42% of the English natural questions collected were classified as causal, with most of them aiming to identify the causes of specific effects. The authors also fine-tuned smaller language models for binary causal classification tasks, achieving an F1-score of 87.7% with the FLAN-T5-XL (LoRA) model fine-tuned for English. These results demonstrate strong potential for leveraging fine-tuning techniques to enhance language models in the identification and classification of causal questions.

In the context of the Portuguese language, recent studies have begun to explore the use of naturally occurring questions for building datasets aimed at computational understanding of causality [Lasheras and Pinheiro, 2025], [Lasheras *et al.*, 2025]. These works propose a structured three-axis taxonomy—including Pearl's Ladder of Causality [Pearl and Mackenzie, 2018]—and introduce a systematic methodology for the annotation and evaluation of a golden collection of natural causal questions in Portuguese. To the best of our knowledge, these are the first efforts to provide a solid methodological and conceptual foundation for the Portuguese-language NLP community to investigate and evaluate the causal reasoning capabilities of LLMs.

Regarding studies that aim to evaluate the extent to which LLMs can generate causal arguments and knowledge, Kıcıman *et al.* [2024] analyze the causal reasoning abilities of LLMs in tasks such as counterfactual reasoning and causal inference. This study benchmarks the causal reasoning abilities of LLMs through a behavioral analysis across multiple tasks. Results show that models like GPT-3.5 and GPT-4 can generate accurate causal arguments, outperforming traditional methods in tasks such as causal discovery, counterfactual reasoning, and event causality. These capabilities persist even on unseen data, indicating generalization beyond memorization. Despite occasional unpredictable errors, LLMs demonstrate human-like reasoning skills—such as identifying background causal context and constructing causal graphs—making them valuable tools for assisting experts in causal analysis. The study also highlights the potential of integrating LLMs with traditional causal inference techniques, as LLMs do not directly operate on raw data.

In the work by Wang [2024], the authors introduce a benchmark designed to evaluate whether LLMs are capable of performing causal reasoning across four dimensions of causality: cause-to-effect, effect-to-cause, cause-to-effect with intervention, and effect-to-cause with intervention. The study provides a robust assessment of various LLMs in terms of their causal reasoning abilities. Furthermore, Wang [2024] explores the relationship between causal reasoning and model hallucinations, showing that models with stronger causal reasoning capabilities tend to produce fewer hallucinations compared to smaller or less capable models.

3 A Framework for a Natural Causal Golden Collection

To guide the development of a Golden Collection (GC) with natural causal questions in Portuguese, we followed the 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 proposed in Lasheras *et al.* [2025]. In Lasheras *et al.* [2025], we present a detailed description of how the framework was used to gather a total of 7,594 natural questions from databases and repositories containing human-generated queries in Portuguese. The paper also describes the process of creating the first Golden Collection of 553 Natural Causal Questions, using a three-axis taxonomy and a human-in-the-loop approach.

In this section, we summarize that process, which served as the foundation for the work extended here.

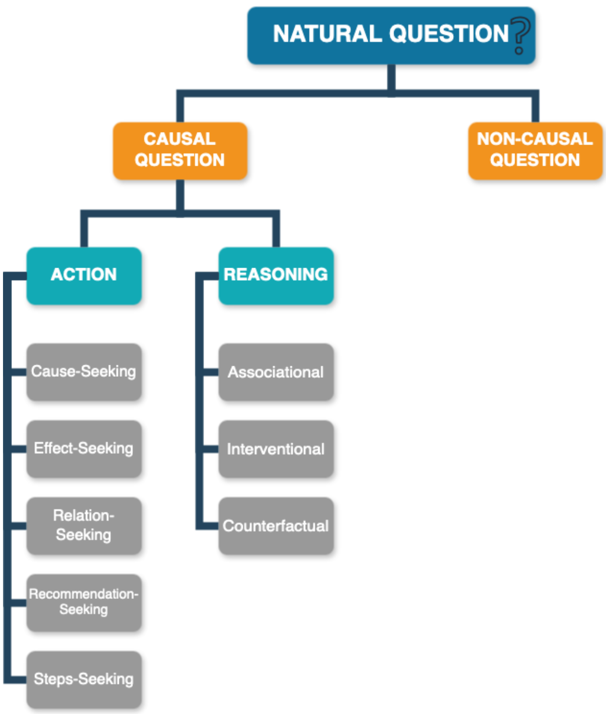

Figure 1. A Three-Axis Taxonomy for Causality Expression proposed in Lasheras *et al.* [2025].

3.1 A Three-Axis Framework for Causal Taxonomy

The 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, the focus is on general causal knowledge about events rather than specific instances where the cause-effect relationship is uniquely determined by a constrained scenario. For example, the question "What are common causes of floods in large cities?" seeks general causal knowledge about events, aiming to identify typical factors that lead to urban flooding, such as poor drainage systems, heavy rainfall, urban sprawl, or deforestation. In contrast, a question like ”What caused the flood in Rio Grande do Sul in 2024?” refers to a specific instance where the cause-effect relationship is tied to a unique and constrained context. In this work, our interest lies in the former type of question—those that involve generalizable causal patterns rather than singular, context-dependent events.

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. This definition of causal questions builds on 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, the causal questions are categorized 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 the 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, the taxonomy incorporates 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 Lasheras *et al.* [2025], the proposed pipeline was applied up to Step 6. In the present work, we build upon that approach to extend the Golden Collection with 2,000 additional natural questions, following Steps 7 through 9.

Steps 1 to 6 define the initial part of the pipeline for building the Golden Collection. In Step 1, questions are gathered from three types of communication sources—Human-to-Human, Human-to-Search Engine, and Human-to-LLMs—through automated filtering and manual review, followed by cataloging by source to capture their distinct linguistic patterns. Step 2 involves selecting a representative sample of natural questions to serve as seed data for annotation and guideline refinement. In Step 3, human annotators classify each seed question along three taxonomy axes: Causal vs. Non-Causal, Action Class, and Type of Causal Reasoning, as described in Section 3.1. Inter-annotator agreement is assessed in Step 4 using metrics such as Cohen’s Kappa, with a minimum threshold required for progression to the next stage. If this level is not reached, Step 5 involves reviewing disagreements and refining the guidelines through alignment sessions. Finally, in Step 6, after achieving strong agreement, a third-party reviewer adjudicates the annotated questions, resulting in the finalized Golden Collection of seed questions.

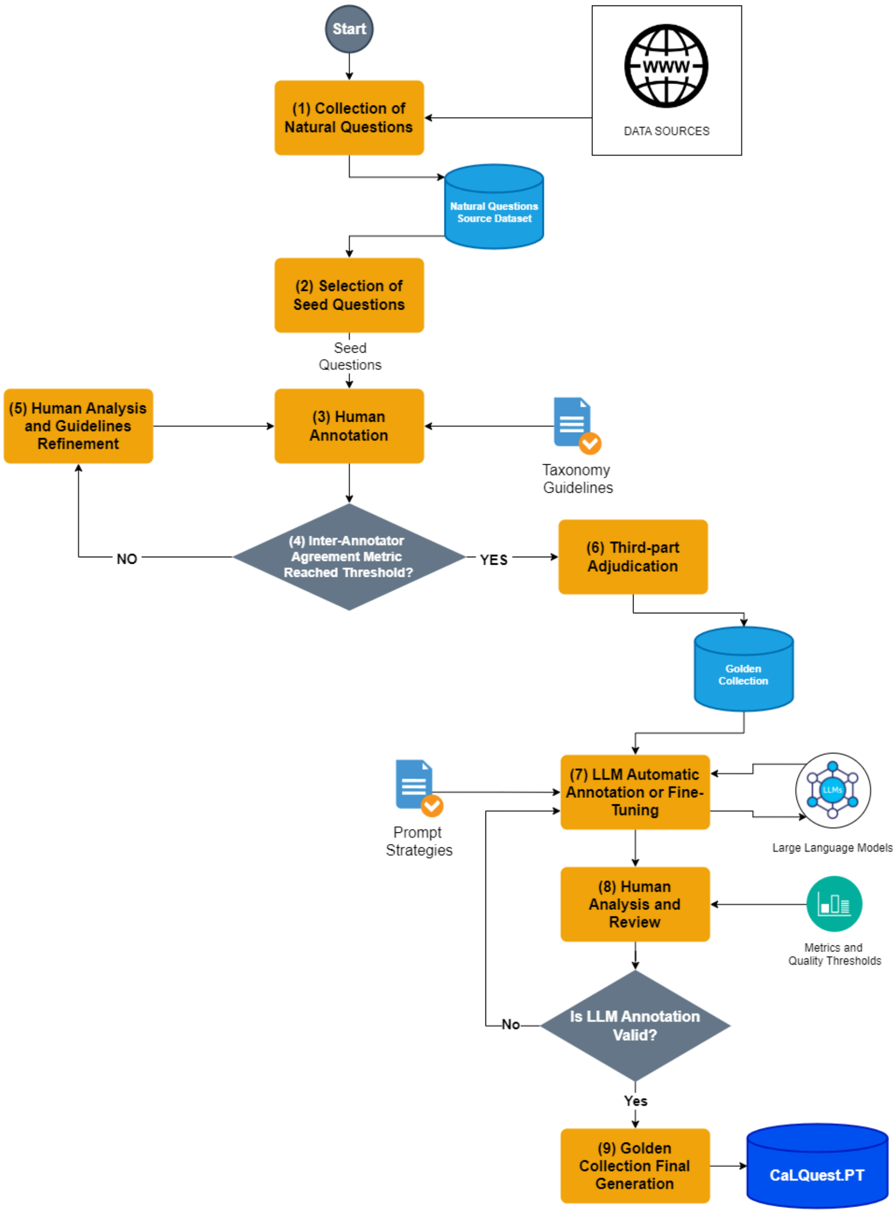

Figure 2. The Human-in-the-loop Approach to Annotation of Causal Datasets, proposed in Lasheras *et al*. [2025].

