ExpliCa: Evaluating Explicit Causal Reasoning in Large Language Models

Anonymous ACL submission

Abstract

Large Language Models (LLMs) are increasingly used in tasks requiring interpretive and inferential accuracy. In this paper, we introduce ExpliCa, a new dataset for evaluating LLMs in explicit causal reasoning. ExpliCa uniquely integrates both causal and temporal relations presented in different linguistic orders and explicitly expressed by linguistic connectives. The dataset is enriched with crowdsourced human acceptability ratings. We tested LLMs on ExpliCa through prompting and perplexity-based metrics. We assessed seven commercial and open-source LLMs, revealing that even top models struggle to reach 0.80 accuracy. Interestingly, models tend to confound temporal relations with causal ones, and their performance is also strongly influenced by the linguistic order of the events. Finally, perplexity-based scores and prompting performance are differently affected by model size.

1 Introduction

Understanding cause-effect relationships is one of the hallmarks of human cognition (Pearl, 2009). The question of whether Large Language Models (LLMs) truly comprehend causal relationships in natural language texts, or merely perform as 'stochastic parrots' (Bender et al., 2021) by replicating statistical associations in their pretraining data, remains a topic of debate (Zečević et al., 2023; Merrill et al., 2024). This question is crucial for the application of LLMs in domains that demand interpretive and inferential accuracy. The recent growth in causal research and benchmarking highlights the fundamental need for more reliable and interpretable models (Chen et al., 2024). LLMs should be able to interpret not only cause-effect relations but also the relations between causal and temporal aspects (Ning et al. (2018) as an example), which are often connected and overlapping (e.g., typically an effect temporally follows its cause).

A common task for evaluating causal reasoning skills is Pairwise Causal Discovery (PCD), which focuses on determining the existence of a causal relation between two events and, if so, establishing which event serves as the cause and which one as the effect (Gao et al., 2023; Wan et al., 2024). However, this formulation does not directly take into account the tight bound between the *cause-effect* and the *before-after* relationships.

In this paper, we present **ExpliCa**, a dataset designed to evaluate LLMs in commonsense causal reasoning through PCD tasks. Our formulation of the PCD task for ExpliCa allows us to take into account also the entanglement of causal and temporal relations between events. This is achieved by considering sentence pairs and connective words that overtly express these relationships. In ExpliCa, each sentence describes an event (e.g. *Martina has less chance of getting the flu* and *Martina has been vaccinated*), and each connective word **explicitly** indicates either a causal relationship (i.e., *so* and *because*) or a temporal one (i.e., *before* and *after*). We collected acceptability ratings for each connective word with respect to each sentence pair, to account for both causal and temporal relations. To our knowledge, ExpliCa is the first dataset containing both causal and temporal explicit relations between events annotated by native speakers via crowdsourcing, rather than expert annotators.

We conducted a nuanced evaluation of a group of LLMs on ExpliCa. Its goal is to shed light on several key aspects. First, we aim to estimate whether and to what extent LLMs can model and distinguish causal and temporal relations, similarly to humans. Second, we want to assess potential differences between LLMs' *competence* and *performance* in our PCD task. Recent studies (Hu and Levy, 2023; Kauf et al., 2024) identified a discrepancy between the linguistic competence of models, as measured by log-probability, and the accuracy of their responses elicited via prompting, with the

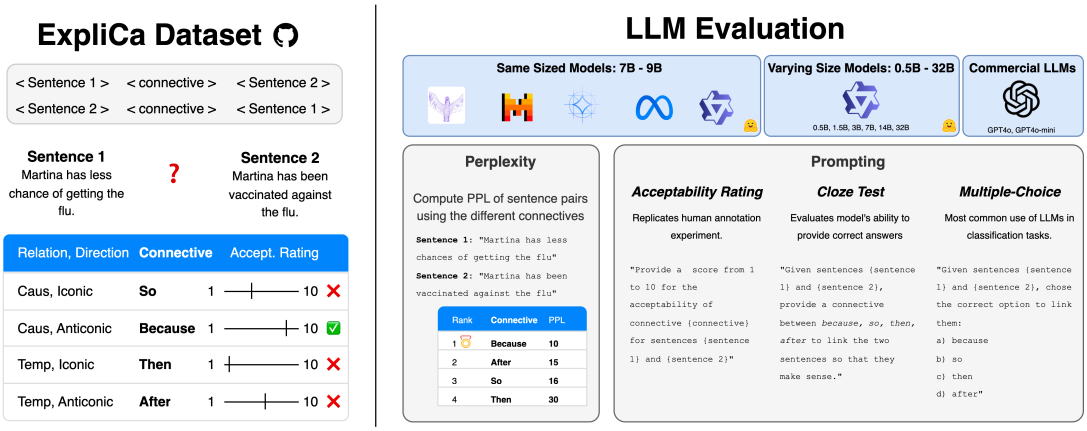

Figure 1: An overview of the contributions of this paper. On the left, the ExpliCa dataset, annotated with human acceptability ratings. On the right, our evaluation framework which leverages LLMs through PPL and prompting.

latter method underestimating the models' actual linguistic knowledge. Finally, we want to study these aspects across models of varying scales.

Contributions. The primary contributions of this paper are illustrated in Fig. 1 and can be summarized as follows:

- we introduce the ExpliCa dataset (Sec. 3), which is syntactically and semantically curated, ensuring high-quality data and lexically balanced in terms of word frequency, making it suitable for comprehensive analyses. ExpliCa has also been extensively annotated by human evaluators, providing a robust foundation for testing various models;
- we offer a framework for analyzing LLMs' ability to model causal and temporal relations. Our approach systematically targets models' *competence* (via perplexity) and *performance* (via prompting), measured across different formulations of the task (Sec. 4);
- we present a comprehensive evaluation of seven LLMs in total, comprising both commercial and open models (Sec. 5).[1]

2 Related Work

The study of causality and its linguistic expressions garnered renewed and intensified interest, particularly in the context of evaluating the reasoning abilities of LLMs. Recent advancements led to the development of several specialized datasets, aimed at testing the causal reasoning of LLMs through hypothetical scenarios. Notable examples include CLadder (Jin et al., 2023), which assesses causal reasoning using questions based on formal rules; CausalBench (Wang, 2024), used for tasks related to mathematics, coding, and textual data; and CausalNet (Ashwani et al., 2024), which covers both causal and counterfactual questions. Unlike ExpliCa, these datasets focus on implicit notions of causality, which are not overtly expressed in the linguistic structures of the text.

To evaluate LLMs, several datasets have also been annotated with explicit causal relationships between events in texts. These range from multilingual educational content, as in MECI (Lai et al., 2022), or financial news (Mariko et al., 2022), to more diverse sources, such as CREST (Hosseini et al., 2021). Although these datasets do not explore the temporal dimensions of causality, many causal annotation schemas are derived from datasets that do annotate temporal relationships between events (e.g., BECauSE, Dunietz et al. 2017) or vice-versa (e.g., Temporal and Causal Reasoning, Ning et al. 2018), including those from news sources (e.g., Causal Time Bank, Mirza et al. 2014; Event StoryLine Corpus, Caselli and Vossen 2017), and from short commonsense narratives (e.g., CaTeRS, Mostafazadeh et al. 2016). However, such datasets do not leverage crowdsourcing annotation by native speakers for both causal and temporal relations as in ExpliCa. In our dataset, the ground truth is given by English native speakers' annotation collected via crowdsourcing, with the aim of addressing the complexity of distinguishing truly causal from merely temporal relations between events. A key challenge in evaluating LLMs using datasets with direct textual annotations of

[1] Data and code at: https://anonymous.4open.science/r/ExpliCa-6473/.

causal relations is, in fact, the inherent ambiguity of their expression in natural language. For instance, linguistic markers such as *and* can signal either causality or temporality, depending on the context. This ambiguity can limit the effectiveness of such datasets in assessing causal reasoning.

To overcome this limitation, ExpliCa has chosen a more controlled approach to causality evaluation by conducting a pairwise analysis of events, each expressed by a single sentence. The same strategy has been adopted in the COPA dataset (Roemmele et al., 2011), where causality detection is framed as a task where the system must choose the most plausible alternative between two options. Similarly, e-CARE (Explainable Causal Reasoning, Du et al. 2022) includes over $21,000$ multiple-choice questions focused on causal reasoning, accompanied by conceptual explanations that clarify the underlying causal logic of each question. While these two datasets present instances of implicit causality, the BIG-Bench (Beyond the Imitation Game, Srivastava et al. 2022) initiative also models explicit causal reasoning. In this framework, the system must select the most plausible causal relationship between *A because B* and *B because A*. Similarly, in ExpliCa pairs of sentences from e-CARE and BIG-bench are joined in both directions ($A > B$; $B > A$), but using both temporal and causal connectives. This allows us to carefully analyze the models' ability to discriminate between the related and yet very different relations of temporal precedence and causality. Furthermore, other linguistic cues, such as anaphoric references, have been removed. This design ensures that models are unlikely to rely on surface features, preventing the correct interpretation of causal markers from being inferred simply from the syntactic context.

Finally, it is worth mentioning that some of the above datasets have become part of a broader evaluation framework called Causal evaluation of Language Models (CaLM, Chen et al. 2024). CaLM serves as a comprehensive benchmark for assessing LLMs' causal reasoning capabilities. It comprises $126,334$ data samples and provides a foundational taxonomy of four modules: causal target, adaptation, metric, and error analysis. In relation to causal discovery, this framework addresses issues distinct from those targeted by ExpliCa, and it focuses solely on the analysis of LLM-generated responses. By contrast, in our work, we evaluated both the model outputs elicited via prompts, and the internal knowledge of LLMs, assessed through perplexity measurements.

