Exploring Large Language Models' World Perception: A Multi-dimensional Evaluation through Data Distribution

Anonymous ACL submission

Abstract

In recent years, large language models (LLMs) have achieved remarkable success across diverse natural language processing tasks. Nevertheless, their understanding of core human experiences remains underexplored. Current benchmarks for LLM evaluation typically focus on a single aspect of linguistic understanding, thus failing to capture the full breadth of its abstract reasoning about the world. To address this gap, we propose a multidimensional paradigm to investigate the capacity of LLMs to perceive the world through temporal, spatial, sentimental, and causal aspects. We conduct extensive experiments by partitioning datasets according to different distributions and employing various prompting strategies. Our findings reveal significant differences and shortcomings in how LLMs handle temporal granularity, multi-hop spatial reasoning, subtle sentiments, and implicit causal relationships. While sophisticated prompting approaches can mitigate some of these limitations, substantial challenges persist in effectively capturing abstract human perception. We aspire that this work, which assesses LLMs from multiple perspectives of human understanding of the world, will guide more instructive research on the LLMs' perception or cognition. The data and code will be released soon.

1 Introduction

Large Language Models (LLMs) have made significant strides in advancing natural language processing (NLP) (Brown et al., 2020; Kojima et al., 2022; Zhao et al., 2024a; Chu et al., 2024a), showcasing impressive abilities in understanding and generating human-like text (Sicilia and Alikhani, 2022; Gao et al., 2023b; Minaee et al., 2024). However, their comprehension of fundamental human experiences—such as time, space, sentiment, and causality—remains largely underexplored. Maurice Merleau-Ponty, a renowned phenomenologist, highlighted the embodied nature of perception, asserting that our bodily and affective experiences are central to how we engage with the world (Merleau-Ponty et al., 2013). He argued that consciousness is deeply intertwined with physical existence, challenging the Cartesian dualism of mind and body. This perspective suggests that a deeper understanding of human perception requires considering the pivotal role of the body in shaping experience.

In recent years, research has started to investigate specific facets of LLMs' world perception. For example, studies have examined their understanding of sentimental scenarios through the framework of appraisal and coping theory, revealing that while LLMs' responses generally align with human patterns in sentimental appraisal and coping dynamics, they differ in their sensitivity to key appraisal dimensions (Yongsatianchot et al., 2023). Additionally, evaluations of their causal reasoning capabilities have uncovered challenges in handling complex causal structures and distinguishing between correlation and causation (Liu et al., 2025; Zhou et al., 2024). To further explore the understanding and cognition of the world in terms of LLMs, we need to comprehensively evaluate their perception in multiple dimensions, including the dimensions emphasized by Merleau-Ponty's phenomenological sense.

This study aims to evaluate the world perception of LLMs through a multi-dimensional framework that encompasses time, space, sentiment, and causality. We have elected two datasets for each dimension and annotated them with relevant features based on different data distributions for evaluation. To guide this assessment, we employ a variety of prompting techniques, including basic, Chain-of-Thought (CoT), few-shot, and few-shot CoT prompting. Few-shot prompting (Dai et al., 2022) involves providing the model with a few examples to help guide its responses, while CoT (Wei et al., 2022) prompting encourages the model to generate intermediate reasoning steps, thereby improving its problem-solving abilities.

The main contributions of this study are as follows. (1) We introduce a novel framework for evaluating LLMs' world perception across four critical dimensions: time, space, sentiment, and causality from the perspective of data distribution. (2) By employing a variety of prompting strategies, this study explores how different prompting methods influence the performance of LLMs across the four dimensions. (3) We reveal the strengths and limitations of current LLMs in handling various reasoning tasks, providing valuable insights for future LLM development and applications.

2 WorldInsight BENCH

2.1 Benchmark Design

WorldInsight BENCH is designed to assess the capacity of large language models to understand the world at the abstract level of human cognition and perception. Given the multifaceted nature of perceptual domains, we structure our evaluation into four critical dimensions: time, space, sentiment, and causality. Each of these dimensions is examined through two specialized datasets. Based on different data distributions, we analyze how LLM interprets and processes the world.