Steps 7 and 8 introduce evaluation cycles that combine LLM-driven annotation 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 of seed questions, 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. Finally, 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 CaLQuest.PT is provided in a machine-readable format (e.g., JSON) and includes the question identifier, its source, the question text, and the classification for each axis. Each classification entry, in turn, contains the assigned class per axis, the maturity level, and the reasoning generated by LLMs to support the classification decision. The classification maturity level can take the following values: 0, when the class is assigned by an LLM; 1, when the class is validated by a human reviewer; and 2, when the class is the result of a human annotation process involving at least two annotators and an adjudication step conducted by a third human reviewer.

3.3 A Golden Collection of Seed Natural Causal Questions in Portuguese

Lasheras *et al*. [2025] described the development of the Golden Collection of seed natural questions in Portuguese using this human-in-the-loop approach. In summary, according to Step 1, 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).

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 collected from the Reddit question and answer forum[1], where interactions are H-to-H. This data was obtained using Apify API[2] for extracting questions (namely, the title of Reddit's posts), with Reddit's consent to make this extraction. No user data was collected, and post's IDs were anonymized. Posts containing nudity, swear words and inappropriate content were also excluded from the data using flags available in the post's data. The other two datasets are from sources where humans interact with LLMs (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. Both datasets are available publicly at HuggingFace.[4]

Table 2 shows the distribution of the datasets by question type according to the 5W-2H question categorization. This classification permits 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, 553 seed questions were selected equally from each dataset. 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. Two iterations, through steps 3, 4, and 5, were conducted to achieve a satisfactory level of agreement, using Cohen's Kappa [Cohen, 1960]. After the second iteration, an agreement between annotators Kappa = 83.8 (Step 4) was achieved.

In Step 6, a third-party adjudication was performed in a few cases of divergence, resulting in the Golden Collection of seed questions. Table 3 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.

The initial CaLQuest.PT was therefore composed of a Golden Collection of 553 seed questions, in which each classification across the axes of the causal taxonomy was recorded with a maturity level of 2, as the class assignment resulted from a human annotation process involving at least two annotators and an adjudication step by a third human reviewer. Appendix E provides examples of natural questions for each class across all axes.

4 Expanding a Causal Question Golden Collection in Portuguese

The human annotation approach did not prove to be effective in building a robust dataset of natural causal questions in Portuguese. The generation process described in Lasheras *et al*. [2025] resulted in only 221 causal questions, with some classes in Axes 2 and 3 remaining underrepresented. For instance, in Axis 2, the classes Cause-Seeking and Relation-Seeking had only 24 and 21 examples each, while Axis 3 contained just 36 counterfactual questions (see Table 3).

As anticipated in the human-in-the-loop approach (see Section 3.2), expanding the Golden Collection of Natural Causal Questions requires the support of robust LLMs and in-context learning (ICL) prompt engineering techniques [Brown *et al*., 2020; Cui *et al*., 2024]. Recent studies have demonstrated the effectiveness of ICL strategies such as few-shot learning in information extraction tasks involving structured document analysis, as in [Almeida and Caminha, 2024].

In this work, we present the process of expanding the Causal Question Golden Collection in Portuguese. We completed two annotation and evaluation cycles that combined LLM-driven annotation with subsequent human analysis and review (Steps 7 and 8).

[1]Reddit: https://www.reddit.com (accessed on 12/11/2024)
[2]Apify API:https://apify.com (accessed on 06/18/2025)
[3]ShareGPT: https://huggingface.co/datasets/anon8231489123/ShareGPT_Vicuna_unfiltered (accessed on 12/11/2024)
[4]Data License: ShareGPT (Apache-2), WildChat (AI2 ImpACT - Low Risk), Reddit (Non-Commercial research only)

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 Natural Questions in Portuguese.

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 2. Analysis of 5W-2H question types in datasets of natural questions in Portuguese.

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 3. Distribution of the seed questions of the CaLQuest.PT across our Three-axis Taxonomy.

4.1 The First Cycle of LLM Annotation (Step 7) and Human Analysis and Review (Step 8)

In the first cycle, we used GPT-4o [OpenAI, 2024], LLaMA 3 [Meta, 2024], and DeepSeek [DeepSeek-AI *et al.*, 2025], with the goal of assessing how well three of the most advanced LLMs currently available could recognize the nature of the causal and non-causal natural questions. These models were evaluated using two prompting strategies: few-shot learning and chain-of-thought (CoT) [Cui *et al.*, 2024], applied in a test set with 300 seed questions (In this work, this test set is referred to as 300_TEST_SET). To ensure class balance, we used three sets of 100 seed questions—one for each axis—with equal representation of each class across the data. Table 4 presents the distribution of the test sets in each axis. The results of causal reasoning by the LLMs are presented in Tables 5, 6, and 7.

Table 5 presents the results for Axis 1. The DeepSeek model, using a Few-Shot prompt, achieved promising results, obtaining an F1-score of 90% for both the Causal and Non-Causal classes. Following closely, LLAMA 3.1-70B and GPT-4o, also using Few-Shot prompts, stood out with strong performance—scoring 89.13% and 88.89% for the Causal class, and 90.74% and 89.11% for the Non-Causal class, respectively. Overall, all evaluated models performed better with the Few-Shot prompt strategy and achieved results above the F1-score threshold of 80%.

For Axis 2, whose results are shown in Table 6, the best performance was achieved by GPT-4o with the Chain-of-Thought (CoT) prompt, reaching a weighted F1-score of 76.05%, closely followed by GPT-4o with the Few-Shot prompt (weighted F1-score = 75.63%).

In Axis 3, as shown in Table 7, GPT-4o using the Few-Shot (CoT) prompt achieved a weighted F1-score of 68.63%, slightly outperforming GPT-4o with the standard CoT prompt (weighted F1-score = 67.46%).

It is important to note that for Axis 2, only GPT-4o and DeepSeek reached performance above the F1-score threshold of 80%, and only in 2 out of 5 classes: Cause-Seeking and Steps-Seeking. In Axis 3, the only model to meet the desired performance threshold was GPT-4o in the Associational class. These findings reinforce our hypothesis that, in general, LLMs face significant limitations when attempting to engage in higher-level causal reasoning, particularly at the interventional and counterfactual levels. Due to its superior performance on Axis 3, the GPT-4o model with the Few-Shot prompt was selected for the expansion of the golden collection.

Figure 3 illustrates the structure of the Few-Shot Learning prompt used in the first evaluation cycle. The system_prompt consists of two main components: the first provides context and defines each category to be used in the classification task, while the second presents examples containing a list of questions ("Pergunta:") and their corresponding answers ("Resposta:"). Each answer includes both the assigned category and the reasoning used by the LLM to determine the classification.

- **Pergunta:** - Qual o pior jeito de ganhar dinheiro?;
- **Resposta:**
  - **categoria:** - 'Causal'
  - **raciocinio:** - 'Esta pergunta é Causal (2) Dado um efeito, prever uma causa, pois busca indentificar quais formas de ganhar dinheiro resultam num pior cenário possível...'

The user_prompt, in turn, contains the actual question ("Pergunta:") to be classified and specifies the required response format, which must include the fields "categoria" and "raciocínio". In the case of Chain of Thought (CoT) prompts, the structure remains the same, with the addition of an instruction encouraging step-by-step reasoning: *"Faça a linha de raciocínio passo a passo."*, (in English, "Think through the reasoning step by step.").

The full versions of the Few-Shot and CoT prompts used for the three classification axes are presented in Appendices A, B e C.

4.2 The Second Cycle of LLM Annotation (Step 7) and Human Analysis and Review (Step 8)

In the second evaluation cycle, our goal was to expand the Golden Collection using the best-performing model from the first cycle. Based on the results—and applying an F1-score threshold of ≥ 80%—we selected GPT-4o with the Few-Shot Learning (FS) prompt configuration in Step 7 as the most suitable LLM for this task. This model and prompt combination achieved the highest performance in 5 out of 8 classes from Axes 2 and 3, and for Axis 1, its performance was comparable to other leading open-source models such as DeepSeek and Llama 3.1-70B. Using this setup, we proceeded to expand the Golden Collection by automatically annotating an additional 2,000 natural questions.

Table 8 presents an analysis of these selected questions, categorized according to the 5W-2H typology (Who, What, When, Where, Why, How, and How much/How many). Interestingly, the distribution of question types among the 2,000 newly selected questions closely mirrors the distribution found in the original dataset of 7,594 natural questions, indicating the consistency and representativeness of the expansion process.