3 The ExpliCa Dataset

ExpliCa[2] is designed to evaluate LLMs on commonsense causal reasoning through PCD tasks. It is composed of sentence pairs, where each sentence describes an event. Sentence pairs were in part adapted from sentences in existing datasets and in part manually crafted. Approximately a third of the sentence pairs were based on sentences from DeScript (Wanzare et al., 2016), e-Care (Du et al., 2022), and BIG-Bench (Srivastava et al., 2022).[3]

ExpliCa includes 600 English sentence pairs, selected to be equally divided into three subsets: i.) CAUSAL subset, where the relationship is most likely causal (and possibly also temporal); ii.) TEMPORAL subset where the relation is expected to be only temporal; iii.) UNRELATED subset, including sentences that are thematically related but neither causally nor temporally.[4] Sentence pairs are linked through words that explicitly signal either a causal or temporal relationship. These connectives act as linguistic cues, enabling a causal or temporal interpretation of events based on the compositional meaning of the sentences. Other than the TYPE of the relation intercurring among the events, the connectives specify the ORDER of the events in the relations, which can be ICONIC — where the effect follows the cause and events are presented in their chronological order — or ANTI-ICONIC, where the order is reversed. We join each sentence pair with each connective, also considering the reverse order. This way, we obtained $4,800$ unique items (600 pairs $\times$ 4 connectives $\times$ 2 orders).

An English native speaker examined a sample of the dataset to validate if the connectives correctly express the nature of the relation and the order between the events. We detail the connectives below:

- *then* - indicates a temporal relation in an iconic order: The first event precedes the second event;
- *after* - indicates a temporal relation in an anti-iconic order: The first event follows the second event;
- *so* - indicates a causal relation in an iconic order: The first event causes the second event;

[2] Release license details are in App. A.

[3] Cf. Sec. 2 for a brief description of these datasets.

[4] Note that the division was done by the authors during the creation of the dataset, but the final annotations provided as gold standard are based on the human ratings from the survey.

- *because* - indicates a causal relation in an anti-iconic order: The first event is a consequence of the second event.

Crucially, the connectives serve as the only linguistic cue for a causal or temporal interpretation of events. Other potential cues, such as causal verbs (i.e., *cause*, *result*, *produce*, *affect*, etc.) were not included. Moreover, to avoid biasing humans' and models' decisions, anaphoric pronouns were not used. For example, instead of *The coffee was bitter, The child spit it out*, we overtly expressed the object in the second sentence: *The coffee was bitter, The child spit the coffee out.*

Then we applied a rigorous procedure to enrich, validate, and refine the dataset. First, we validated the dataset assuring that the statistical association between words in sentence pairs does not significantly differ across the three subsets (i.e., UNRELATED, TEMPORAL, CAUSAL). This validation was performed using Mutual Information (Sec. 3.1). Hence, we enriched each item with human acceptability ratings collected via crowdsourcing, and we set a threshold to the ratings for unrelated ones (Sec. 3.2). Finally, we checked for frequency biases in the dataset by analyzing the triplets {1st sentence verb, connective, 2nd sentence verb} in enTenTen (Jakubíček et al., 2013)(Sec. 3.3).

3.1 Lexical Association Bias

If unrelated sentence pairs presented very different lexical elements compared to causally or temporally related ones, this might affect a LLM behavior, and lead to biased results. In order to address this aspect, we computed the statical association strength via Pointwise Mutual Information (PMI) and Local Mutual Information (LMI) (Church and Hanks, 1990; Evert, 2004) between pairs of lexemes (nouns, verbs, or adjectives),[5] one belonging to the first sentence and the other to the second sentence. MI scores were computed on UkWaC (Ferraresi et al., 2008). We averaged the MI scores of all the possible pairs of lexemes from a single sentence pair to obtain an **item-level MI score**. We applied the Wilcoxon test to check if there were statistically significant differences in the item-level-MI scores across the CAUSAL, TEMPORAL, and UNRELATED groups. We found that the statistical association of the sentence pairs in the three groups was not significantly different, both for LMI (W: $41,576$, p-value: 0.4312) and PMI (W: $38,318$, p-value: 0.4009). Thus, we can conclude that our dataset is free from lexical association biases.

[5]We used Stanza (Qi et al., 2020) for PoS-tagging.

Table 1: The ground truth of ExpliCa according to human acceptability ratings.

Rel. Type	Rel. Order	Connective	# Sentences
Temporal	Iconic	Then	1,040
Temporal	Anti-iconic	After	656
Causal	Iconic	So	820
Causal	Anti-iconic	Because	876
Unrelated	-	-	1,408

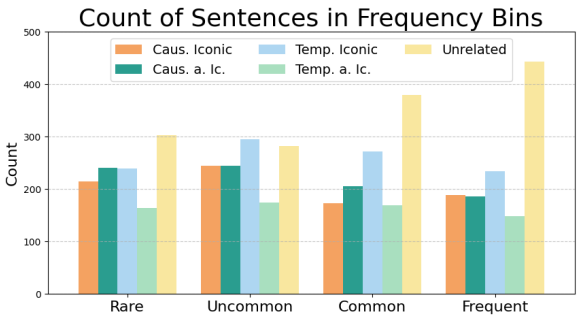

Figure 2: Number of sentence pairs (categorized by relation TYPE and ORDER) in each frequency bin.

3.2 Human Ratings

Each of the $4,800$ items in ExpliCa was annotated via crowdsourcing by 15 native English speakers.[6] We asked the participants to assess the acceptability of each item by giving a rating from 1 to 10. The ratings were then averaged for each item, obtaining the average acceptability rating for each connective in each sentence pair, in both event directions. We assigned to each sentence pair a relation TYPE and a relation ORDER label based on the connective deemed more acceptable for humans. For example, if *then* obtained the highest average rating for a sentence pair (in a specific direction), this is labeled as TEMPORAL TYPE with an ICONIC ORDER. Sentence pairs for which no connective had a rating higher than 6 and with mean rating below 5 are labeled UNRELATED. Tab. 1 summarizes the cardinality of the different classes in the dataset.

3.3 Frequency Bias

The frequency of linguistic constructions affects the performances of LLMs (McCoy et al., 2023). For example, if the construction {*listen*, *after*, *turn up*} were more frequent than {*listen*, *then*, *turn up*}, a model might be biased towards the former.

[6]See App. B for details on the annotation procedure.

To control for such frequency biases in ExpliCa, we proceeded as follows. We PoS-tagged and lemmatized the sentence pairs with SpaCy.[7] Then, we extracted the verb from each sentence, and the connective used to join them. For copulative verbs, we also considered the noun or the adjective following the copula. We retained particles in phrasal verbs (e.g., *turn up*), and in sentences containing multiple verbs (e.g., with embedded clauses), we considered up to two verbs, prioritizing those most salient to the sentence meaning (e.g., *Michele chose the pizza he wanted to eat* [want, eat]). Then, we used the SketchEngine APIs[8] to query the enTenTen21 corpus to compute the co-occurrence of the elements in the triplet {1st sentence verb, connective, 2nd sentence verb}. We divided the co-occurrences into frequency bins based on quartile ranges: RARE, UNCOMMON, COMMON, and FREQUENT. Fig. 2 shows the number of sentence pairs for each bin. While there are differences between classes, their distribution on the frequency spectrum shows no significant trends.[9] This suggests that our dataset is relatively free from frequency biases.

4 Experiments

We evaluated models on ExpliCa using human ratings as ground truth across four tasks: three **prompting tasks**, and one **perplexity evaluation**. Prompting experiments were conducted under various settings, including few-shot and zero-shot setups, and employing either greedy search[10] or the Outlines framework (Willard and Louf, 2023) for generating answers. We used **accuracy** as evaluation metric. To study the effect of the model parameter scale, we then compared several models of the Qwen2.5 family on the acceptability rating task and perplexity. Finally, we compared the models' rating distributions in the acceptability tasks to the human ones, and assessed their **correlation**.

Experiments were ran on a single Nvidia A100 80GB GPU, for around 120 GPU hours. OpenAI models were queried via the proprietary API for an estimate of 7 GPU hours.

Prompting Evaluation. This aimed to assess LLMs' generation abilities and analyze how performance varies based on task modeling. Specifically, we defined three tasks:

i.) **acceptability ratings** - we adopted the same design used in the survey with human participants (Sec. 3.2). Items for which the model failed to provide a rating were assigned a score of -1;

ii.) **cloze test** - given a test item consisting of two sentences linked by a connective, we masked the connective and asked the model to choose the most suitable one out of a list of candidates. An out-of-list answer was considered a miss;

iii.) **multiple-choice task** - the model received a sentence pair with the four connectives marked as A, B, C, D, and tasked to return the letter corresponding to the appropriate connective. Failure to provide one of the options was considered a miss.

We collected data from a single prompt per task, and each underwent a selection process. We drafted a first prompt, and then used ChatGPT to obtain four more variants of it.[11] We averaged the perplexity of all the open models on each prompt and chose the one with the lowest average perplexity.

We randomized the order of few-shot examples, options to choose from, and correct answers during inference. In the few-shot scenario, the models saw one example for each connective.[12]

Perplexity Evaluation. We computed the perplexity (PPL) of each item in the dataset, and grouped those corresponding to the same sentence pair. Then, we chose the connective from the item with the lowest PPL. We derived from the connective the TYPE and ORDER of the relation and computed models' accuracy by comparing these results with the human ground truth obtained with crowdsourcing annotation. We call this accuracy as **Accuracy Perplexity Score (APS).**

4.1 Models

We selected 7 generative LLMs. Specifically, we selected two open-weights models (**Mistral-7B-Instruct-v0.3** and **falcon-7b-instruct**), three partially open models (**Meta-Llama-3.1-8B-Instruct**, **gemma-2-9b-it**, and **Qwen2.5-7B-Instruct**), and two commercial models, **gpt4o** and **gpt4o-mini**. Perplexity evaluation was not performed on commercial models as it is not permitted through the API. We used **Qwen2.5 instruct** models of different sizes (from 0.5B to 32B parameters) for analyzing the impact of model scale.[13]

[7] spacy.io, English model en_core_web_sm.
[8] https://www.sketchengine.eu/apidoc/
[9] Sentence count for each frequency bin in Tab. 3, App. C.
[10] More details on answer cleaning in App. D.
[11] https://chatgpt.com/, used November 2024.
[12] Selected prompts and perplexity scores are in App. E.
[13] More details on models in App. F.