Temporal dimension focuses on the models' ability to understand and reason about the passage of time and the relationships between temporal events. Spatial dimension centers on the model's capacity to grasp and interpret spatial relationships. Sentiment recognition evaluates the model's understanding of human sentiments exposed to various scenes, and its ability to discern sentimental states, intensity, and the underlying psychological dynamics. Causal perception examines the models' ability to infer causal relationships, distinguish between correlation and causation, and reason in counterintuitive causal scenarios.

2.2 Challenges

Complex reasoning tasks in natural language processing mirror real-world cognitive challenges. They require not only language comprehension but also intricate logical inference, recognition of implicit relationships, and the integration of multidimensional information (Niu et al., 2024; Xiang and Wang, 2022; Wang et al., 2024).

Temporal Logic and Event Sequencing Analyzing temporal information involves understanding event ordering, duration, frequency, and typical time. This analysis requires managing several temporal relationships concurrently, inferring implicit logic, and constructing accurate event sequences (Dong et al., 2024). The challenge increases when multiple time frames or ambiguous temporal cues are involved.

Complex Spatial Relationship Inference Inferring spatial relationships entails identifying both direct and indirect cues that determine the relative positions of entities (Hu et al., 2024). This process becomes more difficult as the number of objects and the complexity of their arrangements grow.

Sentiment Analysis with Implicit Context Detecting sentiment in text demands sensitivity to subtle sentimental nuances, including sarcasm and implicit emotions (Wang and Luo, 2023). The task will be further complicated when texts convey mixed emotions or when broader situational factors exist in text (Zhang et al., 2024).

Complex Causal Relationship Analysis Understanding causal relations in text involves tracking multiple events and their interactions (Lyu et al., 2022), particularly when causal links are implied rather than explicitly stated. Moreover, Large models can be confused when reasoning about counterfactual scenarios.

2.3 Datasets

In face of the above challenges, we selected two datasets for each dimension. And every dataset is segmented into different data distributions. The appendix A provides specific examples.

2.3.1 Temporal Cognition

TempNLI (Thukral et al., 2021) contains time-related premise-hypothesis pairs annotated with logical labels: Entailment, Contradiction, and Neutral. It focuses on evaluating temporal reasoning across two primary dimensions, including time granularity and Language complexity.

MCTACO (Zhou et al., 2019) evaluates the models' reasoning ability from multiple temporal relationship types, comprising time frequency, order, duration, stationarity, and typical event time. It presents short contexts followed by temporal reasoning questions with multiple valid answers.

2.3.2 Spatial Intelligence

Multi-hop Space (Li et al., 2024) evaluates the models' capability in reasoning about complex spatial relationships through multiple steps. The dataset presents scenarios of increasing complexity, ranging from 1-hop to 10-hop, in which the model must determine the relative position between two objects based on a series of intermediate spatial relationships.

SpaceTrans (Comsa and Narayanan, 2023) aims to assess the capability of LLMs to process spatial transfer relations conveyed through spatial prepositions in diverse contexts, including physical, metaphorical, and mixed scenarios. The dataset specifically examines whether models can distinguish between cases where spatial transitivity holds (in physical scenarios) versus cases where it breaks down (in metaphorical or hybrid contexts). This helps evaluate LLMs' understanding of how spatial reasoning rules apply differently across contexts.

2.3.3 Sentimental Insight

Yelp-5 (Zhang et al., 2015) contains restaurant reviews labeled with sentimental intensity ratings from 0 to 4, where 0 indicates strong negative sentiment and 4 indicates strong positive sentiment. The reviews discuss various aspects of dining experiences, including food quality, service, ambiance, and value. This dataset enables assessment of models' ability to detect fine-grained sentimental expressions in long-form consumer feedback.