The results of the annotation process across the three taxonomy axes are summarized in Table 9. When comparing the classification distribution of the 2,000 newly added questions to that of the initial Golden Collection (553 questions) (see Table 3), the following patterns emerge:

- In Axis 1, the distribution remains consistent, with approximately 40% causal and 60% non-causal questions;
- In Axis 2, there is a higher proportion of Recommendation-Seeking questions and a lower proportion of Effect-Seeking questions;
- In Axis 3, Associational questions continue to account for around 63% of the total. However, Interventional questions nearly doubled in proportion, while Counterfactual questions became underrepresented, with only 30 additional questions of this type included in the expanded Golden Collection.

Classification	Total	%
AXIS 1 - Causal / Non-Causal		
Causal	50	39.9%
Non-Causal	50	60.1%
	100	100.0%
AXIS 2 - Action Class		
Cause-Seeking	20	20.0%
Effect-Seeking	20	20.0%
Steps-Seeking	20	20.0%
Recommendation-Seeking	20	20.0%
Relation-Seeking	20	20.0%
	100	100.0%
AXIS 3 - Causal Reasoning		
Associational	33	33.0%
Interventional	33	33.0%
Counterfactual	34	34.0%
	100	100.0%

Table 4. Distribution of the 300 seed questions, used as the test set referred to as 300_TEST_SET, across the Three-Axis Taxonomy.

LLM (Prompt)	Causal	Non-Causal
LLAMA3.1 70B (Few-Shot)	89.13%	90.74%
LLAMA3.1 70B (CoT)	84.44%	87.27%
DeepSeek (Few-Shot)	90.00%	90.00%
DeepSeek (CoT)	88.89%	89.11%
GPT-4o (Few-Shot)	88.89%	89.11%
GPT-4o (CoT)	87.50%	88.46%

Table 5. GPT-4o, Llama3.1 and DeepSeek classification results of 100 seed questions of the 300_TEST_SET into Causal and Non-Causal Categories (Axis-1), using Few-Shot Learning and Chain-of-Thought (CoT) Prompting Strategies. Positive class is Causal, and Negative Class is Non-Causal

LLM (Prompt)	Cause-Seek.	Effect-Seek.	Steps-Seek.	Recom-Seek.	Rel-Seek.	F1-Weighted
LLAMA3.1 70B (Few-Shot)	75.00%	61.11%	73.47%	68.18%	48.28%	65,21%
LLAMA3.1 70B (CoT)	75.00%	65.00%	73.47%	63.64%	51.85%	65,79%
DeepSeek (CoT)	89.47%	62.50%	80.00%	69.23%	70.97%	74,43%
GPT-4o (Few-Shot)	86.49%	62.86%	81.82%	72.00%	75.00%	75,63%
GPT-4o (CoT)	91.89%	61.11%	80.95%	72.00%	74.29%	76,05%

Table 6. GPT-4o, Llama3.1 and DeepSeek classification results of 100 seed questions of the 300_TEST_SET into Action Classes (Axis-2), using Few-Shot Learning and Chain-of-Thought (CoT) Prompting Strategies.

LLM (Prompt)	Associational	Interventional	Counterfactual	F1-Weighted
LLAMA3.1 70B (Few-Shot)	64,52%	61,54%	40,91%	55,51%
LLAMA3.1 70B (CoT)	65,57%	63,74%	41,67%	56,84%
DeepSeek (CoT)	68,97%	66,00%	38,10%	57,49%
GPT-4o (Few-Shot)	80,65%	70,33%	55,32%	68,63%
GPT-4o (CoT)	82,54%	68,13%	52,17%	67,46%

Table 7. GPT-4o, Llama3.1 and DeepSeek classification results of 100 seed questions of the 300_TEST_SET into Pearl's Ladder of Causality (Axis-3), using Few-Shot Learning and Chain-of-Thought (CoT) Prompting Strategies.

This imbalance reflects the limited ability of LLMs to accurately identify counterfactual questions, often misclassifying them as interventional.

It is important to note that the automatic annotations were initially assigned a maturity level of 0, indicating that they were generated solely by an LLM.

In Step 8 of the second cycle, we conducted a human evaluation of a sample of the 2,000 questions, focusing on the classes where the LLM showed the weakest performance — *Effect-Seeking, Recommendation-Seeking* and *Relation-Seeking*, into Axis 2; and *Interventional* and *CounterFactual* classes, into Axis 3. The errors identified during this review were adjudicated by a single human reviewer and reclassified with a maturity level of 1, indicating that they were reviewed and validated by a single human evaluator.

Moreover, we conducted an additional analysis using the LLM-as-a-Judge strategy, employing DeepSeek as the Judge LLM (Appendix E presents the prompt used). This analysis aimed to explore the feasibility of leveraging LLMs for evaluation tasks in the future, especially as the dataset is ex-

Question Type	**Reddit**	**WildChat**	**ShareGPT**	**Total**	**%**
What	404	483	477	1364	68.20%
Who	21	8	12	41	2.05%
Why	55	21	12	88	4.40%
Where	31	27	20	78	3.90%
When	13	19	7	39	1.95%
How	120	120	111	351	17.55%
How much	23	8	7	38	1.90%
Others	0	1	0	1	0.05%
Total	667	687	646	2000	100%

Table 8. Analysis of 5W-2H question types in 2,000 questions from the original dataset, selected for the expansion of the CaLQuest.PT Golden Collection.

Classification	**Reddit**	**WildChat**	**ShareGPT**	**Total**	**%**
AXIS 1 - Causal / Non-Causal					
Causal	450	136	278	864	43.20%
Non-Causal	214	551	371	1136	56.80%
	.	.	.	2000	100.0%
AXIS 2 - Action Class					
Cause-Seeking	49	49	25	123	14.24%
Effect-Seeking	40	19	10	69	7.99%
Steps-Seeking	76	50	167	293	33.91%
Recommendation-Seeking	222	16	67	305	35.30%
Relation-Seeking	59	5	10	74	8.56%
	.	.	.	864	100.0%
AXIS 3 - Causal Reasoning					
Associational	278	92	167	537	62.15 %
Interventional	160	31	160	297	34.38 %
Counterfactual	11	16	3	30	3.47 %
	.	.	.	864	100.0%

Table 9. Distribution of the 2,000 newly added CaLQuest.PT questions across the Three-Axis Taxonomy.

```
<system_prompt>
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....

Categoria 2: Não-Causal. ...

Exemplos:
Pergunta: Qual o pior jeito de ganhar dinheiro? Resposta: {'categoria': 'Causal', 'raciocinio': 'Essa pergunta é Causal (2) Dado um efeito, prever uma causa, pois busca identificar quais formas de ganhar dinheiro resultam no pior cenário possível. Para responder, é necessário avaliar diferentes métodos e julgá-los com base em fatores como risco, ética e esforço, estabelecendo uma relação de causa e efeito.'}
...

Solicito que você classifique em uma das duas categorias acima detalhadas: Causal e Não-Causal. Não responda a pergunta apresentada!
</system_prompt>

<user_prompt>
Pergunta: {PERGUNTA} Resposta: {'categoria': [classe], 'raciocinio': [raciocinio do LLM]}
</user_prompt>
```

Figure 3. Design of the Few-Shot Learning prompt applied to causal reasoning tasks in Axis 1: Classification of Causal versus Non-Causal questions.

pected to grow. Specifically, we focused on the 864 causal questions classified under Axis 3 (see Table 9). Out of these, 177 were judged by the Judge LLM as misclassified by GPT-4o (the LLM used as the classifier). Of these, approximately 135 (76% of 177) were originally labeled as Associational, suggesting that LLMs tend to prioritize correlation-based patterns in causal reasoning tasks. This behavior may lead to inconsistencies, particularly when dealing with more complex reasoning types, such as Counterfactuals. However, this finding requires further investigation as the dataset evolves. It is important to highlight that this evaluation was reviewed by human annotators to identify potential errors made by the Judge LLM, and this human-in-the-loop process will be refined in future work.

Figure 4 shows the final output for one of the questions classified by GPT-4o across the three axes of the taxonomy. Each question includes a classification for each axis, along with the corresponding reasoning used by the model to justify its decision. Additional relevant information is also stored in JSON format, such as the data source, the model used, the complete response generated by the model, and the maturity level. For reminder, the classification maturity level can take the following values: 0, when the class is assigned by an LLM; 1, when the class is validated by a human reviewer; and 2, when the class is the result of a hu- man annotation process involving at least two annotators and an adjudication step conducted by a third human reviewer.

5 Fine-Tuning Open Source LLM in Causal Reasoning in Portuguese

Large language models such as GPT-4o, LLaMA 70B and DeepSeek have demonstrated consistent performance in causal classification tasks, as shown by the results presented earlier. However, their practical adoption remains limited by factors such as high computational cost, dependence on proprietary APIs or deployment challenges in low-resource environments. In contrast, smaller open-source models offer significant advantages beyond computational efficiency. Their transparency, enabled by the ability to download, inspect, and modify the models, make them well-suited for scenarios that demand trust and control. Furthermore, local execution eliminates the need to transmit data to external servers, which favors privacy and data security, especially in sensitive contexts. Another crucial aspect is their adaptability to specific domains: through fine-tuning with labeled data, these models can be specialized for targeted tasks, achieving competitive performance at a substantially lower cost.

