Causal_QA.PT: A Human–LLM Co-Curated Benchmark for Causal Question Answering in Portuguese Language

Lia Furtado[1], Cíntia Araripe[1], Jocelani Castilhos[1], Lucas Holanda[1], Vladia Pinheiro[1]

[1]Universidade de Fortaleza, Fortaleza, Brazil
Correspondence: liasucf@gmail.com

Abstract

We present Causal_QA.PT, a human–LLM co-curated benchmark for causal question answering in Portuguese, addressing the lack of high-quality evaluation resources for causal reasoning in non-English languages. The dataset is developed through a hybrid human–LLM process with targeted generation, validation, and evaluation procedures, and is organized according to the PEARL causal typology. Using this resource, we evaluate the ability of Large Language Models to answer causal questions in Portuguese and examine the role of explicitly providing causal class information in prompt design. Our findings show that current LLMs are capable of producing high-quality causal responses in Portuguese, with GPT-5 Mini in particular demonstrating strong performance in judgment-based evaluation. Explicit causal class information yields model- and question-dependent benefits, particularly for interventional and counterfactual questions. Finally, we observe that human reference answers are not always superior, underscoring the importance of careful benchmark curation and robust evaluation for underrepresented languages.

1 Introduction

Large Language Models (LLMs) have profoundly reshaped research in Natural Language Processing (NLP) and Artificial Intelligence (AI), shifting the field's focus from traditional language understanding and generation tasks to deeper questions involving capabilities such as causal and temporal reasoning (Zhang et al., 2023) (Zhang et al., 2020). Recently, Gartner's 2025 Hype Cycle for Artificial Intelligence (Truong, 2025) highlighted Causal AI as an emerging innovation area, emphasizing both its growing relevance and the many open research challenges it presents. Among these challenges, a central question has gained particular attention: when an LLM answers a causal question, is it genuinely performing structured reasoning about causes and effects, or is it merely acting as a "causal parrot" (Zečević et al., 2023), that is, reproducing linguistic patterns present in its training data rather than engaging in authentic causal inference?

A common point of agreement in the literature is that it is far from trivial to determine whether an LLM actually employed causal reasoning when answering a causal question. The big question is *"Do LLMs understand causality?"*. Consider, for instance, an interventional question such as "What would happen to the unemployment rate if the government increased the minimum wage?" A human answer typically grounds the explanation in a causal mechanism, e.g., "If the minimum wage increases, firms facing higher labor costs may reduce hiring, which could increase unemployment depending on market elasticity." In contrast, an LLM might respond with a surface-level generalization such as "Raising the minimum wage usually leads to higher unemployment," offering no explicit causal pathway or conditions under which the effect may reverse. In fact, the core challenge is evaluating whether it performed any structured causal inference or simply reproduced familiar linguistic associations.

To address partially this problem, several datasets have been proposed in English, such as the CLADDER (Jin et al., 2023), CAUSALQA (Bondarenko et al., 2022) and CAUSALQUEST (Ceraolo et al., 2024). For the Portuguese language, the first dataset with causal natural questions - CaLQuest.PT - was introduced by (Lasheras and Pinheiro, 2025) and subsequently extended in multiple follow-up works (Lasheras et al., 2025a) (Lasheras et al., 2025b). Nevertheless, this dataset contains only questions classified according to a proposed causality taxonomy and does not provide answers to support causal question answering (QA) analysis.

As a result, no dataset currently exists in Por-

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 657–666

tuguese that serves as a comprehensive benchmark for evaluating causal reasoning and causal question answering in LLMs, particularly with respect to the three rungs of causality—associational, interventional, and counterfactual (Pearl and Mackenzie, 2018) (Bareinboim et al., 2022).This gap limits the systematic assessment of models capabilities and constrains progress in research on explainability and causal inference in AI systems. To address this need, we introduce Causal_QA.PT, a linguistic resource consisting of causal questions in Portuguese paired with human-generated and human-curated answers. Our goal is to provide a reliable benchmark for investigating, comparing, and improving causal reasoning agents in Brazilian Portuguese, thereby supporting the development of models better aligned with the linguistic, scientific, and sociotechnical contexts of Brazil.

Several research questions guided the development of this work and shaped its main contributions:

- **RQ1**: How can a systematic methodology be developed to build and refine a high-quality causal question answering benchmark in Portuguese?
- **RQ2**:To what extent do LLMs generate accurate and high-quality answers to causal questions in Portuguese?
- **RQ3**: How does explicitly providing causal class information in the prompt affect model performance across different types of causal questions?