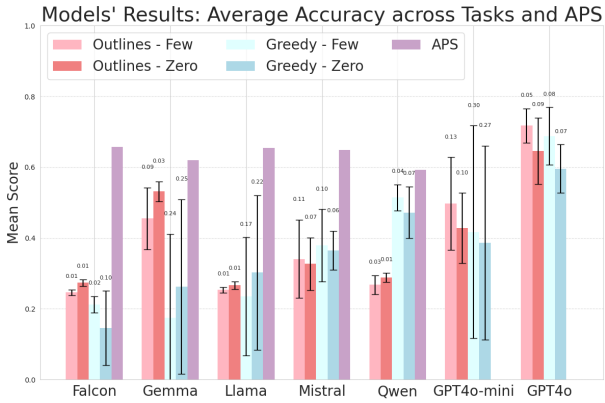

Figure 3: Average models' accuracy (on prompting tasks and perplexity) obtained from all the tasks on causal and temporal related sentence pairs. The numbers on top of each bar represent the standard deviation.

Table 2: Models' accuracy and SD across prompting tasks, reported for i.) all models, ii.) open models, and iii.) GPT variants. The overall best average is in **bold**, the top few/zero-shot result is **underlined and bold**, and the best per task is underlined.

Task	Greedy Search			Outlines		
	All	*Open*	*GPTs*	*All*	*Open*	*GPTs*
			Few-shot			
Acc.	0.46±0.2	0.35±0.1	0.75±0.0	0.38±0.2	0.25±0.1	0.71±0.1
Cloze	0.42±0.2	0.34±0.2	0.62±0.1	0.43±0.2	0.35±0.1	0.62±0.1
M.C.	0.24±0.2	0.22±0.2	0.29±0.3	0.38±0.2	0.34±0.1	0.50±0.2
Avg	0.37±0.2	0.30±0.2	0.55±0.3	**0.40**±0.2	0.31±0.1	**0.61**±0.2
			Zero-shot			
Acc.	0.49±0.2	0.43±0.2	0.66±0.0	0.40±0.2	0.32±0.1	0.62±0.2
Cloze	0.38±0.1	0.32±0.1	0.54±0.0	0.40±0.1	0.35±0.1	0.53±0.0
M.C.	0.21±0.2	0.18±0.2	0.27±0.3	0.38±0.1	0.34±0.1	0.46±0.2
Avg	0.36±0.2	0.31±0.2	0.49±0.2	**0.39**±0.2	0.34±0.1	0.54±0.2

5 Results and Discussion

We carried out an in-depth analysis of several interesting aspects emerging from the experiments, as outlined in the following.

Core results are summarized in Fig. 3,[14] which shows the average accuracy across all three prompting tasks for a specific model in a specific setting (e.g., Outlines in zero-shot). Violet bars represent the APS (a single numerical value) for each open model. These results are complemented by Tab. 2, which compares performances across prompting tasks, aggregated by model type.

The main trends observed are: i.) performance variability is high among models of the same scale in prompting tasks, despite similar APSs across open models; ii.) GPT4o outperforms all models, but its mini variant does not follow this trend and is still far from human performance; iii.) accuracy standard deviation shows that most open models exhibit inconsistent results across tasks, particularly with greedy search (Sec. 5.1); iv.) prompting performance is systematically lower than competence-level APSs in open models (Sec. 5.2).

5.1 Modeling PCD with Prompting

Tab. 2 shows that, on average, PCD performed best when responses were framed using Outlines in a few-shot setting (0.40 overall, 0.61 for GPT models). However, the highest task-specific accuracy was achieved in the acceptability rating task under a zero-shot setting with greedy search (0.49).

Moreover, the Outlines framework proved to be beneficial for all models on the multiple-choice task. Conversely, for the cloze test and acceptability rating task, performances either remained roughly the same or decreased substantially for both open models and GPTs. Strikingly, **the top average accuracy of** 0.61 **in Tab. 2 still reveals a strong gap with human data**.

Fig. 3 shows that each model performs best under different settings. Each of GPT4o, Gemma, and Qwen achieved the highest mean accuracy (0.72, 0.53, and 0.51, respectively) in a distinct setup, surpassing GPT4o-mini (0.50). Standard deviation analysis (sd) indicates that task selection significantly affects smaller models' performance, especially with greedy search (sd 0.2 as in Tab. 2). Notably, GPT4o-mini exhibits a much wider performance range across both few-shot and zero-shot settings.

5.2 Modelling PCD with Perplexity

The APS results of the open models are shown in Fig. 3. The scores obtained by relying solely on perplexity are significantly higher (0.63 overall average) compared to those achieved in the prompting tasks.[15] Although Falcon performs poorly in the latter, it is the model with the best APS (0.66). By leveraging perplexity, the lowest APS too, reached by Qwen with 0.59, surpasses the other open models in prompting tasks and GPT4o-mini as well. Although this result does not match GPT4o performances, it seems to confirm that models' competence about causal relations encoded in their probabilistic predictions is more accurate than their prompting performances (Hu and Levy, 2023).

From these analyses we conclude that ExpliCa

[14] Detailed results are shown in Tab. 5, App. G.

[15] Perplexity values are in Tab. 8, App. G.

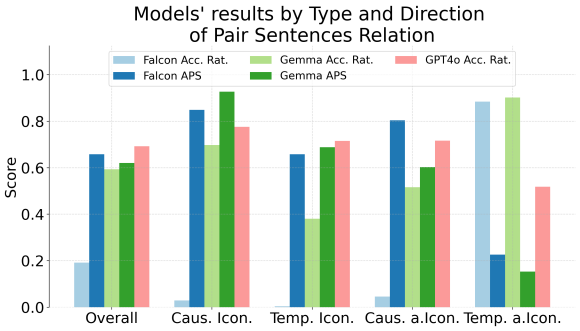

Figure 4: Accuracy of models by relation TYPE and ORDER. APS stands for Accuracy Perplexity score.

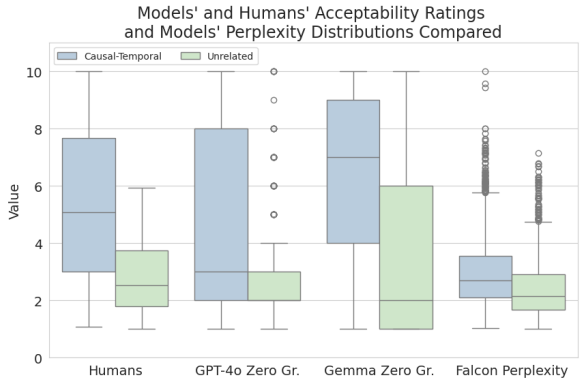

Figure 5: Distribution of acceptability ratings for humans and Gpt4o and Gemma, and Falcon's normalized perplexity.

is a highly challenging dataset for explicit causal reasoning evaluation. State-of-the-art LLMs like GPT4o are not able to fully solve the dataset, while smaller variants perform worse, exhibiting high variability depending on the prompt and generation strategy. Notably, when evaluated via APS, small open models can sometimes outperform larger (commercial) models.

5.3 In-depth Analysis on ExpliCa

The ExpliCa design allows us to provide a more in depth analysis of causal reasoning and LLMs.

Relations' type and order. First, we focus on the relation TYPE between the events expressed in each dataset item, and their ORDER. We report results on the acceptability rating prompting task, with zero-shot and greedy search (i.e., where the models performed best overall).[16] We specifically focus on GPT4o and two open models, namely Gemma and Falcon, with the latter achieving the best APS. In Fig. 4, we see that models perform best on the CAUSAL relations in ICONIC order, with the exception of Falcon.[17] Worse performances are obtained for CAUSAL ANTI-ICONIC and TEMPORAL ICONIC, with the latter often being mistaken for CAUSAL ICONIC.

An interesting pattern emerges by comparing APSs and Accuracy on acceptability ratings. APSs for a model are always higher than the respective acceptability ratings' accuracy, except for TEMPORAL ANTI-ICONIC items.[18] However, by observing the models' behavior,[19] we saw that both Gemma and Falcon tend to favor the TEMPORAL ANTI-ICONIC connective despite showing lower perplexity for the connective expressing a CAUSAL ICONIC relation (i.e., *so*). The performances of GPT4o do not seem to suffer the same biases toward specific cases. Nevertheless, intra-cases differences are up to 26%.

Overall, results show that open models may be rather inconsistent in how they address causal and temporal relations and are highly influenced by how the task is formulated (i.e., by using prompting or perplexity-based scores). Moreover, while LLMs are quite good at identifying causally-related events, **they tend to confound temporal relations with causal ones, and their performance is also strongly influenced by the linguistic order of presentations of the events. This suggests that, compared to humans, they have a less general and abstract knowledge of causal (and temporal) relations**.

Correlation and distribution variation. We compared human ratings with model-generated ones in terms of their distribution and correlation. We show a comparison of models acceptability ratings with humans' for the causal-temporal and unrelated subsets in Fig. 5.[20] We consider ratings of GPT4o and Gemma,[21] and the (normalized) perplexity scores for Falcon, as it obtained the best APSs. A noticeable difference exists between human ratings of CAUSAL and TEMPORAL items vs. UNRELATED and model scores.

Fig. 6 shows Spearman correlation computed among human ratings and models' acceptability ratings collected in zero-shot scenario.[22] We observed

[16] All the models' results, in all settings, are in App. G.
[17] Values are shown in Tab. 9, App. G.
[18] As for most of the models (see Tab. 8, App. G).
[19] See App. H, Fig. 12 for the confusion matrices.
[20] Answers without a rating are not shown in the distribution.
[21] Results for the zero-shot prompt with greedy search. These are the models with the highest accuracy for this task.
[22] Individual results in Tab. 13, App. I.