In this context, we investigate to what extent smaller open-source models (such as LLaMA 8B) can serve as viable alternatives to larger models when properly fine-tuned with high-quality annotated data. This section presents the experiments conducted with this objective and assesses whether smaller models, despite having fewer parameters, can achieve competitive performance in structured causal reasoning tasks. This analysis directly addresses the central research question raised in the introduction concerning the competitiveness of smaller models when properly adapted: *Can smaller open-source LLMs achieve competitive performance when properly fine-tuned?*.

```
{
 "unid": "0001",
 "question": "Afinal como arranjar um emprego?",
 "source": "reddit",
 "model": "gpt-4o",
 "human_response": "-",
 "llm_axis1_response": "Causal",
 "llm_axis1_reasoning": "Essa pergunta é Causal (2) Dado um efeito, prever uma causa, pois busca identificar quais ações ou estratégias (causas) são necessárias para alcançar o objetivo de conseguir um emprego (efeito). Para respondê-la, é necessário analisar diferentes abordagens, como networking, elaboração de currículo, preparo para entrevistas, entre outros, estabelecendo uma relação de causa e efeito.",
 "completion_axis1": "{'categoria': 'Causal', 'raciocinio': 'Essa pergunta é Causal (2) Dado um efeito, prever uma causa, pois busca identificar quais ações ou estratégias (causas) são necessárias para alcançar o objetivo de conseguir um emprego (efeito). Para respondê-la, é necessário analisar diferentes abordagens, como networking, elaboração de currículo, preparo para entrevistas, entre outros, estabelecendo uma relação de causa e efeito.'}",
 "llm_axis2_response": "Busca-Passos",
 "llm_axis2_reasoning": "A pergunta se encaixa na classe Busca-Passos. Ela busca uma solução prática e direta para o problema de como conseguir um emprego, ou seja, um conjunto de etapas ou estratégias que devem ser seguidas para alcançar esse objetivo.",
 "completion_axis2": "{'categoria': 'Busca-Passos', 'raciocinio': 'A pergunta se encaixa na classe Busca-Passos. Ela busca uma solução prática e direta para o problema de como conseguir um emprego, ou seja, um conjunto de etapas ou estratégias que devem ser seguidas para alcançar esse objetivo.'}",
 "llm_axis3_response": "Associacional",
 "llm_axis3_reasoning": "A pergunta é associacional porque busca entender métodos, estratégias ou fatores que estão correlacionados com a obtenção de emprego, ou seja, está tratando da relação entre ações ou contextos específicos e o resultado desejado de conseguir um emprego.",
 "completion_axis3": "{'categoria': 'Associacional', 'raciocinio': 'A pergunta é associacional porque busca entender métodos, estratégias ou fatores que estão correlacionados com a obtenção de emprego, ou seja, está tratando da relação entre ações ou contextos específicos e o resultado desejado de conseguir um emprego.'}",
 "maturity_level": 0
}
```


Figure 4. Exported JSON file containing the classification results for the question: "After all, how can one get a job?" (in Portuguese, "Afinal como arranjar um emprego")

The dataset used for fine-tuning consisted of 1,728 samples, evenly distributed between causal (864) and non-causal (864) instances. These samples were drawn from the CaLQuest.PT corpus and organized according to the proposed three-axis taxonomy. Table 10 summarizes the distribution of classes across Axes 1, 2, and 3 for both training configurations (800 and 1,728 samples), where the smaller dataset is a subset of the larger one.

For each of the three axes, we conducted two fine-tuning experiments: one using the 800-sample dataset and another using the 1,728-sample dataset. This setup was designed to assess how the size of the training set influences model performance. In Axis 1, the first model was fine-tuned using the complete set of 1,728 examples (864 causal and 864 non-causal) and trained for 3 epochs, limited by computational cost. The second model used the reduced set of 800 examples (400 causal and 400 non-causal) and was trained for 5 epochs. For Axes 2 and 3, both the 400 and 864-sample configurations were trained for 5 epochs. All training datasets were prepared according to the official LLaMA documentation, using the recommended formatting tags, including (begin_of_text, start_header_id, end_header_id, end_of_text) to ensure proper structuring of the prompt during the fine-tuning process.

The fine-tuning experiments were carried out using the

Classification	Dataset 800	Dataset 1728
AXIS 1 - Causal / Non-Causal		
Causal	400	864
Non-Causal	400	864
AXIS 2 - Action Class		
Causal-Seek	85	123
Effect-Seek	69	69
Steps-Seek	86	293
Recomm.-Seek	86	305
Relation-Seek	74	74
AXIS 2 - Action Class		
Associational	185	537
Interventional	185	297
Counterfactual	30	30

Table 10. Distribution of the training dataset across our Three-Axis Taxonomy.

PyTorch framework with default hyperparameter settings. Three distinct prompts were employed (one for each thematic axis) and, due to the length of some prompts, the maximum input context was set to 4,096 tokens. The training was executed on cloud platforms such as *Runpod* and *Vast.ai*, using NVIDIA H200 GPUs with 140 GB of VRAM. The duration of fine-tuning varied depending on the size of the dataset, ranging from 8 to 35 minutes per epoch.

Table 11 presents the fine-tuning results of Axis 1, which assesses the ability to classify questions as causal or non-causal. The fine-tuned LLAMA 8B models showed a significant performance improvement compared to the base version. While the non-fine-tuned LLAMA 8B achieved only 77.66% F1 in causal questions and 74.46% in non-causal ones, the FINETUNED 864 models reached 84.11% (causal) and approximately 81,17% (non-causal). These results significantly narrow the gap compared to large models such as GPT-4o, LLAMA 70B, and DeepSeek, which ranged between 88% and 90%. It is important to note that all models were evaluated using the same test set, consisting of 100 questions, which ensures a fair comparison between the fine-tuned small models and the large-scale models. These results indicate that even with moderately sized supervised datasets (800 or 1,728 examples), small models can achieve highly competitive performance.

For Axis 2, which involves a more fine-grained classification of causal questions by action type, the performance improvement after fine-tuning was also notable, as shown in Table 12. The base version of LLaMA 8B performed poorly in several categories, with F1 scores of 17.4% in Relation-Seeking and 48.88% in Recommendation-Seeking. However, after fine-tuning with 400 and 864 examples of causal questions, the model surpassed LLaMA 70B (COT) in multiple categories. In Causal-Seeking questions, both fine-tuned models achieved F1 scores above 77%, closely approaching the performance of large models. In the Effect-Seeking category, the 864-sample model reached 63.15%, outperforming DeepSeek (62.5%) and GPT-4o (61,11%). These results suggest that, with carefully annotated and well-balanced training data, small models can be optimized to match (and in some cases exceed) the performance of much larger alternatives.

In Axis 3 (Table 13), based on Judea Pearl's Ladder of Causation (with the associational, interventional, and counterfactual levels), small models also showed remarkable progress. The base LLAMA 8B achieved only 50.0% F1 in the associational class, but its fine-tuned version with 864 examples (of causal questions) reached 80.59%, outperforming even the LLAMA 70B COT (65.57%) and DeepSeek COT (68.97%), ranking just behind GPT-4o FS (80.65%). At the interventional level, the fine-tuned model with 864 examples reached 64.36%, a performance superior to all evaluated large models. In the counterfactual class, however, improvements were limited. The fine-tuned model with 864 examples performed identically to the base model, both reaching 52.17%, while the version trained on 400 examples (of causal questions) obtained a lower score of 45.45%. These results remain below the performance of the GPT-4o model (70.33%), but despite being low, the LLAMA 8B FINE-TUNED 864 (52.17%) performed better than larger models such as LLaMA 70B (41.67%) and DeepSeek (38.1%). This limitation is likely related to the small number of training examples available for this class (only 30 instances were used during fine-tuning). As with the other axes, all models were evaluated on the same test set of 100 questions, ensuring fair comparisons across model sizes.

The results presented consistently demonstrate that smaller open-source models, such as LLAMA 8B, can achieve competitive performance in complex causal reasoning tasks when subjected to a careful supervised fine-tuning process. In several scenarios, the fine-tuned models not only significantly reduced the gap with large-scale models but also outperformed some of them in specific tasks. The use of the same test set to evaluate all models in each axis further strengthens the reliability of the comparison. These findings reinforce the feasibility of using small LLMs in contexts with computational or budgetary constraints, provided there is access to representative and well-annotated datasets.

6 Conclusion

The research questions that guided this study were: *"Do LLMs understand causality?"* and *"Can smaller open-source LLMs achieve competitive performance when properly fine-tuned?"*. In this context, the present work introduces an expanded version of CaLQuest.PT, a golden collection in Portuguese composed of 2,500 natural causal and non-causal questions, annotated according to a structured three-axis taxonomy: causality, action type, and reasoning type. The creation of this resource fills a relevant gap in the causal AI literature by providing a dataset in Portuguese suitable for evaluating language models on tasks that require causal inference. The dataset was built through a structured annotation process, combining human validation and support from LLMs, resulting in a diverse collection that spans associational, interventional, and counterfactual questions. This foundation enabled the execution of comparative experiments involving different language models, including both proprietary and open-source LLMs, with and without supervised fine-tuning.