Following the proposed methodology, we developed the first version of the Causal_QA.PT benchmark, employing both open-source and proprietary LLMs to generate and analyze the answer set. The quality of these answers was assessed using reference-based and reference-free metrics, enabling a comprehensive evaluation of LLM's performance. The results indicate that LLMs can produce high-quality causal answers in Portuguese, with GPT-5 Mini delivering the most consistent performance. Providing explicit causal class information improves answer quality in some cases, but the effect depends on both the model and the type of question: GPT-5 Mini benefits noticeably, especially for interventional and counterfactual questions, while Gemma-27B shows limited improvement. The results also indicate that human reference answers are not always superior, highlighting the need for careful benchmark curation and robust evaluation methods, particularly for underrepresented languages such as Portuguese.

2 Related Works

The CLADDER dataset (Jin et al., 2023) was proposed as a means to rigorously evaluate formal causal reasoning in Large Language Models (LLMs). Its key innovation lies in grounding each causal question expressed in English natural language in a corresponding symbolic formulation whose ground-truth answer is computed by an oracle causal inference engine. This engine adheres strictly to Pearl's causal inference framework—based on graphical models and structural causal models (SCMs), thereby ensuring that all answers reflect formally valid causal reasoning rather than human annotation or surface linguistic patterns. However the questions of this dataset were artificially generated

CAUSALQA (Bondarenko et al., 2022) is the first large-scale benchmark for causal question answering in English, comprising approximately 1.1 million causal question–answer pairs extracted from ten widely used QA datasets. The authors propose a novel, theory-grounded typology of causal questions along semantic (antecedent, consequent, and causal chain) and pragmatic (intent-based) dimensions, and use high-precision lexical rules to identify causal questions at scale. Their analysis shows that around 5% of real web search queries are causal, underscoring the practical relevance of the task.

CAUSALQUEST (Ceraolo et al., 2024) is a dataset of 13,500 naturally occurring questions sourced from social networks, search engines, and AI assistants, in English language. The authors formalized the definition of causal questions and establish a taxonomy for finer-grained classification. Through a combined effort of human annotators and large language models (LLMs), CAUSALQUEST was carefully label. This work found that 42% of the questions humans ask are indeed causal, with the majority seeking to understand the causes behind given effects. Using this dataset, efficient classifiers (up to 2.85B parameters) for the binary task of identifying causal questions, achieving high performance with F1 scores of up to 0.877.

In Portuguese language, CaLQuest.PT (Lasheras and Pinheiro, 2025) was the first corpus of 8,000

natural questions in Portuguese, designed to support causal evaluation at multiple levels. The dataset allows for three classification layers: (1) causal versus non-causal questions, (2) types of causal action, such as cause-seeking, effect-seeking, and recommendation-seeking, and (3) types of reasoning based on Pearl's Ladder of Causality—associative, interventionist, and counterfactual. An improved few-shot learning strategy is also presented, and the performance of open-source models fitted to this corpus is evaluated (Lasheras et al., 2025b). However, this dataset contains only questions classified according to a proposed causality taxonomy but does not provide answers to support causal question answering (QA) analysis.

3 Causal_QA.PT: A Benchmark for Causal Question Answering in Portuguese Language

The proposed methodology for building and systematically generating, curating, evaluating, and refining the Causal_QA.PT dataset and benchmark is illustrated in Figure 1.

3.1 Data Sources

The initial dataset was the CaLQuest.PT (Lasheras and Pinheiro, 2025), a dataset with $2,585$ natural questions collected from four sources: *Reddit* (881), *ShareGPT* (814), *WildChat* (890), containing causal and non-causal questions. The causal questions are annotated with one of the Pearl's classes, indicating the causal reasoning: associational, counterfactual, and interventional.

From this initial dataset, we extracted all causal questions (1,196 in total) and removed duplicate entries, resulting in a final set of 915 unique items. To standardize question length, we applied a lower bound by removing questions with fewer than four words and defined an upper bound using the interquartile range (IQR). After applying these filters, a total of 810 questions remained in the dataset. Table 1 contains representative examples from the dataset.

3.2 Clustering and Topic Selection

To identify coherent thematic groups and support dataset organization and analysis, we constructed a topic-modeling pipeline based on BERTopic (Grootendorst, 2022). Texts were embedded, reduced in dimensionality using UMAP, and clustered with HDBSCAN. BERTopic then assigned interpretable topics to each cluster via its class-based TF-IDF representation, guided by a set of seed words to improve interpretability. This process yielded seven clusters and one residual topic. Based on the most representative terms, we manually assigned semantic labels to each cluster to enhance explainability. Table 2 summarizes the resulting categories, their top-3 words, and the number of questions per cluster.

Subsequently, questions can be selected from one or more specific topics. In this first version, we chose the largest and most representative topic identified by our topic-modeling analysis—Education and Career—as the starting point for annotation, comprising a total of 395 natural causal questions.