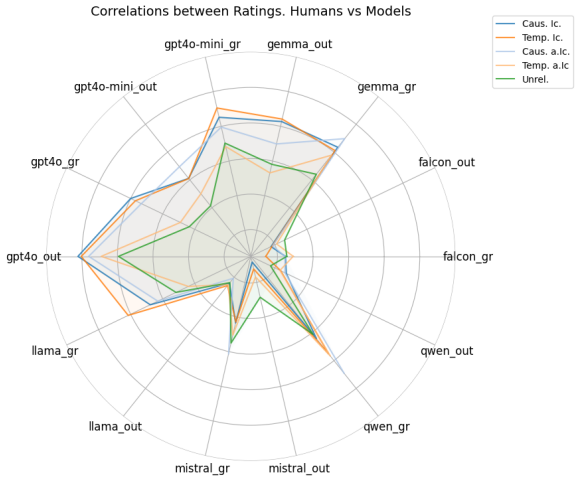

Figure 6: Spearman correlation between humans' ratings and models' results. Ratings from the zero-shot task, both with greedy (_gr) and Outlines (_out).

that even though GPT4o, Gemma, and Qwen have a strong correlation with human ratings, this greatly varies according to the items' condition. In particular, the model scores for TEMPORAL ANTI-ICONIC and UNRELATED items have a much lower correlation with human ratings.

Size Effect. We analyzed how the model size affects performance on PCD. We selected the Qwen model family, available in a wide range of sizes, and used the acceptability rating and perplexity tasks as a testbed.[23] Fig. 7 shows that Qwen's performance and APS improve with model size, except for a slight drop for the latter in the largest variant. Nonetheless, the improvement rates are quite different. Performances seem to linearly scale with size, whereas the APS growth curve is flatter and eventually plateaus. The initial APS is markedly higher (2x) than the respective accuracy; accuracy and APS are near equal at the 14B mark; the accuracy of the 32B keeps improving while the APS is stale. A similar trend holds true also for each relation,[24] despite showing a higher variability. For example, the 0.5B variant is quite proficient with ANTI-ICONIC relations but almost incapable of modelling ICONIC ones. Model's scale seem to correlate with less differences in performances among relations.

This might suggest that size mainly affects the model's performance on prompting tasks, rather than their competence about causal and temporal relations. These are likely to be already encoded – though partially, as observed above – in smaller models, although they lack the necessary scale to properly use such competence in generation.

[23] Few-shot and greedy search setup, i.e. the best for Qwen.
[24] Results for each are shown in App. J.

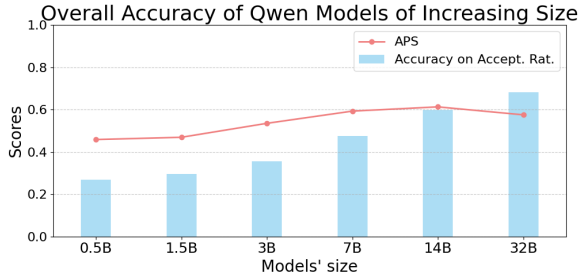

Figure 7: Results on Accuracy for Acceptability Ratings and APS for Qwen models of increasing size.

6 Conclusion and Future Work

In this paper, we introduced ExpliCa, a dataset designed to evaluate LLMs on PCD tasks with the aim to assess their reasoning abilities on explicitly expressed relations. ExpliCa is the first dataset to incorporate both causal and temporal relations between events, with acceptability ratings provided by native speakers via crowdsourcing.

Results indicate that ExpliCa is particularly challenging, even for commercial models like GPT4o-mini, which is outperformed by open models when leveraging probabilistic scores, and GPT4o, which struggles to reach 0.80 accuracy. We observed that LLMs exhibit variable performance according to both the evaluation setup and the relation TYPE and ORDER. The models' rating distribution do not approximate the humans' ratings, especially for certain types of relations and linguistic orders. Moreover, our results suggest that the competence on causal relations measured with perplexity is significantly more accurate than prompting performance (especially in smaller models).

For future works, we aim to further explore the effect of model size on different model families, and we plan to adopt ExpliCa also to investigate how models interpret implicit causality.

So far, the presented results reveal that, despite their increasing size, **the knowledge of causal relations is still suboptimal in LLMs**, which also show a strong tendency to confound temporal relations with causal ones, compounded with a limited abstraction from the surface linguistic presentation order of the events. e can conclude that, although LLMs have undoubtedly progressed along the 'ladder of causation' (Pearl, 2009), they have still several steps to climb.

7 Limitations

This study has several limitations that should be acknowledged. The prompts used across different models were not specifically optimized for each of them. However, this decision was necessary to maintain the feasibility of our experiments and ensure a fair comparative evaluation among all models and tasks. Computational constraints played a significant role in shaping our methodology, influencing aspects such as batch size, model capacity, and the overall scope of the analysis. While we examined causal reasoning through various tasks, the selection of prompts, possible choices, and connectives represents only a subset of all potential analytical strategies. Additionally, the number of closed-source models included in the analysis was limited, and the dataset size was relatively small, further constraining the generalizability of the findings. Experiments on model scaling were confined to the Qwen model, where we observed that larger models tend to equalize performance on tasks and competence. However, this observation requires further validation to confirm that it is not influenced by factors such as the specific prompt or model used, which would bolster the robustness of our claims. Among the CAUSAL sentence pairs in ExpliCa, 50 are labeled as 'socially challenging', indicating content that touches on sensitive or potentially offensive topics such as religion, abortion, immigration, gender identity, drug abuse, and bribery. We acknowledge that some sentences may be offensive to certain groups, but these themes were added to evaluate whether bias-mitigation strategies in LLMs would impact PCD performance. Due to space constraints, we plan to explore such aspects more in depth in future works. Finally, we acknowledge that we focused only on LLMs and we did not include the so-called Large Reasoning Models (LRM) like OpenAI o1 (OpenAI, 2024b) or the very recent DeepSeek-R1 (DeepSeek-AI, 2025), which may have an advantage in such a task. This choice was mainly due to financial constraints, but they will be explored in future works.

Acknowledgments

Acknowledgments are anonymized during review.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The Falcon Series of Open Language Models. *Preprint*, arXiv:2311.16867.

Swagata Ashwani, Kshiteesh Hegde, Nishith Reddy Mannuru, Mayank Jindal, Dushyant Singh Sengar, Krishna Chaitanya Rao Kathala, Dishant Banga, Vinija Jain, and Aman Chadha. 2024. Cause and effect: Can large language models truly understand causality? *arXiv preprint arXiv:2402.18139*.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623. ACM.

Tom B. Brown, Benjamin Mann, Nick Ryder, and Melanie Subbiah et al. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Tommaso Caselli and Piek Vossen. 2017. The event storyline corpus: A new benchmark for causal and temporal relation extraction. In *Proceedings of the Events and Stories in the News Workshop*, pages 77–86.

Sirui Chen, Bo Peng, Meiqi Chen, Ruiqi Wang, Mengying Xu, Xingyu Zeng, Rui Zhao, Shengjie Zhao, Yu Qiao, and Chaochao Lu. 2024. Causal evaluation of language models. *arXiv preprint arXiv:2405.00622*.

Kenneth Church and Patrick Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational linguistics*, 16(1):22–29.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

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

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jesse Dunietz, Lori Levin, and Jaime G Carbonell. 2017. The because corpus 2.0: Annotating causality and overlapping relations. In *Proceedings of the 11th Linguistic Annotation Workshop*, pages 95–104.

Stefanie Evert. 2004. *The statistics of word cooccurrences: word pairs and collocations*. Ph.D. thesis, University of Stuttgart.

Adriano Ferraresi, Eros Zanchetta, Marco Baroni, Silvia Bernardini, et al. 2008. Introducing and evaluating ukwac, a very large web-derived corpus of english. In *Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google*, pages 47–54.

Jinglong Gao, Xiao Ding, Bing Qin, and Ting Liu. 2023. Is chatgpt a good causal reasoner? a comprehensive evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11111–11126.

Pedram Hosseini, David A Broniatowski, and Mona Diab. 2021. Predicting directionality in causal relations in text. *arXiv preprint arXiv:2103.13606*.

Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. In *Proceedings of EMNLP*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Miloš Jakubíček, Adam Kilgarriff, Vojtěch Kovář, Pavel Rychlỳ, and Vít Suchomel. 2013. The tenten corpus family. In *7th international corpus linguistics conference CL*, pages 125–127. Valladolid.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, and Diego de las Casas et al. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, LYU Zhiheng, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, et al. 2023. Cladder: Assessing causal reasoning in language models. In *Thirty-seventh conference on neural information processing systems*.

Carina Kauf, Emmanuele Chersoni, Alessandro Lenci, Evelina Fedorenko, and Anna A Ivanova. 2024. Log probabilities are a reliable estimate of semantic plausibility in base and instruction-tuned language models. In *Proceedings of the EMNLP BlackBoxNLP Workshop on Analysing and Interpreting Neural Networks*.

Viet Dac Lai, Amir Pouran Ben Veyseh, Minh Van Nguyen, Franck Dernoncourt, and Thien Huu Nguyen. 2022. Meci: A multilingual dataset for event causality identification. In *Proceedings of the 29th international conference on computational linguistics*, pages 2346–2356.

Dominique Mariko, Hanna Abi Akl, Kim Trottier, and Mahmoud El-Haj. 2022. The financial causality extraction shared task (fincausal 2022). In *Proceedings of the 4th Financial Narrative Processing Workshop@ LREC2022*, pages 105–107.

R Thomas McCoy, Shunyu Yao, Dan Friedman, Matthew Hardy, and Thomas L Griffiths. 2023. Embers of autoregression: Understanding large language models through the problem they are trained to solve. *ArXiv*.

William Merrill, Zhaofeng Wu, Norihito Naka, Yoon Kim, and Tal Linzen. 2024. Can you learn semantics through next-word prediction? the case of entailment. *arXiv preprint arXiv:2402.13956*.

Paramita Mirza, Rachele Sprugnoli, Sara Tonelli, and Manuela Speranza. 2014. Annotating causality in the tempeval-3 corpus. In *Proceedings of the EACL 2014 Workshop on Computational Approaches to Causality in Language (CAtoCL)*, pages 10–19.

Nasrin Mostafazadeh, Alyson Grealish, Nathanael Chambers, James Allen, and Lucy Vanderwende. 2016. Caters: Causal and temporal relation scheme for semantic annotation of event structures. In *Proceedings of the Fourth Workshop on Events*, pages 51–61.

Qiang Ning, Zhili Feng, Hao Wu, and Dan Roth. 2018. Joint reasoning for temporal and causal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2278–2288.

OpenAI. 2024a. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI. 2024b. Openai o1 system card. *Preprint*, arXiv:2412.16720.