Using CaLQuest.PT, fine-tuning experiments were conducted with the LLaMA 8B model, whose results demonstrate that smaller-scale open-source models can achieve competitive performance compared to larger models when fine-tuned with high-quality supervised data. In particular, the model fine-tuned with 864 examples achieved an F1 score of 80.59% in the associational class, outperforming larger models such as LLaMA 70B (65.57%) and DeepSeek (68.97%). At the interventional level, the same model reached 64.36% F1, again surpassing all larger models evaluated. These results indicate that even with moderately sized datasets, it is possible to adapt smaller models to perform well on complex causal classification tasks.

The study also revealed that model performance varies significantly across reasoning levels. While gains were substantial in the associational and interventional classes, counterfactual reasoning remains a challenge, with all models (including large-scale ones) showing more modest results. Nonetheless, the experiments showed that the performance gap between small and large models can be significantly

Model	Causal (F1)	Non-Causal (F1)
Large Models		
GPT-4o (Few-Shot)	88,89%	89,11%
LLAMA 70B (Few-Shot)	89,13%	**90,74%**
DeepSeek (Few-Shot)	**90,0%**	90,0%
Small Models		
LLAMA 8B (Base Version) (Few-Shot)	77,66%	74,46%
LLAMA 8B FT 400 (Few-Shot)	**84,11%**	80,43%
LLAMA 8B FT 864 (Few-Shot)	**84,11%**	**81,17%**

Table 11. Fine-Tuning Results for Axis 1 (Causal vs. Non-Causal). The table presents F1-Score values for the classification task on the first axis of the taxonomy. For large-scale models (GPT-4o, LLAMA 70B, and DeepSeek), the best results per model type are reported, based on the experiments presented in Table 5. For smaller models, we report the performance of LLAMA 8B in its base version and after supervised fine-tuning with 800 and 1,728 examples.

Model	Causal-Seek (F1)	Effect-Seek (F1)	Steps-Seek (F1)	Recomm.-Seek (F1)	Relation-Seek (F1)
Large Models					
GPT-4o (CoT)	**91,89%**	61,11%	**80,95%**	**72,00%**	**74,29%**
LLAMA 70B (CoT)	75,00%	**65,00%**	73,47%	63,64%	51,85%
DeepSeek (CoT)	89,47%	62,50%	80,00%	69,23%	70,97%
Small Models					
LLAMA 8B (Base Version) (Few-Shot)	57,14%	57,14%	65%	48,88%	17,39%
LLAMA 8B FINETUNED 400 (Few-Shot)	**79,06%**	52,63%	63,15%	56,00%	51,61%
LLAMA 8B FINETUNED 864 (Few-Shot)	77,77%	**63,15%**	**71,11%**	**64,00%**	**58,06%**

Table 12. Fine-Tuning Results for Axis 2. The table presents F1-Score values for the classification task on the second axis of the taxonomy. For large-scale models (GPT-4o, LLAMA 70B, and DeepSeek), the best results per model type are reported, based on the experiments presented in Table 6. For smaller models, we report the performance of LLAMA 8B in its base version and after supervised fine-tuning with 400 and 864 examples of causal questions.

Model	Associational (F1)	Interventional (F1)	Counterfactual (F1)
Large Models			
LLAMA 70B (CoT)	65,57%	63,74%	41,67%
DeepSeek (CoT)	68,97%	**66,0%**	38,1%
GPT-4o (Few-Shot)	**80,65%**	55,32%	**70,33%**
Small Models			
LLAMA 8B (Base Version) (Few-Shot)	50,0%	61,7%	**52,17%**
LLAMA 8B FINETUNED 400 (Few-Shot)	68,96%	61,22%	45,45%
LLAMA 8B FINETUNED 864 (Few-Shot)	**80,59%**	**64,36%**	**52,17%**

Table 13. Fine-Tuning Results for Axis 3. The table presents F1-Score values for the classification task on the third axis of the taxonomy. For large-scale models (GPT-4o, LLAMA 70B, and DeepSeek), the best results per model type are reported, based on the experiments presented in Table 7. For smaller models, we report the performance of LLAMA 8B in its base version and after supervised fine-tuning with 400 and 864 examples of causal questions.

narrowed, highlighting viable and accessible specialization strategies with practical implications for real-world applications that demand robust causal reasoning.

6.1 Limitations and Challenges

Although the results demonstrate the promising potential of fine-tuning smaller language models for causal classification tasks in Portuguese, our study faces certain limitations and highlights important challenges for future research.

Firstly, one limitation lies in the scarcity and imbalance of data for specific causal classes within the CaLQuest.PT dataset. As observed in the results for Axis 3 (Pearl's Ladder of Causality), the performance in classifying Counterfactual questions was notably lower and showed little improvement with fine-tuning. This is directly correlated to the extremely limited number of counterfactual examples (only 30 instances) available in the training sets of 800 and 1,728 samples. While the Associational and Interventional classes benefited from the data increase, the lack of sufficient examples for the counterfactual level prevented the fine-tuned Llama 8B model from reaching the performance levels of larger models or even significantly improving compared to its base version in this specific category. This scenario underscores the inherent challenge of collecting a substantial volume of natural examples that adequately represent the higher and more complex levels of causal reasoning, which may be less frequent or harder to articulate explicitly in natural language.

Secondly, related to the previous point, the overall size and diversity of the fine-tuning dataset, although expanded compared to previous work, may still impose limits on the model's robustness and generalization capability. Despite using up to 1,728 examples in fine-tuning, ensuring that this volume is sufficient to cover the wide range of linguistic nuances, knowledge domains, and structural complexities inherent in naturally formulated causal questions remains a challenge. The model's performance on causal questions that differ substantially (in style, topic, or implicit complexity) from those present in CaLQuest.PT might not reach the same levels observed in our test set. Further increasing the size and, crucially, the diversity of the dataset is fundamental for improving generalization.

Faced with these challenges of collecting and ensuring the availability of annotated data, the question arises of exploring strategies for knowledge distillation, where larger and more capable models (like GPT-4o or Llama 70B, which demonstrated better performance) could be used to generate synthetic training examples for the smaller models. The idea would be to instruct these larger models to produce questions (and their correct causal classifications) that could artificially augment the fine-tuning datasets, especially for underrepresented classes like the Counterfactual one. However, this approach introduces its own set of complex challenges. Ensuring the high quality, diversity, and causal fidelity of the synthetically generated examples is non-trivial; larger models might perpetuate or even amplify biases, or generate examples that do not genuinely reflect the complexity of the intended causal reasoning. Developing robust methodologies to control the quality of this synthetic generation and evaluate its real impact on the smaller model's learning constitutes an active and necessary area of research, which was not addressed within the scope of this work, but represents a promising avenue to mitigate data limitations.

Declarations

Acknowledgements

The authors would like to thank Fundação Edson Queiroz for providing the computational infrastructure necessary for conducting the experiments presented in this study. All experiments were carried out using resources made available through the Núcleo de Ciência de Dados e Inteligência Artificial (NCDIA) of the University of Fortaleza.

Funding

This study was funded by Fundação Edson Queiroz, as part of its ongoing support for academic research and innovation.

Authors' Contributions

VP, CC, and UL contributed to the conception and design of the study. All authors equally participated in conducting the experiments, performing validation, and writing the manuscript. All authors read and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Availability of data and materials

CalQuest.PT source code, datasets and documentation are available on https://github.com/GhosTheKaos3150/CalQuest_PT

References

Almeida, F. C. and Caminha, C. (2024). Evaluation of entry-level open-source large language models for information extraction from digitized documents. In *Symposium on Knowledge Discovery, Mining and Learning (KDMiLe)*, pages 25–32. SBC. DOI: 10.5753/kdmile.2024.243859.

Bondarenko, A., Wolska, M., Heindorf, S., Blübaum, L., Ngonga Ngomo, A.-C., Stein, B., Braslavski, P., Hagen, M., and Potthast, M. (2022). CausalQA: A benchmark for causal question answering. In *COLING:2022:1*, pages 3296–3308, Gyeongju, Republic of Korea. International Committee on Computational Linguistics. Available at: `https://aclanthology.org/2022.coling-1.291/`.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. DOI: 10.48550/arXiv.2005.14165.

Ceraolo, R., Kharlapenko, D., Reymond, A., Mihalcea, R., Sachan, M., Schölkopf, B., and Jin, Z. (2024). Causalquest: Collecting natural causal questions for ai agents. DOI: 10.48550/arxiv.2405.20318.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46. DOI: 10.1177/001316446002000104.

Cui, Y., He, P., Tang, X., He, Q., Luo, C., Tang, J., and Xing, Y. (2024). A theoretical understanding of chain-of-thought: Coherent reasoning and error-aware demonstration. DOI: 10.48550/arXiv.2410.16540.