3.3 Human-Question Answering of Natural Questions

From the 395 selected natural causal questions of the topic "Education and Career", human annotators were able to answer 254 of these questions, because some proved too difficult to resolve manually.

The response generation process involved five human annotators with expertise in computer science, engineering, law, and medicine. Each annotator answered, on average, 50 questions, which were assigned, whenever possible, according to their areas of expertise. Three annotators are doctoral students, contributing academic expertise to the task. Annotators manually produced complete and contextually appropriate responses, avoiding overly simplified formulations. To guide the structure and level of detail, they consulted responses previously generated by Artificial Intelligence systems, although the final answers were written independently. After the initial drafting phase, inconsistencies or ambiguities were discussed among the participants until consensus was reached, resulting in a validated set of human-authored responses used in the study.

The distribution of these questions across Pearl Classes was imbalanced, particularly for the counterfactual class: 134 associational, 105 interventional and 15 counterfactual.

3.4 LLM-Generation of Question–Answer Pairs

To address the imbalance across Pearl's causal hierarchy, particularly the underrepresented counterfactual questions, our methodology includes a ded-

Type	Question	Answer
Associational	Faz sentido clean architecture em frameworks como Rails e Laravel? *Does clean architecture make sense in frameworks such as Rails and Laravel?*	Sim, faz sentido, pois melhora a manutenibilidade, legibilidade e colaboração do código. *Yes, it does, as it improves code maintainability, readability, and collaboration.*
Counterfactual	Se você pudesse descobrir a cura de uma doença, qual você escolheria? *If you could discover the cure for a disease, which one would you choose?*	Escolheria a cura para o câncer, pois é uma doença que afeta milhões de pessoas em todo o mundo e causa sofrimento e perda para muitas famílias. *I would choose the cure for cancer, as it is a disease that affects millions of people worldwide and causes suffering and loss for many families.*
Interventional	Você aceitaria um emprego de fim de semana por 20 mil dólares por ano para ser um espantalho humano e espantar pássaros dos campos? *Would you accept a weekend job for 20 thousand dollars a year to act as a human scarecrow and scare birds away from fields?*	Sim, considerando o valor oferecido, parece um trabalho relativamente simples e bem remunerado. *Yes, considering the amount offered, it seems like a relatively simple and well-paid job.*

Table 1: Illustrative questions and answers organized by causal type.

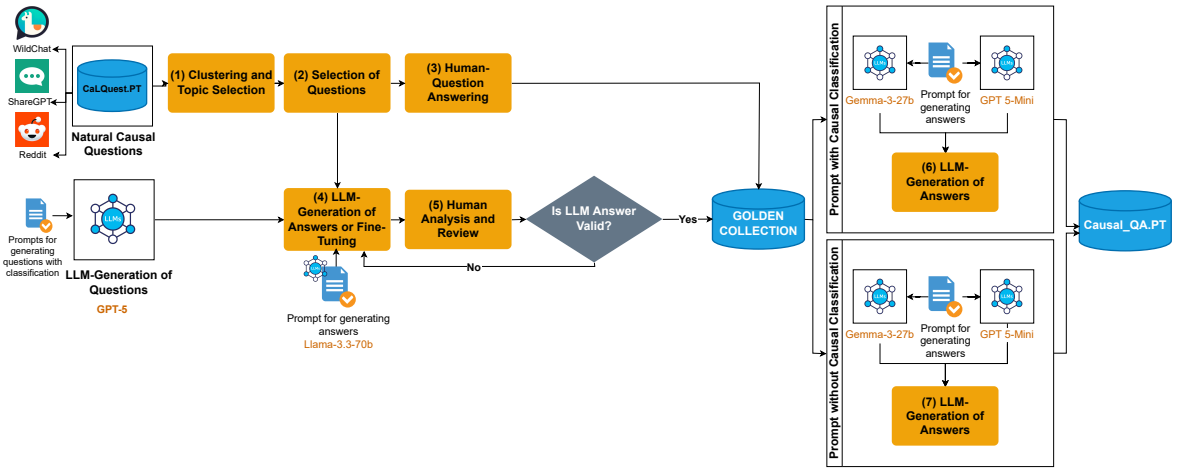

Figure 1: Methodology pipeline for constructing Causal_QA.PT, a human–LLM co-curated benchmark for causal question answering in Portuguese.

ID	Category	Top Words	Count
0	Education/Career	university, job, study	395
1	Pets/Animal Care	pet food, cat, dog	175
2	Technology	python, code, website	98
3	Motorsport	F1, Honda, McLaren	36
4	Business	strategy, finance, sell	33
5	Family/Language	Brazil, children, English	30
6	Science/Nature	universe, biology, science	23
-1	Residual Topic	–	20
Total			**810**

Table 2: Summary of clusters, their category labels, top representative words, and size.

icated dataset expansion stage in which additional items are generated using an LLM.