Judea Pearl. 2009. *Causality*. Cambridge University Press, New York, NY, USA.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI spring symposium series*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. 2024a. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024b. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Guangya Wan, Yuqi Wu, Mengxuan Hu, Zhixuan Chu, and Sheng Li. 2024. Bridging causal discovery and large language models: A comprehensive survey of integrative approaches and future directions. *arXiv preprint arXiv:2402.11068*.

Zeyu Wang. 2024. Causalbench: A comprehensive benchmark for evaluating causal reasoning capabilities of large language models. In *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*, pages 143–151.

Lilian DA Wanzare, Alessandra Zarcone, Stefan Thater, and Manfred Pinkal. 2016. A crowdsourced database of event sequence descriptions for the acquisition of high-quality script knowledge. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3494–3501.

Brandon T Willard and Rémi Louf. 2023. Efficient guided generation for llms. *arXiv preprint arXiv:2307.09702*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Matej Zečević, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. 2023. Causal parrots: Large language models may talk causality but are not causal. *arXiv preprint arXiv:2308.13067*.

Table 3: Sentences count for each frequency bin across different item TYPES and ORDERS.

Category Bin	C. Ic.	C. a.Ic.	T. Ic.	T. a.Ic	Unrel.
Rare (0-4)	214	241	239	164	303
Uncommon (5-56.4)	244	244	295	174	282
Common (56.5-498)	173	205	272	169	380
Frequent (499-7M)	189	186	234	149	443

Appendix

A Dataset License

The dataset is made publicly available under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (`CC BY-NC-SA 4.0`) license, which allows redistribution, adaptation, and reuse for non-commercial purposes, provided proper attribution is given and derivative works are shared under the same terms. However, the dataset cannot be used for training artificial intelligence models or other machine learning systems.

B Human Annotation

We collected acceptability ratings from human annotators for each of the 4800 items in ExpiCa. We used the crowdsourcing platform Prolific[25], which ensures fair compensation for participants. For our study, we recruited native English speakers born and residing in the UK or USA. Each item was examined by 15 participants. Prolific facilitated the distribution of Google Forms, through which participants provided formal consent to take part in the study and received the following instructions:

Acceptable use of the connective

This survey aims to investigate the acceptability of sentences each describing two events linked by the use of a temporal or causal connective among: "then", "after", "because", and "so". You must rate on a scale from 1 to 10 how acceptable the connective is to express the relation between the events in the sentences. For example:

- Jude walked under the rain for an hour, so Jude got sick: Rating 10 (highly acceptable)
- Mary bought some flowers, because Jean went to the dentist: Rating 1 (not acceptable)

IMPORTANT: There are three check questions in the survey. You must answer these check questions correctly to receive payment.
Read the following sentences, then rate each sentence by answering this question: **How acceptable is the sentence from 1 (unsuitable connective) to 10 (suitable connective)?**

C Frequency Analysis

Tab. 3 reports the number of sentences for each frequency bin computed on the enTenTen21 corpus. The Table highlights a slight difference in the distribution of UNRELATED items across the frequency spectrum compared to CAUSAL and TEMPORAL

[25] https://www.prolific.com/

Table 4: Perplexity of the models on five prompt variants for each task.

Task	Prompt	Mistral	Falcon	Qwen	Llama	Gemma	Avg
Accept. Rat.	0	9.99	23.2	14.7	13.74	18.85	**16.10**
	1	12.23	24.38	21.76	16.14	32.39	21.38
	2	15.59	25.05	14.37	17.7	38.81	22.30
	3	12.81	29.52	20.89	20.09	31.34	22.93
	4	14.62	29.86	18.25	17.81	30.62	22.23
Multiple-Choice	0	3.34	3.66	3.58	4.34	4.39	**3.86**
	1	4.06	4.6	5.87	6.6	7.57	5.74
	2	3.42	4.15	5.89	7.26	7.86	5.72
	3	3.99	6.09	11.42	10.68	8.51	8.14
	4	4.41	4.75	7.46	7.27	8.03	6.38
Cloze Test	0	13.41	15.52	20.41	14.2	21.38	16.98
	1	9.03	12.42	24.07	20.05	19.98	17.11
	2	9.87	12.06	16.64	13.39	15.02	**13.40**
	3	12.72	13.91	21.18	21.34	23.73	18.58
	4	13.52	14.93	17.65	12.13	14.95	14.64

items in both ICONIC and ANTI-ICONIC orders. This pattern was expected, as verbs in unrelated sentences tend to be higher-frequency common English verbs. However, we account for their topical relatedness by measuring PMI and LMI among the lexemes in the sentence pairs.

D Models Answer Cleaning

In order to maximize the models' accuracy in greedy search, we processed the models' answers with regular expressions, to clean the returned text (e.g., from tags used to mark the end of the generated text, spaces, tabulations, symbols, punctuation, and motivation of their choices).

E Prompt Engineering

The prompts used during our experiments were selected by computing models' perplexity over five prompts per task, as described in Sec. 4. Tab. 4 shows the perplexity score given by the models to each prompt variant in zero-shot setting. Then the prompts with the lower average perplexity for each task were selected. The selected prompts are reported in the boxes below in the few-shot setting.

Acceptability Rating Prompt

Evaluate the acceptability of sentences that describe two events linked by connectives: 'so', 'because', 'after', and 'then'.
Rate each sentence on a scale from 1 to 10 based on how well the connective expresses the relationship between the events.
Examples:
- So (effect): "Jude walked under the rain for an hour, so Jude got sick." (Rating: 10)
- Because (cause): "Mary bought some flowers, because Jean went to the dentist." (Rating: 1)
- After (preceding event): "The girl finished her homework, after the girl put her books in the backpack." (Rating: 1)
- Then (following event): "James took the phone, then James called Clara." (Rating: 10)

Sentence: <Sentence>
Rating:

Cloze test Prompt

Select the word that best describes the relationship between the events in these two sentences.
Use this template: event in sentence 1 <word> event in sentence 2.
Choose from: ['thus', 'then', 'because', 'after'].
Provide only one word, no explanation.
Examples:
- Sentence 1: "Jude walked under the rain for an hour."
  Sentence 2: "Jude got sick."
  Answer: "so"
- Sentence 1: "Mary bought some flowers."
  Sentence 2: "Mary wants to give a present to her mom."
  Answer: "because"
- Sentence 1: "The girl put her books in the backpack."
  Sentence 2: "The girl finished her homework."
  Answer: "after"
- Sentence 1: "James took the phone."
  Sentence 2: "James called Clara."
  Answer: "then"

Sentences:
- Sentence 1: "sentence_1"
- Sentence 2: "sentence_2"

Answer:

Multiple-choice Prompt

Task Description:
You are given two sentences, Sentence A and Sentence B, and a list of words. Your task is to select the most appropriate word to connect the two sentences logically and coherently. The chosen word should fit grammatically and contextually.
Instructions:
1. Read Sentence A and Sentence B carefully.
2. Review the list of words provided.
3. Select the word that best connects the two sentences.

Format:
1. Sentence A: [Insert Sentence A here]
2. Sentence B: [Insert Sentence B here]
3. Words:
   - A. [Insert word A here]
   - B. [Insert word B here]
   - C. [Insert word C here]
   - D. [Insert word D here]
4. Answer: [Provide the letter of the correct word]

Examples:
- 1. Sentence A: "Jude walked under the rain for an hour."
  2. Sentence B: "Jude got sick."
  3. Words:
     - A. then
     - B. after
     - C. because
     - D. so
  4. Answer: D
- 1. Sentence A: "Mary bought some flowers."
  2. Sentence B: "Mary wants to give a present to her mom."
  3. Words:
     - A. because
     - B. then
     - C. after
     - D. so
  4. Answer: A
- 1. Sentence A: "The girl put her books in the backpack."
  2. Sentence B: "The girl finished her homework."
  3. Words:
     - A. because
     - B. after
     - C. so
     - D. then
  4. Answer: B
- 1. Sentence A: "James took the phone."
  2. Sentence B: "James called Clara."
  3. Words:
     - A. after
     - B. because
     - C. then
     - D. so
  4. Answer: C

Sentence Connection Task:
1. Sentence A: "sentence_a"
2. Sentence B: "sentence_b"
3. Words:
"multiple_choices"
4. Answer:

F Models Details

Open and partially open models used for our experiments are available via the Hugging Face model library (https://huggingface.co/models); all

models are used as per their licenses. We used the instruction-tuned version of the models as follows:

- **Mistral** (Jiang et al., 2023) leverages grouped-query attention (GQA) for faster inference, and sliding window attention (SWA) to effectively handle sequences of arbitrary length with a reduced inference cost. We used the 0.3 version of 7B parameters, which is provided with a larger vocabulary compared to the previous version. This model is available at: `https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3`.

- **Falcon** (Almazrouei et al., 2023) series of models was developed by the Technology Innovation Institute (TII). Pre-training data was collected from dumps from CommonCrawl after significant filtering (to remove machine-generated text and adult content) and deduplication. We used the 7B parameter model, which is available at: `https://huggingface.co/tiiuae/falcon-7b-instruct`.

- **Llama** (Large Language Model Meta AI) is a family of open-source LLMs developed by Meta AI. The first version of LlamA was released in February 2023. We adopted the version 3.1 of Llama (Dubey et al., 2024), which counts 8B parameters. This model is available at: `https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct`.

- **Qwen** is a LLM family built by Alibaba Cloud. We used the 2.5 version (Yang et al., 2024) which underwent an optimization process during both the pre-training and post-training stages. This model is available at: `https://huggingface.co/Qwen/Qwen2.5-7B-Instruct`. For experiments on size effect, we used Qwen 2.5 instruction-tuned models of increasing size, according to their number of parameters: 0.5, 1.5, 3, 7, 14, and 32 billions.[26]

- **Gemma** is a model released by Google (Team et al., 2024a) and it is built from the research and technology used to create Gemini models. We adopted version 2 of this model (Team et al., 2024b), with a size of 9B parameters. This is the biggest among the small open models we adopted for our experiments. This model is available at: `https://huggingface.co/google/gemma-2-9b-it`.