DeepSeek-AI, Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., Dai, D., Guo, D., Yang, D., Chen, D., Ji, D., Li, E., Lin, F., Dai, F., Luo, F., Hao, G., Chen, G., Li, G., Zhang, H., Bao, H., Xu, H., Wang, H., Zhang, H., Ding, H., Xin, H., Gao, H., Li, H., Qu, H., Cai, J. L., Liang, J., Guo, J., Ni, J., Li, J., Wang, J., Chen, J., Chen, J., Yuan, J., Qiu, J., Li, J., Song, J., Dong, K., Hu, K., Gao, K., Guan, K., Huang, K., Yu, K., Wang, L., Zhang, L., Xu, L., Xia, L., Zhao, L., Wang, L., Zhang, L., Li, M., Wang, M., Zhang, M., Zhang, M., Tang, M., Li, M., Tian, N., Huang, P., Wang, P., Zhang, P., Wang, Q., Zhu, Q., Chen, Q., Du, Q., Chen, R. J., Jin, R. L., Ge, R., Zhang, R., Pan, R., Wang, R., Xu, R., Zhang, R., Chen, R., Li, S. S., Lu, S., Zhou, S., Chen, S., Wu, S., Ye, S., Ye, S., Ma, S., Wang, S., Zhou, S., Yu, S., Zhou, S., Pan, S., Wang, T., Yun, T., Pei, T., Sun, T., Xiao, W. L., Zeng, W., Zhao, W., An, W., Liu, W., Liang, W., Gao, W., Yu, W., Zhang, W., Li, X. Q., Jin, X., Wang, X., Bi, X., Liu, X., Wang, X., Shen, X., Chen, X., Zhang, X., Chen, X., Nie, X., Sun, X., Wang, X., Cheng, X., Liu, X., Xie, X., Liu, X., Yu, X., Song, X., Shan, X., Zhou, X., Yang, X., Li, X., Su, X., Lin, X., Li, Y. K., Wang, Y. Q., Wei, Y. X., Zhu, Y. X., Zhang, Y., Xu, Y., Xu, Y., Huang, Y., Li, Y., Zhao, Y., Sun, Y., Li, Y., Wang, Y., Yu, Y., Zheng, Y., Zhang, Y., Shi, Y., Xiong, Y., He, Y., Tang, Y., Piao, Y., Wang, Y., Tan, Y., Ma, Y., Liu, Y., Guo, Y., Wu, Y., Ou, Y., Zhu, Y., Wang, Y., Gong, Y., Zou, Y., He, Y., Zha, Y., Xiong, Y., Ma, Y., Yan, Y., Luo, Y., You, Y., Liu, Y., Zhou, Y., Wu, Z. F., Ren, Z. Z., Ren, Z., Sha, Z., Fu, Z., Xu, Z., Huang, Z., Zhang, Z., Xie, Z., Zhang, Z., Hao, Z., Gou, Z., Ma, Z., Yan, Z., Shao, Z., Xu, Z., Wu, Z., Zhang, Z., Li, Z., Gu, Z., Zhu, Z., Liu, Z., Li, Z., Xie, Z., Song, Z., Gao, Z., and Pan, Z. (2025). Deepseek-v3 technical report. DOI: 10.48550/arxiv.2412.19437.

Du, L., Ding, X., Xiong, K., Liu, T., and Qin, B. (2022). e-CARE: a new dataset for exploring explainable causal reasoning. In *ACL:2022:long*, pages 432–446, Dublin, Ireland. acl. DOI: 10.18653/v1/2022.acl-long.33.

Feder, A., Keith, K. A., Manzoor, E., Pryzant, R., Sridhar, D., Wood-Doughty, Z., Eisenstein, J., Grimmer, J., Reichart, R., Roberts, M. E., Stewart, B. M., Veitch, V., and Yang, D. (2022). Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158. DOI: 10.1162/tacl$_{a0}$0511.

Gusev, I. and Tikhonov, A. (2022). HeadlineCause: A dataset of news headlines for detecting causalities. In *LREC:2022:1*, pages 6153–6161, Marseille, France. European Language Resources Association. DOI: 10.48550/arXiv.2108.12626.

Jin, Z., Chen, Y., Leeb, F., Gresele, L., Kamal, O., LYU, Z., Blin, K., Gonzalez Adauto, F., Kleiman-Weiner, M., Sachan, M., and Schölkopf, B. (2023). Cladder: Assessing causal reasoning in language models. In Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., and Levine, S., editors, *Advances in Neural Information Processing Systems*, volume 36, pages 31038–31065. Curran Associates, Inc.. DOI: 10.48550/arXiv.2312.04350.

Jin, Z., Liu, J., LYU, Z., Poff, S., Sachan, M., Mihalcea, R., Diab, M. T., and Schölkopf, B. (2024). Can large language models infer causation from correlation? In *The Twelfth International Conference on Learning Representations*. DOI: 10.48550/arXiv.2306.05836.

Kejriwal, M., Santos, H., Mulvehill, A. M., Shen, K., McGuinness, D. L., and Lieberman, H. (2024). Can ai have common sense? finding out will be key to achieving machine intelligence. *Nature*, 634:291–294. DOI: 10.1038/d41586-024-03262-z.

Kıcıman, E., Ness, R., Sharma, A., and Tan, C. (2024). Causal reasoning and large language models: Opening a new frontier for causality. DOI: 10.48550/arxiv.2305.00050.

Lasheras, U., Alves, E., and Pinheiro, V. (2025). Interventional and counterfactual causal reasoning for llm-based ai agents a dataset and evaluation in portuguese. *Procesamiento del Lenguaje Natural*, 74. Available at: `http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6687`.

Lasheras, U. A. and Pinheiro, V. (2025). CaLQuest.PT: Towards the collection and evaluation of natural causal ladder questions in Portuguese for AI agents. In *LORESLM:2025:1*, pages 325–343, Abu Dhabi, United Arab Emirates. acl. Available at: `https://aclanthology.org/2025.loreslm-1.26/`.

Liu, X., Xu, P., Wu, J., Yuan, J., Yang, Y., Zhou, Y., Liu, F., Guan, T., Wang, H., Yu, T., McAuley, J. J., Ai, W., and Huang, F. (2024). Large language models and causal inference in collaboration: A comprehensive survey. *ArXiv*, abs/2403.09606. DOI: 10.18653/v1/2025.findings-naacl.427.

McClure, J., Hilton, D. J., Cowan, J., Ishida, L., and Wilson, M. (2001). When people explain difficult actions, is the causal question how or why? *Journal of Language and Social Psychology*, 20(3):339–357. DOI: 10.1177/0261927X01020003004.

Meta (2024). Introducing llama 3.1: Our most capable models to date. Available at:`https://ai.meta.com/blog/meta-llama-3-1/`.

Mostafazadeh, N., Kalyanpur, A., Moon, L., Buchanan, D., Berkowitz, L., Biran, O., and Chu-Carroll, J. (2020). GLUCOSE: GeneraLized and COntextualized story explanations. In *EMNLP:2020:main*, pages 4569–4586, Online. acl. DOI: 10.18653/v1/2020.emnlp-main.370.

OpenAI (2024). Hello gpt-4o. Available at:`https://openai.com/index/hello-gpt-4o/`.

OpenAI and et al., J. A. (2024). Gpt-4 technical report. DOI: 10.48550/arxiv.2303.08774.

Pearl, J. and Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. Basic Books, Inc., USA, 1st edition. DOI: 10.1090/noti1912.

Schank, R. C. (1995). The structure of episodes in memory. DOI: 10.1016/b978-0-12-108550-6.50014-8.

Tandon, N., Dalvi, B., Sakaguchi, K., Clark, P., and Bosselut, A. (2019). WIQA: A dataset for "what if..." reasoning over procedural text. In *EMNLP:2019:1*, pages 6076–6085, Hong Kong, China. acl. DOI: 10.18653/v1/D19-1629.

Wang, Z. (2024). CausalBench: A comprehensive benchmark for evaluating causal reasoning capabilities of large language models. In *SIGHAN:2024:1*, pages 143–151, Bangkok, Thailand. acl. Available at:`https://aclanthology.org/2024.sighan-1.17/`.

Zhang, L., Xu, H., Yang, Y., Zhou, S., You, W., Arora, M., and Callison-Burch, C. (2023). Causal reasoning of entities and events in procedural texts. In *FINDINGS:2023:eacl*, pages 415–431, Dubrovnik, Croatia. acl. DOI: 10.18653/v1/2023.findings-eacl.31.