To ensure a controlled expansion of the dataset, we continued generating additional questions exclusively within the selected topic, e.g. the "Education and Career" topic. For each causal class, we provided explicit prompts defining the causal types and instructed the model to produce a fixed number of questions for each class while remaining within this topic domain. In this work, using these targeted prompts, a more robust LLM (e.g. GPT-5) generated a total of 148 new questions (119 counterfactual and 29 interventional), which were then answered by a smaller LLM (e.g. Llama 70B) to complete the corresponding entries in the golden collection. See the proposed prompts in (PROMPTS, 2025).

All question–answer pairs were subsequently reviewed by human annotators to identify potential

errors, inconsistencies, or inadequacies. When a pair was deemed valid, it was incorporated into the curated gold collection. Otherwise, the question was resubmitted to the LLM to generate a revised answer, after which the new question–answer pair underwent another round of human validation. This iterative human-in-the-loop process continued until the pair satisfied the established quality criteria.

Our final dataset combines two sources: (1) natural questions answered by humans, and (2) artificially LLM-generated questions answered by an LLM with human validation. In this work, the final golden collection includes 402 causal questions, where 254 are human-answered natural questions and 148 LLM-answered artificial questions. It is important highlight that all questions were reviewed by a human annotator. Table 3 presents the distribution of the collection across the three PEARL classes.

Table 3: Final distribution of Causal_QA.PT dataset across PEARL classes and Human-Answered x LLM-Answered questions.

Class	Count	Human-Answered	LLM-Answered
Associational	134	134	0
Interventional	134	105	29
Counterfactual	134	15	119
Total	**402**	**254**	**148**

3.5 Class-agnostic vs. class-informed strategies

One of our research questions (RQ3) aims to determine how explicitly providing the Pearl causal class of a question affects the quality of the answers produced by different models. In other words, this research question investigates whether explicitly revealing if a question is associational, interventional or counterfactual in the prompt sent to an LLM alters the quality of the generated responses. To enable this analysis in future experiments, the final version of the Causal_QA.PT benchmark includes answers generated under two distinct prompting conditions: (i) a class-agnostic setting, in which no causal classification is provided, and (ii) a class-informed setting, in which the causal class of the question is explicitly included in the prompt. Table 4 presents example responses for both settings.

The examples were translated into English; however, they were placed in the supplementary materials (PROMPTS, 2025).

Across all conditions, the model is guided by the same high-level prompts (PROMPTS, 2025) to produce answers that are (i) coherent, maintaining internal logical structure and appropriate reasoning; (ii) fluent, expressed clearly in natural language; consistent, avoiding contradictions or shifts in perspective; and (iii) relevant, addressing the question directly without unnecessary digressions. These criteria were defined in the prompts as the expected characteristics of a high-quality answer.

The distinction between conditions lies solely in whether the question's class is disclosed. In the class-agnostic setting, the model receives only the general instruction and the question itself, with no reference to the underlying class. In the class-informed setting, the prompt is augmented with a brief indication of the class, providing a short, high-level description of the type of reasoning the question belongs to. As noted earlier, the causal classification corresponds to the type of causal reasoning required by the question, following Pearl's causal hierarchy: associational, interventional, and counterfactual.

In the final two stages of our methodology, we employ two LLMs for each response condition: a smaller open-source model, Gemma-3-27B, and a larger proprietary model, GPT-5-Mini. Together, the four responses produced by the two models across the two prompting conditions constitute our proposed benchmark for causal question answering in Portuguese - Causal_QA.PT.

4 Causal_QA.PT Evaluation

4.1 Evaluation Methodology

To evaluate the proposed Causal_QA.PT benchmark, we designed a three-stage evaluation pipeline (Figure 2). First, in a reference-based evaluation, model-generated responses were compared against gold answers, either human-written or LLaMA-70B–generated and human-curated. Second, following the LLM-as-a-Judge paradigm, each response was individually scored for adherence, coherence, fluency, and relevance. Finally, in a comparative ranking stage, an LLM ranked all responses to the same question from best to worst, capturing relative performance differences.

This multi-stage evaluation provides both absolute and relative perspectives on model behavior and establishes the foundation for analyzing the effects of the class-informed and class-agnostic prompting strategies, which we presented in Section 3.5.