- The GPT (Generative Pre-trained Transformer) series of models (Brown et al., 2020) was developed by OpenAI. It is the commercial model under the ChatGPT platform. For our experiments, we used the state-of-the-art model GPT4o (Hurst et al., 2024), and a lighter version, GPT4o-mini (OpenAI, 2024a). The experiments were conducted between November 2024 and January 2025, by means of the APIs.[27]

G More Models' Results

Tab. 5 shows the results plotted in Fig. 3. Similar to GPT4o-mini, Gemma displays a high instability with greedy search, although it is notably reduced when using the Outlines framework. Llama follows a similar trend, whereas Falcon is quite stable also when using greedy search in zero-shot settings. In contrast, Mistral's results remain nearly the same across all scenarios, with a range similar to GPT4o. On the contrary, Qwen shows the opposite pattern, with more stable performances using greedy search.

Results on all tasks and setups with greedy search are reported in Fig. 8 and Tab. 6. Results with Outlines are reported in Fig. 9, Tab. 7.

Tab. 8 shows the overall APSs achieved by the models and those achieved according to various relations' conditions. APS changes according to the relation between the events contained in each item. Gemma and Llama best detect events in ICONIC ORDER, whereas Falcon, Mistral, and Qwen are better in identifying CAUSAL relations. All the models struggle at recognizing TEMPORAL ANTI-ICONIC relations, achieving really low results for this condition (0.19).

Tab. 9, shows the results plotted in Fig. 4, Sec. 5.3. In the table are reported APS and results on the acceptability rating task, with zero-shot and greedy search, i.e. where the models performed best overall. We specifically focus on GPT4o and two open models, namely Gemma and Falcon, with the latter achieving the best APS.

Post-hoc analyses were conducted on the obtained results to see if they might be affected by the frequency of the triplets {*1st sentence verb, connective, 2nd sentence verb*} in each item. Fig. 10

[26] The whole model family is available in this Hugging Face collection: `https://huggingface.co/collections/Qwen/qwen25-66e81a666513e518adb90d9e`.

[27] `https://platform.openai.com/docs/overview`

Table 5: Models' Greedy Search and Outlines accuracy results on all tasks (including APS when applicable) are averaged in few and zero-shot scenarios.

Model	Outlines				Greedy				APS
	Few-shot		Zero-shot		Few-shot		Zero-shot		
	Avg.	SD	Avg.	SD	Avg.	SD	Avg.	SD	
Falcon	0.25	0.01	0.27	0.01	0.21	0.02	0.15	0.10	0.66
Gemma	0.46	0.09	0.53	0.03	0.18	0.24	0.26	0.25	0.62
GPT4o-mini	0.50	0.13	0.43	0.10	0.42	0.30	0.39	0.27	-
GPT4o	0.72	0.05	0.65	0.09	0.69	0.08	0.60	0.07	-
Llama	0.25	0.01	0.27	0.01	0.23	0.17	0.30	0.22	0.65
Mistral	0.34	0.11	0.33	0.07	0.38	0.10	0.36	0.06	0.65
Qwen	0.27	0.03	0.29	0.01	0.51	0.04	0.47	0.07	0.59

Figure 8: Accuracy scores for all models in all prompting tasks in zero and few-shot scenarios with greedy search. APSs are in blue as *perpl*.

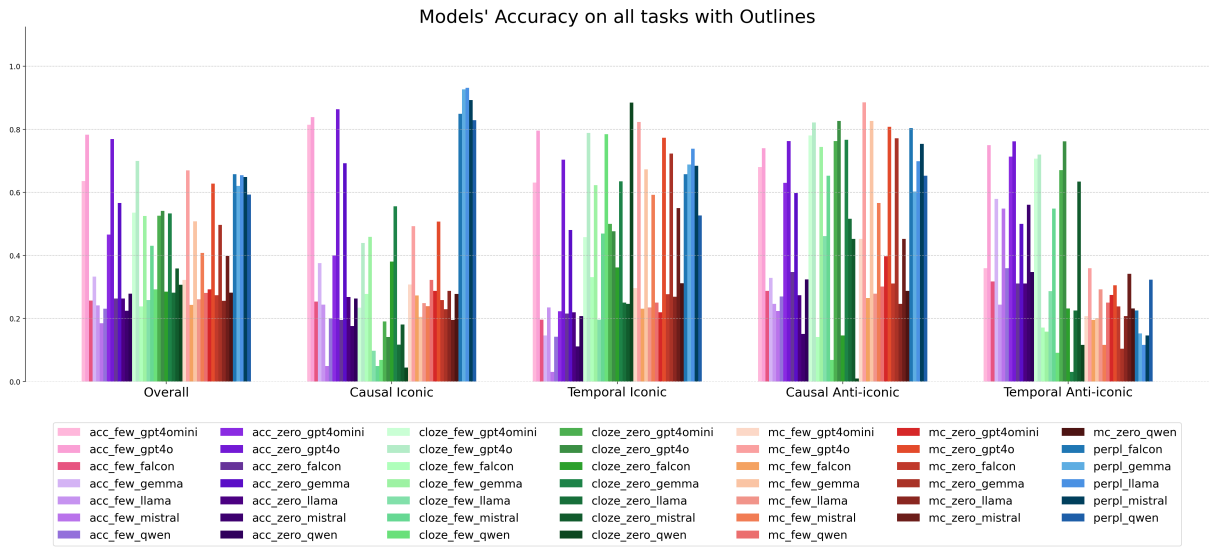

Figure 9: Accuracy scores for all models in all prompting tasks in zero and few-shot scenarios with Outlines. APSs are in blue as *perpl*.

Table 6: Performance comparison across Acceptability, Cloze test, and Multiple-choice tasks for few and zero-shot settings with greedy search.

GREEDY SEARCH										
Acceptability Rating	FEW					ZERO				
	All	Caus. Ic.	Temp. Ic.	Caus. a.Ic.	Temp. a.Ic.	All	Caus. Ic.	Temp. Ic.	Caus. a.Ic.	Temp. a.Ic.
GPT4o-mini	0.71	0.81	0.82	0.73	0.37	0.63	0.63	0.58	0.63	0.73
GPT4o	0.78	0.83	0.8	0.74	0.75	0.69	0.78	0.72	0.72	0.52
Falcon	0.2	0.01	0.0	0.03	0.99	0.19	0.03	0.0	0.05	0.88
Gemma	0.51	0.42	0.41	0.5	0.79	0.59	0.7	0.38	0.52	0.9
Llama	0.33	0.76	0.23	0.1	0.24	0.52	0.46	0.48	0.5	0.66
Mistral	0.25	0.13	0.02	0.1	0.95	0.29	0.12	0.01	0.27	0.95
Qwen	0.48	0.5	0.36	0.56	0.52	0.54	0.38	0.37	0.78	0.71
Cloze test	FEW					ZERO				
	All	Caus. Ic.	Temp. Ic.	Caus. a.Ic.	Temp. a.Ic.	All	Caus. Ic.	Temp. Ic.	Caus. a.Ic.	Temp. a.Ic.
GPT4o-mini	0.54	0.25	0.48	0.79	0.66	0.52	0.19	0.5	0.75	0.66
GPT4o	0.69	0.44	0.77	0.8	0.74	0.55	0.16	0.49	0.83	0.76
Falcon	0.24	0.87	0.0	0.01	0.16	0.25	0.52	0.05	0.21	0.26
Gemma	0.02	0.02	0.02	0.02	0.0	0.2	0.45	0.1	0.21	0.01
Llama	0.38	0.06	0.38	0.57	0.51	0.39	0.01	0.43	0.83	0.21
Mistral	0.49	0.4	0.48	0.57	0.53	0.41	0.32	0.32	0.44	0.64
Qwen	0.56	0.37	0.78	0.65	0.35	0.37	0.1	0.86	0.08	0.32
Multiple-chioce	FEW					ZERO				
	All	Caus. Ic.	Temp. Ic.	Caus. a.Ic.	Temp. a.Ic.	All	Caus. Ic.	Temp. Ic.	Caus. a.Ic.	Temp. a.Ic.
GPT4o-mini	0.0	0.0	0.01	0.0	0.01	0.0	0.0	0.0	0.0	0.01
GPT4o	0.59	0.47	0.68	0.77	0.33	0.54	0.42	0.62	0.72	0.36
Falcon	0.19	0.2	0.17	0.23	0.17	0.0	0.0	0.0	0.0	0.0
Gemma	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Llama	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.01	0.0
Mistral	0.4	0.2	0.58	0.57	0.12	0.39	0.18	0.55	0.44	0.34
Qwen	0.5	0.42	0.65	0.52	0.35	0.5	0.29	0.84	0.56	0.16

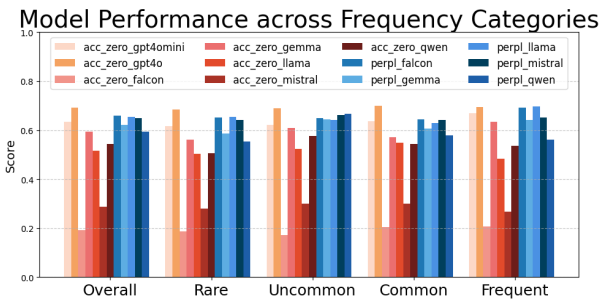

Figure 10: Accuracy scores grouped by frequency bin. The results are shown in red for acceptability ratings in zero-shot scenario with greedy search (i.e., the task where models obtained the best cumulative performance); in blue, the APSs of open models (as *perpl*).

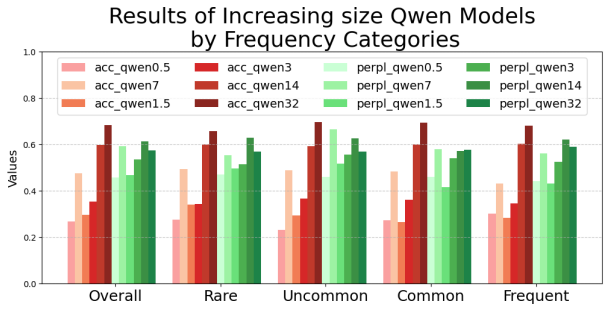

Figure 11: Accuracy scores grouped by frequency bin of increasing size of Qwen Models. The results are shown in red for acceptability ratings in few-shot scenario with greedy search (i.e., the setting where Qwen-7B obtained the best cumulative performance for all tasks); in green, the APSs of such models (as *perpl*).