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? Resposta: {'categoria': 'Causal', 'raciocinio': 'Essa pergunta é Causal (2) Dado um efeito, prever uma causa, pois busca identificar quais formas de ganhar dinheiro resultam no pior cenário possível. Para responder, é necessário avaliar diferentes métodos e julgá-los com base em fatores como risco, ética e esforço, estabelecendo uma relação de causa e efeito.'}

Pergunta: Por que casas de apostas como tigrinho ou blazer não são derrubadas? Resposta: {'categoria': 'Causal', 'raciocinio': 'Essa pergunta é Causal (2) Dado um efeito, prever uma causa, pois investiga o motivo pelo qual um fenômeno ocorre (casas de apostas não serem derrubadas). Para responder, é necessário analisar fatores como regulamentação, influência política, brechas legais e estratégias dessas empresas para continuar operando, estabelecendo uma relação de causa e efeito.'}

Pergunta: Tem alguma coisa que você faz por obrigação? Resposta: {'categoria': 'Não-Causal', 'raciocinio': 'Essa pergunta é Não-Causal, pois não implica uma relação de causa e efeito. Ela apenas busca uma informação sobre meu comportamento, sem exigir análise de causas ou previsões de efeitos.'}

Pergunta: Qual sua lembrança mais feliz? Resposta: {'categoria': 'Não-Causal', 'raciocinio': 'Essa pergunta é Não-Causal, pois busca apenas uma informação direta sobre uma memória específica, sem envolver relações de causa e efeito ou necessidade de raciocínio causal.'}

Pergunta: Qual é a temperatura adequada para um rack de pabx e de um rack de rede switch? Resposta: {'categoria': 'Causal', 'raciocinio': 'Essa pergunta é Causal (3) Dado um conjunto de variáveis, julgar a relação causal entre elas, pois busca entender a relação entre temperatura e o funcionamento adequado de equipamentos em racks de PABX e de rede. Para responder, é necessário analisar como a temperatura afeta esses dispositivos, identificando possíveis impactos como superaquecimento, falhas e desempenho reduzido, estabelecendo assim uma relação de causa e efeito.'}

Pergunta: Qual será a reação química se '(NH2)2CO' for adicionado a 'NaCl'? Resposta: {'categoria': 'Causal', 'raciocinio': 'Essa pergunta é Causal (1) Dado a causa, prever o efeito, pois solicita a previsão de uma reação química ao misturar dois compostos específicos. A pergunta busca entender o que acontece (efeito) quando uma causa (adicionar os dois compostos) é aplicada. Para responder, é necessário conhecer como esses compostos reagem entre si.'}

Pergunta: Em que países atualmente vigoram a monarquia eletiva? Resposta: {'categoria': 'Não-Causal', 'raciocinio': 'Essa pergunta é Não-Causal, pois está buscando uma informação objetiva sobre os países que adotam a monarquia eletiva, sem envolver uma análise de causa e efeito. É uma consulta direta sobre o sistema político vigente em determinados países.'}

Pergunta: Qual é a ultima versão do pytorch lançada? Resposta: {'categoria': 'Não-Causal', 'raciocinio': 'Essa pergunta é Não-Causal, pois solicita uma informação objetiva sobre a versão mais recente do PyTorch, sem envolver qualquer relação de causa e efeito. Ela busca apenas um dado específico.'}

Pergunta: Qual é a melhor forma de retirar o fundo de uma fotografia no photoshop? Resposta: {'categoria': 'Causal', 'raciocinio': 'Essa pergunta é Causal (2) Dado um efeito, prever uma causa, pois busca saber qual ação (causa) no Photoshop leva ao efeito desejado (remover o fundo de uma fotografia). Para responder, é necessário identificar os passos ou ferramentas mais eficazes para alcançar esse resultado.'}

Pergunta: Qual o trajeto de carro eu posso fazer entre São Paulo e Brasília? Resposta: {'categoria': 'Causal', 'raciocinio': 'Essa pergunta é Causal (2) Dado um efeito, prever uma causa, pois busca saber qual trajeto de carro (causa) pode ser tomado para alcançar o destino desejado (Brasília), partindo de São Paulo. Embora o efeito (chegar a Brasília) seja implícito, a pergunta está buscando identificar a melhor causa (roteiro ou caminho) para atingir esse objetivo.'}

Pergunta: Você consegue achar no nosso histórico de conversas por um assunto especifico? Resposta: {'categoria': 'Não-Causal', 'raciocinio': 'A pergunta é Não-Causal, pois busca apenas uma informação direta sobre a capacidade de realizar uma tarefa, sem envolver qualquer análise de causa e efeito.'}

Pergunta: O que é um disco de vinil? Resposta: {'categoria': 'Não-Causal', 'raciocinio': 'A pergunta é Não-Causal, pois busca apenas uma definição ou descrição objetiva sobre o que é um disco de vinil, sem envolver qualquer análise de causa e efeito.'}

Solicito que você classifique em uma das duas categorias acima detalhadas: Causal e Não-Causal. Não responda a pergunta apresentada!

Figure 5. 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. Não responda a pergunta apresentada! Faça a linha de raciocínio passo-a-passo.

Figure 6. 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? Resposta: {'categoria': 'Busca-Causa', 'raciocinio': 'Essa pergunta se encaixa na classe **Busca-Causa**. Ela busca entender a razão ou justificativa para a diferença de tratamento entre as formas de pagamento (Pix e boleto) e procura explicar por que as lojas oferecem descontos para uma forma e não para a outra.'}

Pergunta: Quais são os sinais de que um relacionamento é feliz e saudável na opinião de vocês? Resposta: {'categoria': 'Busca-Efeito', 'raciocinio': 'A pergunta é Busca-Efeito porque ela está procurando entender os sinais ou efeitos de um relacionamento saudável e feliz, ou seja, os resultados visíveis que indicam que o relacionamento é bom.'}

Pergunta: Existe uma idade mínima ou ideal para aprender sobre política e economia? Resposta: {'categoria': 'Busca-Relação', 'raciocinio': 'A pergunta é Busca-Relação porque está buscando entender se há uma relação entre a idade e a capacidade de aprender sobre política e economia.'}

Pergunta: Qual midia da mais liberdade criativa pro criador?livro,filme,serie ou quadrinho? Resposta: {'categoria': 'Busca-Recomendação', 'raciocinio': 'A pergunta busca uma recomendação sobre qual mídia oferece mais liberdade criativa para o criador, considerando um objetivo (maximizar a liberdade criativa). O indivíduo deseja saber qual opção (livro, filme, série ou quadrinho) é a melhor para atingir esse objetivo, o que caracteriza uma Busca-Recomendação.'}

Pergunta: Como mudo meu nome no reddit? Resposta: {'categoria': 'Busca-Passos', 'raciocinio': 'Essa pergunta se encaixa na classe Busca-Passos. Ela busca uma solução prática e direta para um problema, ou seja, um conjunto de etapas que devem ser seguidas para alterar o nome no Reddit.'}

Pergunta: Por que a Globo corta os filmes na programação? Resposta: {'categoria': 'Busca-Causa', 'raciocinio': 'Essa pergunta se encaixa na classe Busca-Causa. Ela busca entender a razão ou justificativa para a Globo cortar os filmes na programação, investigando o motivo dessa prática.'}

Pergunta: Como seria se cada arroz fosse feijão? Resposta: {'categoria': 'Busca-Efeito', 'raciocinio': 'Essa pergunta se encaixa na classe Busca-Efeito. Ela busca entender os efeitos ou consequências hipotéticas de uma mudança específica — no caso, o efeito de os grãos de arroz se tornarem feijão.'}

Pergunta: Tosse seca é sintoma de Dengue? Resposta: {'categoria': 'Busca-Relação', 'raciocinio': 'Essa pergunta se encaixa na classe Busca-Relação. Ela questiona se há uma relação causal entre a tosse seca e a dengue, ou seja, se a tosse seca pode ser um sintoma da doença.'}

Pergunta: Você pode fazer uma estratégia da Ferrari para voltar a vencer corridas na F1? Resposta: {'categoria': 'Busca-Recomendação', 'raciocinio': 'A questão busca uma recomendação sobre a melhor estratégia para a Ferrari voltar a vencer corridas na F1. O objetivo é encontrar a opção mais eficaz para alcançar esse resultado, ou seja, uma recomendação para maximizar as chances de vitória da equipe. Isso caracteriza uma Busca-Recomendação.'}

Pergunta: Me dê uma introdução para um manual de passo a passo de organização digital. Resposta: {'categoria': 'Busca-Passos', 'raciocinio': 'A pergunta se encaixa na classe Busca-Passos. Ela pede para criar uma introdução que sirva de guia em um manual de organização digital, o que implica um conjunto de passos para estruturar o conteúdo e organizar as informações de forma eficaz.'}

Pergunta: Como acontecem invasões em sites que foram criados usando WordPress? Resposta: {'categoria': 'Busca-Causa', 'raciocinio': 'Essa pergunta se encaixa na classe Busca-Causa. Ela busca entender as causas ou razões pelas quais invasões acontecem em sites criados com WordPress, investigando os fatores que tornam esses sites vulneráveis a ataques.'}

Pergunta: Eu tenho 1,75m e peso 117Kg, sou obeso? Resposta: {'categoria': 'Busca-Efeito', 'raciocinio': 'A pergunta está buscando entender os efeitos de uma pessoa ter 1,75m de altura e pesar 117kg, ou seja, o impacto dessa combinação de peso e altura no diagnóstico de obesidade. Ela busca prever ou entender as consequências dessa situação em termos de saúde ou classificação (IMC). Isso caracteriza uma Busca-Efeito, pois se está buscando entender os efeitos dessa condição específica.'}

Pergunta: Os refrigerantes com 50% de fruta são saudáveis? Resposta: {'categoria': 'Busca-Relação', 'raciocinio': 'Essa pergunta se encaixa na classe Busca-Relação. Ela busca entender se há uma relação causal entre a quantidade de fruta no refrigerante e sua saúde, questionando se o fato de conter 50% de fruta faz com que o refrigerante seja saudável.'}

Pergunta: Qual a melhor forma de usar o ChatGTP para criar conteúdo? Resposta: {'categoria': 'Busca-Recomendação', 'raciocinio': 'Essa pergunta se encaixa na classe Busca-Recomendação. Ela busca uma recomendação sobre a melhor forma de utilizar o ChatGPT para criar conteúdo, ou seja, qual abordagem ou estratégia seria mais eficaz para alcançar o objetivo de criar conteúdo.'}

Pergunta: Como implementar uma instancia de leitura e uma de escrita do banco de dados no Laravel 7.4? Resposta: {'categoria': 'Busca-Passos', 'raciocinio': 'Essa pergunta se encaixa na classe Busca-Passos. Ela busca um conjunto de etapas ou um guia para implementar duas instâncias de banco de dados (leitura e escrita) no Laravel 7.4, ou seja, uma solução prática com um passo a passo para realizar a implementação.'}

Solicito que você classifique em uma das cinco categorias acima detalhadas: Busca-Causa, Busca-Efeito, Busca-Relação, Busca-Recomendação e Busca-Passos; não responda a pergunta apresentada!