Type	Questions	Class-agnostic answers	Class-informed answers
Associational	Why are so many people in graduate school seemingly uncommitted?	Because several factors combine: many graduate students prioritize career and research, have intense workloads and commuting demands, face financial uncertainty and mobility issues (moving between cities or countries), and are also at an age when people may postpone relationships or value independence; the academic environment also reduces opportunities and social rhythms conducive to dating.	Because many graduate students have profiles and circumstances associated with lower academic commitment: they are professionals balancing work and study, enroll for credentials more than vocation, attend flexible programs (in-person, hybrid, or online), face overload/burnout or financial difficulties, and there is also a selection effect involving people with diverse goals (short-term objectives, networking, reskilling). These factors increase the likelihood of behavior perceived as "uncommitted".
Counterfactual	If you could discover the cure for one disease, which would you choose?	Alzheimer's disease.	I would choose to cure Alzheimer's disease. If there were a cure, millions of people would retain their autonomy and memory, the burden on caregivers and healthcare systems would drop drastically, and there would be a major reduction in social costs and suffering; if instead I had cured cancers or malaria, more immediate lives might be saved in some age groups, but the psychological, economic, and quality-of-life impact associated with Alzheimer's would make this cure especially transformative socially.
Interventional	Two reais or a mysterious startup?	A mysterious startup.	If you need immediate money or avoid risk: take the two reais. If you can wait, tolerate risk, and intend to seek information/control: prefer the startup — but only after due diligence or negotiating clear terms (equity, contracts, rights, deadlines), or by combining a small cash payment with an option on the startup.

Table 4: Illustrative questions and answers organized by causal type and prompt strategy.

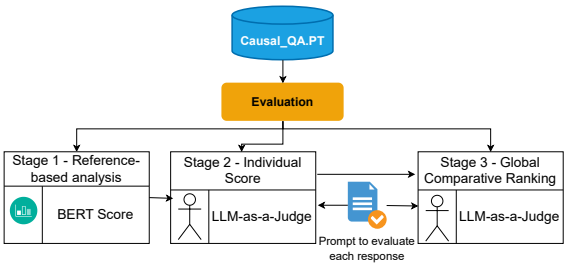

Figure 2: Evaluation Methodology of the Causal_QA.PT.

4.1.1 Reference based metrics

In the first evaluation stage, we assess response quality by comparing model outputs to a gold reference from our curated collection. Each question is paired with a single reference answer, either human-written or generated by LLaMA-70B and human-evaluated, allowing us to measure how closely responses produced under different prompting conditions approximate an accepted solution.

We quantify similarity using a reference-based BERTScore computed with the F1 variant, which measures semantic alignment through contextualized token embeddings rather than surface overlap. As all texts are in Portuguese, we use a pretrained Portuguese model to compute embeddings, yielding a continuous score that reflects semantic and expressive similarity between each response and its reference.

4.1.2 Individual Quality Scoring

In individual evaluation, we adopt the LLM-as-a-Judge paradigm (Chang et al., 2024), in which each response is examined independently, without direct comparison to alternatives. The goal is to measure the intrinsic quality of the answer produced under each experimental condition.

Responses are assessed along four dimensions: adherence to the central question, which evaluates whether the answer correctly addresses the problem posed; fluency and linguistic correctness, capturing clarity, sentence structure, and absence of grammatical errors; coherence and logical progression of ideas, assessing whether the answer maintains a consistent line of reasoning; and relevance and objectivity, which measures the extent to which the

answer avoids unnecessary digressions and focuses on essential information.

Each dimension is rated on a 1–5 Likert scale, ranging from very inadequate to excellent. This procedure provides a consistent quantification of response quality across both prompting settings while preserving the independence of each judgment.

4.1.3 Comparative Ranking

To complement the individual analysis, we conduct a comparative evaluation in which an independent judging agent (GPT-5) directly compares the three responses associated with each question: the reference answer, the class-agnostic response, and the class-informed response. This procedure allows us both to assess the quality of the reference dataset and to quantify the influence of class information on model behavior.

The three responses are presented to the judge in random order to mitigate position bias. The model is then required to produce a full ranking, from 1st (best) to 3rd (worst).

Even in cases where two answers obtain similar scores in the individual evaluation, the judge is instructed to break ties using fine-grained qualitative criteria such as higher objectivity, clarity, factual precision, and tighter adherence to the question's focus. This ensures a unique and consistent ranking for every set of responses.

4.2 Evaluation Results

4.2.1 Reference-Based Performance and the Reliability of Gold Answers

To understand how closely the two model responses (one from a smaller open-source model, Gemma-3-27B, and the other from a larger proprietary model, GPT-5-Mini) align with the expected answers, we first analyze performance in a reference-based evaluation setting.