Table 7: Performance comparison across Acceptability, Cloze test, and Multiple-choice tasks for few and zero-shot settings with Outlines.

OUTLINES										
Acceptability Rating	FEW					ZERO				
	All	Caus. Ic.	Temp. Ic.	Caus. a.Ic.	Temp. a.Ic.	All	Caus. Ic.	Temp. Ic.	Caus. a.Ic.	Temp. a.Ic.
GPT4o-mini	0.64	0.81	0.63	0.68	0.36	0.47	0.4	0.22	0.63	0.71
GPT4o	0.78	0.84	0.8	0.74	0.75	0.77	0.86	0.7	0.76	0.76
Falcon	0.26	0.25	0.2	0.29	0.32	0.26	0.2	0.22	0.35	0.31
Gemma	0.33	0.38	0.15	0.33	0.58	0.57	0.69	0.48	0.6	0.5
Llama	0.24	0.24	0.23	0.25	0.24	0.26	0.27	0.22	0.27	0.31
Mistral	0.19	0.05	0.03	0.22	0.55	0.22	0.18	0.11	0.15	0.56
Qwen	0.23	0.2	0.14	0.27	0.36	0.28	0.26	0.21	0.32	0.35
Cloze test	FEW					ZERO				
	All	Caus. Ic.	Temp. Ic.	Caus. a.Ic.	Temp. a.Ic.	All	Caus. Ic.	Temp. Ic.	Caus. a.Ic.	Temp. a.Ic.
GPT4o-mini	0.54	0.23	0.46	0.78	0.71	0.53	0.19	0.5	0.76	0.67
GPT4o	0.7	0.44	0.79	0.82	0.72	0.54	0.14	0.48	0.83	0.76
Falcon	0.24	0.28	0.33	0.14	0.17	0.29	0.38	0.36	0.15	0.23
Gemma	0.52	0.46	0.62	0.74	0.16	0.53	0.56	0.63	0.77	0.03
Llama	0.26	0.1	0.2	0.46	0.29	0.28	0.12	0.25	0.52	0.23
Mistral	0.43	0.05	0.47	0.65	0.55	0.36	0.18	0.25	0.45	0.63
Qwen	0.29	0.07	0.78	0.07	0.09	0.31	0.04	0.88	0.01	0.12
Multiple-chioce	FEW					ZERO				
	All	Caus. Ic.	Temp. Ic.	Caus. a.Ic.	Temp. a.Ic.	All	Caus. Ic.	Temp. Ic.	Caus. a.Ic.	Temp. a.Ic.
GPT4o-mini	0.32	0.31	0.3	0.45	0.21	0.29	0.29	0.22	0.4	0.27
GPT4o	0.67	0.49	0.82	0.89	0.36	0.63	0.51	0.77	0.81	0.3
Falcon	0.24	0.27	0.23	0.26	0.2	0.27	0.26	0.28	0.31	0.24
Gemma	0.51	0.2	0.67	0.83	0.2	0.5	0.23	0.72	0.77	0.1
Llama	0.26	0.25	0.23	0.28	0.29	0.26	0.29	0.27	0.25	0.21
Mistral	0.41	0.24	0.59	0.57	0.12	0.4	0.2	0.55	0.45	0.34
Qwen	0.28	0.32	0.25	0.3	0.25	0.28	0.28	0.31	0.29	0.23

Table 8: APS results for all the open models. The table shows results by relation ORDER (ICONIC as Ic. vs ANTI-ICONIC as a.Ic) and TYPE (CAUSAL as C. vs TEMPORAL as T.). **In bold the best-averaged result**, the overall result of the best model is underlined.

Model	Overall	C. Ic.	T. Ic.	C. a.Ic.	T. a.Ic.
Falcon	0.66	0.85	0.66	0.80	0.23
Gemma	0.62	0.93	0.69	0.60	0.15
Llama	0.65	0.93	0.74	0.70	0.12
Mistral	0.65	0.89	0.68	0.75	0.15
Qwen	0.59	0.83	0.53	0.65	0.32
Avg.	0.63	**0.89**	0.66	0.70	0.19

(Tab. 10) shows that, on the items in each frequency bin, models perform about the same way, i.e., tasks results have, with slight differences, the same distribution.

We also tested the frequency effect in models of increasing size. In Fig. 11 (Tab. 11), as in the previous case, the distribution of Qwen's results reached by models of different sizes is about the same.

H Error Analysis

Other than what is described in Sec. 5.3, from the confusion matrixes in Fig. 12, we observed that GPT tends to consider TEMPORAL RELATIONS as CAUSAL ones more often. However, most of the mistakes are within the same ORDER (i.e., ANTI-ICONIC), whereas, open models tend to make more mistakes under both TYPE and ORDER of the relation. Gemma, in the acceptability rating task, interprets CAUSAL relations as TEMPORAL, but it tends to confuse also the ORDER of the TEMPORAL ones. Differently, Falcon is inclined to interpret all the items as TEMPORAL in ANTI-ICONIC ORDER. We also can observe that Falcon and Gemma errors in a different manner according to the way the PCD task is modeled. These results further underline a discrepancy between models' internal representation and prompted knowledge, not only across different models but also referring to the same one.

I Correlation & Distribution

In Fig. 5, Sec. 5.3, we see that for both humans and models, the acceptability ratings are lower for the unrelated sentence pairs, showing that models appear to understand this difference. For the CAUSAL-TEMPORAL subset, in the case of normalized perplexities for Falcon, all values are heavily skewed toward the lower end of the spectrum. On the contrary, ratings from Gemma are gener-

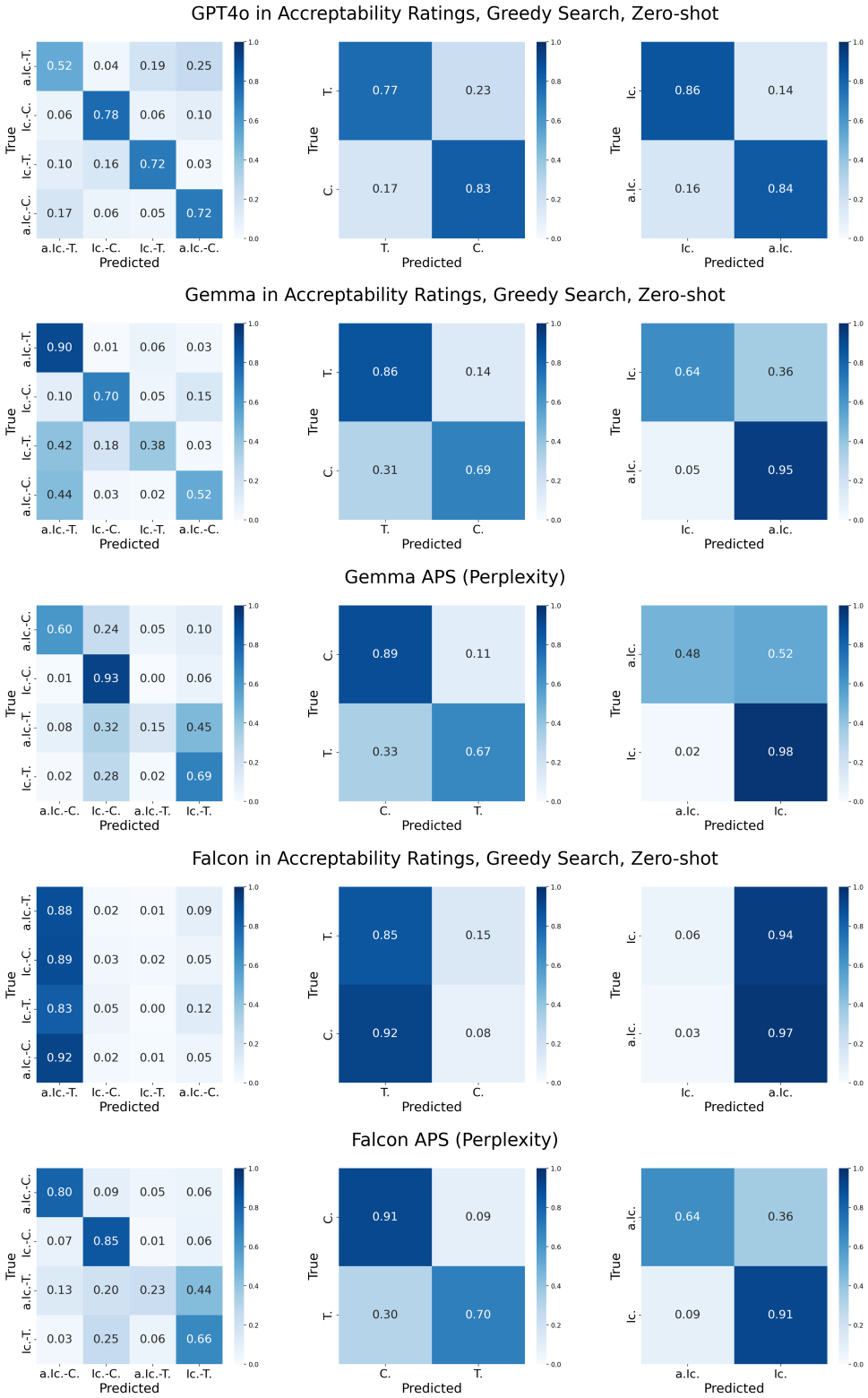

Figure 12: Comparison of different plots

Table 9: Performance metrics for the best models in the best settings and tasks for PCD: zero-shot with greedy search on acceptability rating task and APS. Results are reported with an overview over items' RELATION and ORDER, i.e., direction of the relation.

Model	Overall		Causal Iconic		Temporal Iconic		Causal Anti-iconic		Temporal Anti-iconic	
	Acc. Rat.	APS	Acc. Rat.	APS	Acc. Rat.	APS	Acc. Rat.	APS	Acc. Rat.	APS
Falcon	0.19	0.66	0.03	0.85	0.00	0.66	0.05	0.80	0.88	0.23
GPT-4o	0.69	-	0.78	-	0.72	-	0.72	-	0.52	-
Gemma	0.59	0.62	0.70	0.93	0.38	0.69	0.52	0.60	0.90	0.15

Table 10: LLMs' accuracy on the acceptability rating task and APS across different frequency bins.