Figure 7. 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; Não responda a pergunta apresentada! Faça uma linha de raciocínio passo-a-passo.

Figure 8. 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:
Pergunta: Por que entrevista de emprego virou tortura? Resposta: {'categoria': 'Associacional', 'raciocinio': 'Essa pergunta se encaixa na categoria associacional, pois ela busca entender uma correlação ou relação entre a entrevista de emprego e a sensação de sofrimento ou pressão que as pessoas podem experienciar, sugerindo uma relação estatística entre esses dois eventos.'}
Pergunta: Consigo fazer mestrado me graduando em EAD? Resposta: {'categoria': 'Intervencional', 'raciocinio': 'Essa pergunta se encaixa na categoria intervencional, pois está buscando entender a possibilidade de uma intervenção (fazer o mestrado) a partir de uma situação específica (estar se graduando em EAD), questionando o efeito dessa intervenção sobre a viabilidade ou aceitação do mestrado.'}
Pergunta: Eu teria ótimas oportunidades de emprego com estes cursos no currículo + minha experiência? Resposta: {'categoria': 'Contrafactual', 'raciocinio': 'A pergunta é contrafactual porque o interlocutor está refletindo sobre o passado e se perguntando como teria sido sua situação de emprego caso tivesse feito cursos e adquirido experiência antes, ou seja, considerando um cenário alternativo do que poderia ter acontecido.'}
Pergunta: Essa nova geração é realmente pior que a passada? Resposta: {'categoria': 'Associacional',

'raciocinio': 'A pergunta é associacional porque busca entender a correlação entre as características da nova geração e da passada, investigando se existe uma relação de "melhor" ou "pior" entre elas.'}
Pergunta: Vocês acham que perderiam suas amizades se descobrissem tudo o que você pensa?
Resposta: {'categoria': 'Intervencional', 'raciocinio': 'Essa pergunta se encaixa na categoria intervencional, pois está considerando uma ação hipotética (revelar tudo o que pensa) e perguntando sobre o impacto que essa ação teria nas amizades. O foco é entender o efeito de uma intervenção no relacionamento social.'}
Pergunta: O que teria acontecido se nunca tivesse existido exploração no mundo? Resposta: {'categoria': 'Contrafactual', 'raciocinio': 'Essa pergunta é contrafactual, pois está explorando uma realidade alternativa no passado, questionando como o mundo teria sido diferente se a exploração nunca tivesse ocorrido, ou seja, o que teria acontecido em uma linha do tempo onde esse evento não existisse.'}

Solicito que você classifique em uma das três categorias acima detalhadas: Associacional, Intervencional ou Contrafactual. Não responda a pergunta apresentada!

Figure 9. 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").

…

Solicito que você classifique em uma das três categorias acima detalhadas: Associacional, Intervencional ou Contrafactual. Não responda a pergunda apresentada! Faça uma linha de raciocínio passo-a-passo.

Figure 10. 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**
Ser empreendedor compensa?	Is being an entrepreneur worth it?	Causal
Utilizar fontes digitais e compiladas no mestrado pega mal?	Is it bad to use digital and compiled sources in a master's degree?	Causal
Como arrumar tempo ?	How to make time?	Causal
Você tem medo de parecer ser o que não é?	Are you afraid of appearing to be something you are not?	Non-Causal
Qual seu desenho animado favorito?	What's your favorite cartoon?	Non-Causal
Você é uma pessoa estrategista?	Are you a strategic person?	Non-Causal

Table 14. Examples of Causal / Non-Causal Questions on the 2000 Questions Dataset, classified according to the Axis-1 of the taxonomy.

Class of Action		
Question(PT)	**Question(EN)**	**Class**
Porque o mundo é tão desigual financeiro?	Why is the world so financially unequal?	Cause-seek.
Por que temos tão pouco conteúdo em pt-br sobre FPGA/SoC?	Why do we have so little content in pt-br about FPGA/SoC?	Cause-seek.
Meu cachorro late pro nada desesperado o que pode ser?	My dog barks at nothing in despair, what could it be?	Cause-seek.
As empresas ainda contratam Júnior?	Do companies still hire Juniors?	Effect-Seek.
Você já teve consequências por mentir para faltar aula?	Have you ever had consequences for lying to skip class?	Effect-Seek.
O quanto saber alemão e francês ajuda na carreira acadêmica?	How much does knowing German and French help in your academic career?	Effect-Seek.
Certificados da Udemy têm valor para as empresas?	Are Udemy certificates valuable to companies?	Relation-Seek.
Vocês acham que canal de games deixam crianças mais bobas?	Do you think gaming channels make kids more silly?	Relation-Seek.
É possível uma pessoa amar trabalhar como suporte técnico?	Is it possible for a person to love working as a technical support worker?	Relation-Seek.
Meu gato me arranhou, devo me preocupar?	My cat scratched me, should I be worried?	Recomm.-Seek.
É possível aprender a gostar de algo, ou é melhor deixar isso de lado e buscar outra coisa?	Is it possible to learn to like something, or is it better to leave it aside and l ook for something else?	Recomm.-Seek.
Vale a pena seguir meu sonho e cursar o que sempre quis?	Is it worth following my dream and studying what I always wanted?	Recomm.-Seek.
Afinal como arranjar um emprego?	So how do you get a job?	Steps-Seek.
Como conseguir clientes para meus trabalhos?	How do I get clients for my work?	Steps-Seek.
Como iniciar uma conversa com alguém por mensagem?	How to start a conversation with someone via message?	Steps-Seek.

Table 15. Examples of 2000 Questions Dataset of the CaLQuest.PT, classified according to the Axis-2 of the taxonomy.

Pearl's Ladder of Causality		
Question(BR)	**Question(EN)**	**Class**
Como saber onde acontece esse eventos de tecnologia?	How do I know where these technology events take place?	Associat.
Essa ração para cachorro é boa?	Is this dog food good?	Associat.
Em quais casos preciso devolver a bolsa PIBIC?	In which cases do I need to return the PIBIC grant?	Associat.
Migrar para Europa e continuar em vaga confort-zone ou buscar algo mais desafiador?	Migrate to Europe and remain in a vague comfort zone or seek something more challenging?	Interven.
Para estágio, vale a pena focar em C#?	For an internship, is it worth focusing on C#?	Interven.
Sou biólogo, vale a pena ser bioinformata ?	I'm a biologist, is it worth being a bioinformatician?	Interven.
O que fariam se faculdade presencial não fosse possível?	What would you do if in-person college wasn't possible?	Counterf.
Especule como seria um universo . deletromagnético	Speculate what an electromagnetic universe would be like.	Counterf.
O que aconteceria se o Presidente Kennedy ainda estivesse vivo?	What would happen if President Kennedy was still alive?	Counterf.

Table 16. Examples of 2000 Questions Dataset of the CaLQuest.PT, classified according to the Axis-3 of the taxonomy.

E Examples of Causal Seed Questions

Você é um avaliador de um classificador de perguntas feitas por um humano nas classes <CLASSES>. A definição destas classes são as seguintes:

CLASSE 1: <DEFINIÇÃO DA CLASSE 1>…

CLASSE N <DEFINIÇÃO DA CLASSE N>

Você deve julgar se a pergunta foi classificada corretamente como<CLASSES>, marcando como CORRETA se classificado corretamente, e como INCORRETA se classificado incorretamente. Seguem alguns exemplos:

EXEMPLOS:

EXEMPLO 1:

Pergunta: <PERGUNTA 1>

Classificação: <CLASSIFICAÇÃO DA PERGUNTA 1>

Raciocínio: <RACIOCÍNIO DA PERGUNTA 1>

Resposta do Avaliador: <CORRETO ou INCORRETO>

...

Não responda à pergunta apresentada, apenas responda no campo "Resposta do Avaliador:" sua avaliação sobre a classificação como CORRETA ou INCORRETA, e apresente o raciocínio passo-a-passo para sua resposta, no seguinte formato:

Resposta do Avaliador: "<Resposta>"

Raciocínio da avaliação: "<Raciocínio da Resposta>"

Figure 11. Base prompt for Causal Judge on any Axis of Taxonomy