Figure 3 summarizes the effect of providing class-specific information on the semantic similarity of the model outputs to the reference answers, as measured by BERTScore. The performance of the model was generally low according to BERTScore. GPT-5 Mini achieved mean scores below 0.5 for interventional and associational questions, while performance on counterfactual questions was slightly higher, approaching 0.6. When comparing class-informed and class-agnostic prompting strategies, we observed no substantial differences between them. The two prompting conditions produced highly similar results, suggesting that explicitly providing the causal class does not yield consistent gains in this reference-based setting. Across all question types and both prompting strategies, Gemma-27B achieves higher overall BERTScores than GPT-5-Mini, indicating closer alignment with the reference answers under this metric and highlighting the competitiveness of smaller open-source models.

To better understand these outcomes, we analyze the quality of the reference answers themselves. We independently evaluated the quality of the reference answers themselves using our Likert-based scoring procedure (bottom panel of Figure 4).

Human-written reference answers received an average score around 2 ("inadequate/confusing"), whereas LLaMA-70B reference answers scored between 3 ("acceptable/partially correct") and 4 ("good, with minor issues"). This indicates a potential misalignment between reference-based metrics and perceived answer quality: if human references are weaker, similarity-based evaluation may underestimate the true performance of the evaluated models. These findings highlight the limitations of relying solely on reference-based evaluation and motivate the need for complementary LLM-judgment-based assessments.

4.2.2 Individual Quality Evaluation Across Prompting Strategies

We next evaluated whether the effects observed in the reference-based setting were reflected in a direct assessment of response quality. Using individual Likert-scale scoring, GPT 5.1, acting as a judge and following a prompt-based evaluation strategy (PROMPTS, 2025), rated each model response independently against predefined quality criteria. Figure 5 summarizes the performance of the prompting strategies across models and PEARL causal classes. Consistent with the BERTScore analysis, we observe that for Gemma 27B there is virtually no difference between the class-informed and class-agnostic prompting strategies for any causal class. In contrast, GPT-5 Mini shows a small improvement when the causal class is provided in the prompt for Interventional and Contrafactual questions, although the magnitude of this gain remains limited.

Moreover, the overall Likert scores reveal that responses generated by GPT-5 Mini received higher mean quality ratings (4.50) compared to Gemma 27B (3.66), indicating that the GPT model gen-

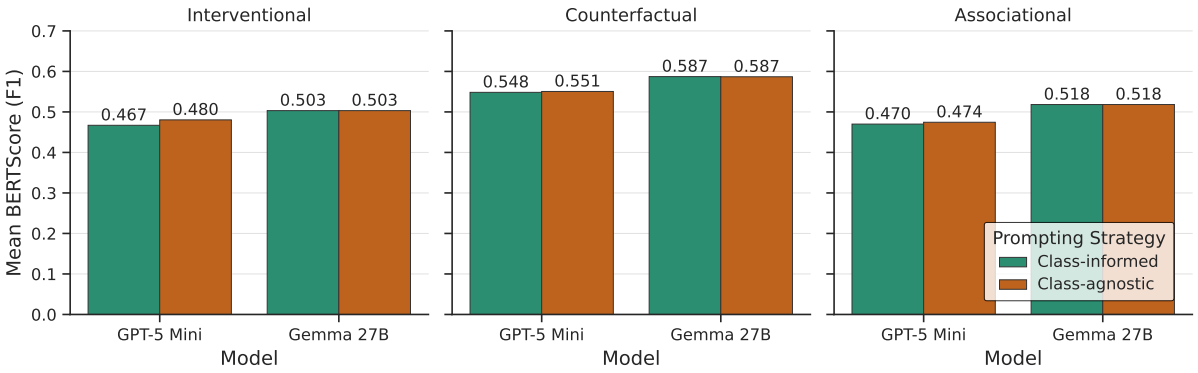

Figure 3: Impact of class-specific prompting on BERTScore across question types. Mean BERTScore is reported for *Class-agnostic* and *Class-informed* prompts within each PEARL causal category (associational, interventional, and counterfactual).

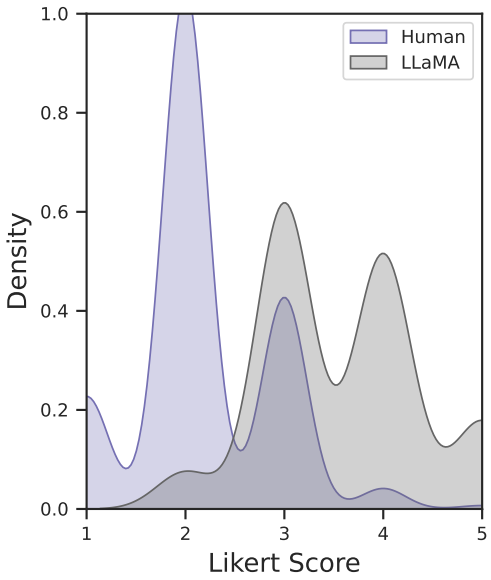

Figure 4: Kernel density estimation (KDE) of Likert quality scores comparing reference answers written by humans and by the LLaMA model. The plot shows the distribution of perceived answer quality within the golden reference collection, independently of the evaluated model.