Model	Overall		Rare		Uncommon		Common		Frequent	
	Acc. Rat.	APS	Acc. Rat.	APS	Acc. Rat.	APS	Acc. Rat.	APS	Acc. Rat.	APS
GPT-4o Mini	0.63	-	0.62	-	0.62	-	0.64	-	0.67	-
GPT-4o	0.69	-	0.69	-	0.69	-	0.70	-	0.70	-
Falcon	0.19	0.66	0.19	0.65	0.17	0.65	0.20	0.64	0.21	0.69
Gemma	0.59	0.62	0.56	0.59	0.61	0.64	0.57	0.61	0.63	0.64
Llama	0.52	0.65	0.50	0.66	0.52	0.64	0.55	0.63	0.48	0.70
Mistral	0.29	0.65	0.28	0.64	0.30	0.66	0.30	0.64	0.27	0.65
Qwen	0.54	0.59	0.51	0.55	0.58	0.67	0.54	0.58	0.54	0.56

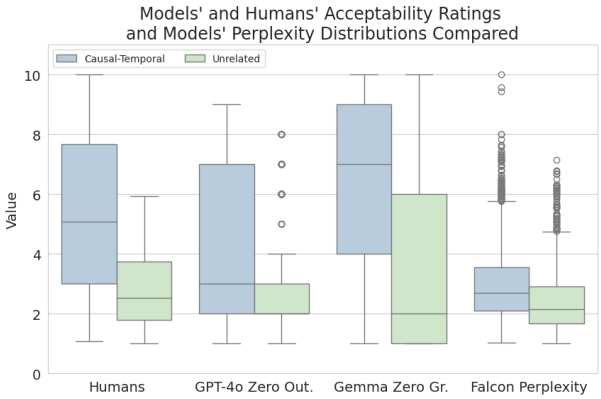

Figure 13: This plot shows the

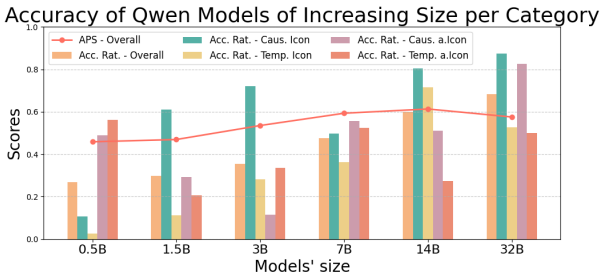

Figure 14: This plot shows the

ally distributed towards higher values than the human ratings, whereas GPT4o's ratings tend to be lower. GPT4o shows this behavior in zero-shot with greedy search, but also in zero-shot with Outlines as shown in Fig. 13, even if the model's ratings strongly correlate with humans' ones. Tab. 12 shows the correlation of open models' perplexity with results in acceptability ratings with greedy search and Oulines. In none of the setups, the accuracy of the internal representation of the models and the prompting accuracy show some kind of correlation.

Correlation scores for all the models (in zero-shot setting) are reported in Tab. 13. We see that GPT4o obtains a high correlation with humans only when using Outlines ($\rho = 0.77$), while the zero-shot greedy setting is markedly worse. We observe an opposite trend for GPT4o-mini and open models, which seem to be closer to human ratings when prompted without using Outlines. The best-performing open model is Gemma, on par with GPT4o-mini, followed by Qwen. All other models have a correlation lower than 0.5. Gemma is also the only model that is relatively resistant to the decoding method, i.e. greedy or with Outlines. Tab. 13 also contains correlation values computed on results obtained with prompting task according to the different relation TYPE and ORDER of the events in each dataset item.

J Size effect

Fig. 7 in Sec. 5.3 shows how Qwen models' of different sizes perform according to two evaluation settings, accuracy on acceptability rating and APS. In Fig. 14 (Tab. 14 for the individual results), the overall accuracy on the acceptability rating task is divided into groups related to the relation condition of each item, according to the ground truth we obtained by human ratings. We observed that the same pattern as APS is also followed by TEMPORAL ICONIC relations, the most frequent condition in our dataset according to human ratings. TEM-

Table 11: Performance of Qwen models with increasing size across different frequency bins. Accuracy for the acceptability rating task and APSs are reported.

Model	Overall		Rare		Uncommon		Common		Frequent	
	Acc. Rat.	AP. S.	Acc. Rat.	AP. S.	Acc. Rat.	AP. S.	Acc. Rat.	AP. S.	Acc. Rat.	AP. S.
Qwen-0.5	0.27	0.46	0.28	0.47	0.23	0.46	0.27	0.46	0.30	0.44
Qwen-1.5	0.30	0.47	0.34	0.50	0.29	0.52	0.26	0.42	0.28	0.43
Qwen-3	0.35	0.54	0.34	0.51	0.37	0.56	0.36	0.54	0.35	0.53
Qwen-7	0.48	0.59	0.49	0.55	0.49	0.67	0.48	0.58	0.43	0.56
Qwen-14	0.60	0.61	0.60	0.63	0.59	0.63	0.60	0.57	0.60	0.62
Qwen-32	0.68	0.58	0.66	0.57	0.70	0.57	0.69	0.58	0.68	0.59

Table 12: Spearman correlation computed on models' perplexity and acceptability ratings in a zero-shot scenario, for greedy search and outlines. Results are on items of all conditions.

Model	Greedy search	Outlines
Falcon	-0.030 (0.039)	-0.062 (0)
Gemma	-0.115 (0)	-0.115 (0)
Llama	-0.236 (0)	-0.091 (0)
Mistral	-0.150 (0)	0.082 (0)
Qwen	-0.215 (0)	0.002 (0.904)

PORAL ICONIC relations are naturally the most frequent if we consider the chronological order of events in each narrative and the fact that a causal relation between events (most of the times) implies a temporal relation. Thus, the APS might rely more on the frequency of observation of a certain phenomenon in the dataset. On the contrary, the accuracy in the detection of relations in ANTI-ICONIC ORDER seems not to follow a clear pattern.

Table 13: Spearman correlation values between models' results and human ratings on causal, temporal, and unrelated sentence pairs (human ratings as ground truth). Models' results are computed on the basis of acceptability ratings in a zero-shot scenario, for greedy search and outlines, and perplexity.

Model	Caus. Ic.		Temp. Ic.		Caus. a.Ic.		Temp. a.Ic	
	Greedy	Outlines	Greedy	Outlines	Greedy	Outlines	Greedy	Outlines
GPT4o	0.60 (0.0)	0.82 (0.0)	0.57 (0.0)	0.80 (0.0)	0.53 (0.0)	0.76 (0.0)	0.29 (0.0)	0.69 (0.0)
GPTo-mini	0.65 (0.0)	0.41 (0.0)	0.71 (0.0)	0.41 (0.0)	0.60 (0.0)	0.49 (0.0)	0.48 (0.0)	0.30 (0.0)
Gemma	0.63 (0.0)	0.63 (0.0)	0.61 (0.0)	0.64 (0.0)	0.70 (0.0)	0.50 (0.0)	0.58 (0.0)	0.33 (0.0)
Qwen	0.44 (0.0)	0.07 (0.041)	0.55 (0.0)	0.04 (0.233)	0.69 (0.0)	0.06 (0.065)	0.57 (0.0)	0.03 (0.413)
Llama	0.48 (0.0)	0.04 (0.224)	0.62 (0.0)	0.06 (0.052)	0.43 (0.0)	0.01 (0.816)	0.24 (0.0)	0.05 (0.171)
Mistral	0.23 (0.0)	-0.12 (0.001)	0.23 (0.0)	-0.08 (0.013)	0.41 (0.0)	-0.18 (0.0)	0.30 (0.0)	-0.03 (0.425)
falcon	0.04 (0.2)	-0.02 (0.497)	-0.07 (0.035)	-0.02 (0.594)	0.04 (0.215)	-0.01 (0.691)	0.09 (0.023)	0.01 (0.784)

Model	Unrel.		All		Perplexity
	Greedy	Outlines	Greedy	Outlines	
GPT4o	0.23 (0.0)	0.59 (0.0)	0.46 (0.0)	0.77 (0.0)	-
GPT4o-mini	0.50 (0.0)	0.21 (0.0)	0.66 (0.0)	0.38 (0.0)	-
Gemma	0.44 (0.0)	0.38 (0.0)	0.65 (0.0)	0.55 (0.0)	-0.150 (0)
Qwen	0.41 (0.0)	-0.03 (0.312)	0.59 (0.0)	0.04 (0.009)	-0.282 (0)
Llama	0.32 (0.0)	0.04 (0.141)	0.47 (0.0)	0.05 (0.002)	-0.265 (0)
Mistral	0.35 (0.0)	0.09 (0.001)	0.42 (0.0)	-0.12 (0.0)	-0.273 (0)
falcon	0.05 (0.049)	0.06 (0.022)	0.05 (0.001)	0.00 (0.784)	-0.251 (0)

Table 14: Performance metrics for Qwen models of increasing size, measured in billions of parameters. Acc. Rat. refers to the accuracy rating task, and APS refers to Accuracy Perplexity Score.

Model's Size	Overall		Caus. Icon.		Temp. Icon.		Caus. a.Icon.		Temp. a.Icon.	
	Acc. Rat.	APS	Acc. Rat.	APS	Acc. Rat.	APS	Acc. Rat.	APS	Acc. Rat.	APS
0.5	0.27	0.46	0.11	0.87	0.03	0.43	0.49	0.42	0.56	0.03
1.5	0.30	0.47	0.61	0.94	0.11	0.40	0.29	0.44	0.21	0.02
3	0.35	0.54	0.72	0.85	0.28	0.54	0.11	0.60	0.34	0.05
7	0.48	0.59	0.50	0.83	0.36	0.53	0.56	0.65	0.52	0.32
14	0.60	0.61	0.80	0.88	0.72	0.58	0.51	0.72	0.27	0.20
32	0.68	0.58	0.87	0.87	0.53	0.52	0.83	0.67	0.50	0.16