IronyEval (Van Hee et al., 2018) comprises social media posts labeled as either sarcastic or non-sarcastic. Each post is classified as "explicit" and "implicit" based on whether it contains overt sarcasm markers or contextual cues that suggest sarcasm. This dataset tests models' capability to identify both overt and subtle forms of sarcastic expression common in social media communication.

2.3.4 Causal Comprehension

ECI (Gao et al., 2023a) consists of sentences containing event pairs, where the model must identify whether one event causes another. The dataset is categorized into man-made causality and natural causality based on different types of causal features. Additionally, the textual distance between event entities within the context is classified into close-range and far-range.

FantasyR (Srivastava et al., 2023) presents scenarios involving fictional elements like magic, supernatural beings, and fantastical situations, and is categorized based on the explicitness of causal relationships depicted in the text. It tests whether LLMs can maintain causal coherence and apply consistent logic within hypothetical worlds.

2.4 Evaluation Metrics

In this work, we utilize a range of evaluation metrics to assess the performance of LLMs on chosen tasks. The evaluation metrics include accuracy, F1-score, exact match, tolerant accuracy, etc. However, due to space limitations, we only report the accuracy in the main body, while the detailed scores for other metrics are provided in the Appendix B.

3 Approaches

3.1 Model Setup and Implementation

We evaluate a range of widely used LLMs, encompassing both open-source and proprietary models. The open-source models included in this evaluation range from the Llama 2 series to Llama 3.3 (Touvron et al., 2023; Grattafiori et al., 2024), with parameter sizes varying from 8B to 70B. Additionally, the proprietary GPT-4o model is also assessed.

The open-source models (Llama 2, Llama 3, Llama 3.1 and Llama3.3) are deployed locally across 8 x NVIDIA A800 80GB PCIe, while the GPT-4o model is accessed via API. For all experiments, we configure the temperature to 0.0 to enforce greedy decoding (Prabhu, 2024).

3.2 Evaluation Methods

In this study, we evaluate the LLMs using four distinct prompting strategies: Basic prompting, Chain of Thought (CoT) prompting, and their combination with Few-Shot setting. The aim is to investigate the competence of LLMs to understand the world in an abstract dimension, and whether different prompting methods can enhance their relevant reasoning.

Basic Prompting, also denoted as zero-shot (ZS), provide the model with specific instructions for each task. And in the few-shot (FS) setting, the model receives several QA pairs as demonstrations to guide the responses to new questions. The prompts P can be formulated as follows

$$\mathrm{P}_{\mathrm{ZS}} = \{\mathrm{INST}\} \oplus \{Q\} \tag{1}$$

$$\mathrm{P}_{\mathrm{FS}} = \{\mathrm{INST}\} \bigoplus_{i=1}^{n} (\{Q_i\} \oplus \{A_i\}) \oplus \{Q\} \tag{2}$$

where INST, Q, A represent the instruction, question, and answer, respectively. And i is the index of instance.

CoT Prompting builds on standard prompting by adding guidance for reasoning steps. In specific, we append a reasoning trigger "Let's think step by