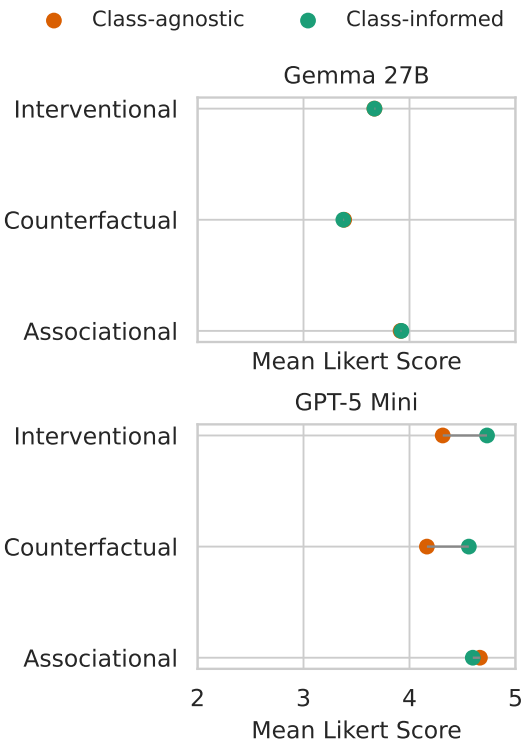

Figure 5: Comparison of mean Likert scores across PEARL classes for both prompting conditions (class-informed vs. class-agnostic) and models.

erally produced higher-quality answers under this evaluation setting.

Taken together, these findings indicate that the limited impact of class-informed prompting is not merely a consequence of reference-based similarity metrics, but is also reflected in independent judgment-based quality evaluations.

4.2.3 Comparative Based Evaluation of Prompting Strategies

To better understand which responses are genuinely preferred in a direct comparative setting, we conducted a ranking-based evaluation in which the LLM GPT 5.1, acting as a judge and, following a prompt-based evaluation strategy (PROMPTS, 2025), compared the three candidate answers for each question: the class-informed output, the class-agnostic output, and the reference answer. Unlike the individual quality scoring analysis, this setup forces the LLM-evaluator to make relative judgments, explicitly selecting the best, second-best, and worst response for each case. This allows us to assess not only whether prompting strategies (class-agnostic and class-informed) improve qual-

ity in absolute terms, but also how they perform when directly competing with one another and with the golden reference set.

Figure 6 reports the ranking-based evaluation comparing the responses generated under each prompting strategy (class-informed or class-agnostic) for each model and the reference answers. For Gemma 27B answers, the distribution of rankings shows no consistently dominant source of answers. Ground truth references, class-informed outputs, and class-agnostic outputs each achieved first place in approximately 30% of the cases. However, an important observation is that in the majority of remaining instances, the reference answers were ranked as the worst response, frequently being placed last when compared to both prompting strategies. This highlights potential concerns about relying exclusively on the golden reference set when assessing performance.

In contrast, GPT-5 Mini answers displays a clearer and more stable pattern. More than 60% of the time, the class-informed responses were ranked as the best answer, while the class-agnostic responses were ranked second in approximately 57% of the cases. Ground truth answers rarely appeared as the top-ranked response for GPT-5 Mini, further reinforcing the gap between model performance and reference quality.

Taken together, these results indicate that while Gemma 27B does not strongly benefit from class-conditioning in a comparative setting, GPT-5 Mini shows clear advantages when the causal class is explicitly provided. Moreover, the frequent poor ranking of human or LLaMA reference answers suggests that the golden set may not always reflect the strongest available responses, which has important implications for reference-based evaluation frameworks.

5 Conclusion

This work introduced a systematic methodology for building Causal_QA.PT, a benchmark for causal question answering in Portuguese. The pipeline combines human-authored questions, controlled synthetic augmentation, structured validation, and multi-stage evaluation, addressing the lack of reliable evaluation resources for Portuguese NLP in this domain (RQ1).

Regarding response quality (RQ2), the results indicate that Large Language Models can produce high-quality causal answers in Portuguese. In the individual Likert evaluation, GPT-5 Mini responses were frequently rated between "Good" and "Excellent", suggesting it can generate accurate, contextually appropriate, and well-structured answers.

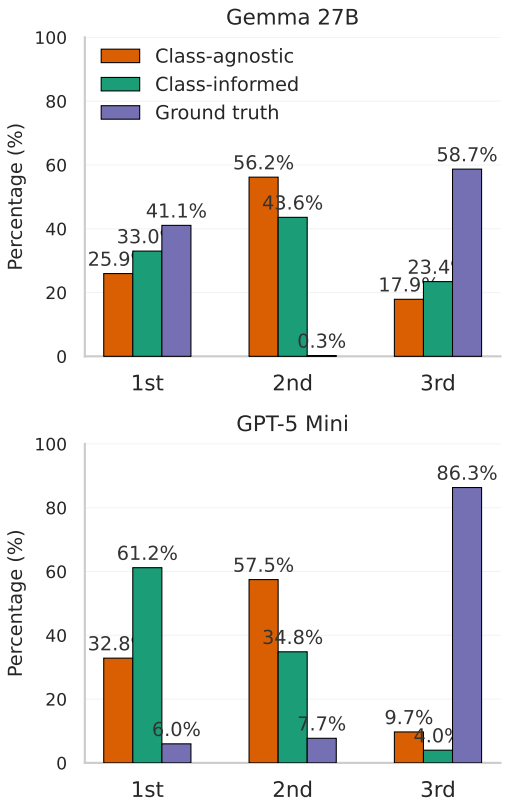

Figure 6: Comparative ranking results for Gemma-27B and GPT-5 Mini. Bars show the percentage of times responses were ranked in *1st*, *2nd*, or *3rd* place during LLM-based comparative evaluation. Rankings are reported for *class-agnostic*, *class-informed*, and *ground-truth* (reference) responses.

We also investigated whether providing causal class information improves model performance (RQ3). Results showed limited and model-dependent effects: while Gemma 27B showed little improvement with class-informed prompting, GPT-5 Mini benefited from it, particularly for interventional and counterfactual questions.

Finally, reference answers were often ranked as the worst option, highlighting the need for stronger gold standards. Future work should improve the reference dataset with richer human-written answers, expand the benchmark to additional domains, and increase coverage across PEARL causal classes.

References

Elias Bareinboim, Juan D. Correa, Duligur Ibeling, and Thomas Icard. 2022. *On Pearl's Hierarchy and the Foundations of Causal Inference*, 1 edition, page

507–556. Association for Computing Machinery, New York, NY, USA.

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

Roberto Ceraolo, Dmitrii Kharlapenko, Amélie Reymond, Rada Mihalcea, Bernhard Schölkopf, Mrinmaya Sachan, and Zhijing Jin. 2024. Causalquest: Collecting natural causal questions for AI agents. In *Causality and Large Models @NeurIPS 2024*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3).

Maarten R. Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *ArXiv*, abs/2203.05794.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. 2023. Cladder: assessing causal reasoning in language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Uriel Lasheras, Elioenai Alves, and Vladia Pinheiro. 2025a. Interventional and counterfactual causal reasoning for llm-based ai agents: A dataset and evaluation in portuguese. *Procesamiento del Lenguaje Natural*, 74(0):363–382.

Uriel Lasheras, Elioenai Alves, Caio Ponte, Carlos Caminha, and Vládia Pinheiro. 2025b. Open llms meet causality in portuguese: A corpus-based fine-tuning approach. *Journal of the Brazilian Computer Society*, 31(1):1004–1029.

Uriel Anderson Lasheras and Vladia Pinheiro. 2025. CaLQuest.PT: Towards the collection and evaluation of natural causal ladder questions in Portuguese for AI agents. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 325–343, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

J. Pearl and D. Mackenzie. 2018. *The Book of Why: The New Science of Cause and Effect*. Penguin Books Limited.

PROMPTS. 2025. Causal_qa.pt repository. `https://l1nq.com/casual-QA-PT`. Anonymous for blind review.

Vinh Truong. 2025. Hype and adoption of generative artificial intelligence applications. *arXiv preprint arXiv:2504.18081*.

Matej Zečević, Moritz Willig, {Devendra Singh} Dhami, and Kristian Kersting. 2023. Causal parrots: Large language models may talk causality but are not causal. *Transactions on Machine Learning Research*, 2023(8).

Cheng Zhang, Stefan Bauer, Paul Bennett, Jiangfeng Gao, Wenbo Gong, Agrin Hilmkil, Joel Jennings, Chao Ma, Tom Minka, Nick Pawlowski, and James Vaughan. 2023. Understanding causality with large language models: Feasibility and opportunities. *arXiv preprint arXiv:2304.05524*.

Li Zhang, Qing Lyu, and Chris Callison-Burch. 2020. Reasoning about goals, steps, and temporal ordering with WikiHow. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4630–4639, Online. Association for Computational Linguistics.