Method	Temporal		Spatial		Sentimental		Causal		Overall Score				
	TempNLI	MCTACO	M-h Space	SpaceT	Yelp-5	IronyEval	ECI	FantasyR	Temp.	Spat.	emot.	Causal	Avg.
GPT-4o	63.50	53.75	48.75	88.25	61.50	79.00	35.25	80.00	58.63	68.50	70.25	57.63	63.75
+COT	70.25	60.00	42.50	89.50	59.25	77.50	59.00	81.00	65.13	66.00	68.38	70.00	67.38
+FS	70.25	57.25	46.75	89.25	**63.50**	**90.25**	64.75	81.00	63.75	68.00	76.88	72.88	70.38
+FS CoT	**70.75**	**74.50**	**52.75**	**92.00**	60.25	81.75	**66.50**	**91.50**	**72.63**	**72.38**	**71.00**	**79.00**	**73.75**
Llama-3.3-70b	53.50	54.75	36.00	82.50	57.75	74.00	58.50	75.50	54.13	59.25	65.88	67.00	61.56
+COT	70.00	63.25	48.25	87.25	**58.00**	76.25	54.25	80.00	66.63	67.75	67.13	67.13	67.16
+FS	71.25	58.50	**54.75**	85.75	57.50	**82.25**	31.75	79.50	64.88	**70.25**	**69.88**	55.63	65.16
+FS CoT	**74.50**	**72.75**	45.00	**88.75**	55.75	78.50	**59.50**	**83.00**	**73.63**	66.88	67.13	**71.25**	**69.72**
Llama-3.1-70b	50.50	49.25	38.00	86.25	**58.25**	73.75	43.75	78.50	49.88	62.13	66.00	61.13	59.78
+COT	64.50	57.50	44.00	87.50	52.75	72.50	55.50	76.00	61.00	65.75	62.63	65.75	63.78
+FS	63.00	44.75	**50.00**	87.50	56.50	**83.00**	55.75	**84.00**	53.88	68.75	**69.75**	69.88	65.56
+FS CoT	**72.00**	**66.50**	44.00	**91.75**	53.50	78.50	**68.00**	82.00	**69.25**	**67.88**	66.00	**75.00**	**69.53**
Llama-3-70b	50.25	33.25	25.25	79.75	55.00	72.50	**70.25**	63.00	41.75	52.50	63.75	66.63	56.16
+COT	48.25	31.25	31.75	85.25	57.75	73.75	49.75	76.50	39.75	58.50	65.75	63.13	56.78
+FS	51.75	**48.75**	**40.25**	83.00	**59.50**	**81.00**	28.75	76.00	50.25	**61.63**	**70.25**	52.38	58.63
+FS CoT	**70.75**	47.00	28.25	**89.00**	56.25	79.50	56.50	**77.00**	**58.88**	58.63	67.88	**66.75**	**63.03**
Llama-3-8b	46.25	37.75	**23.25**	71.50	46.25	59.75	**71.00**	70.50	42.00	47.38	53.00	**70.75**	53.28
+COT	41.00	18.25	15.50	75.00	50.75	56.75	47.25	70.50	29.63	45.25	53.75	58.88	46.88
+FS	50.00	**41.50**	20.25	70.50	51.75	73.75	38.75	61.50	**45.75**	45.38	62.75	50.13	51.00
+FS CoT	**50.75**	28.50	22.75	**84.00**	**57.50**	**77.50**	46.75	**74.00**	39.63	**53.38**	**67.50**	60.38	**55.22**
Llama-2-70b	45.50	**24.50**	22.75	65.25	29.50	61.50	19.00	61.50	**35.00**	44.00	45.50	40.25	41.19
+COT	47.25	19.25	**25.25**	76.00	**59.50**	52.00	**45.75**	**75.00**	33.25	50.63	55.75	**60.38**	50.00
+FS	**48.50**	14.25	21.00	63.25	50.25	**70.00**	21.50	64.00	31.38	42.13	60.13	42.75	44.09
+FS CoT	45.75	23.00	24.25	**85.50**	58.50	69.50	38.75	73.00	34.38	**54.88**	**64.00**	55.88	**52.28**
Llama-2-13b	**49.50**	7.75	9.00	51.50	47.25	42.00	31.75	**66.50**	28.63	30.25	44.63	49.13	38.16
+COT	47.00	13.25	17.75	**75.00**	39.50	49.50	**38.75**	64.50	30.13	46.38	44.50	**51.63**	43.16
+FS	44.25	**15.50**	12.50	57.25	33.00	57.75	21.25	**66.50**	29.88	34.88	45.38	43.88	38.50
+FS CoT	49.00	15.00	**23.50**	71.25	**60.50**	**71.50**	37.75	60.50	**32.00**	**47.38**	**66.00**	49.13	**48.63**

Table 1: Main experimental results over 8 datasets. All models are alignment mdoels (-chat or -instruct). Accuracy is reported here, and additional evaluation metrics can be found in Appendix B.

step" to encourage the model to break down the problem into logical steps before providing an answer. In the few-shot CoT setting, we also provide demonstrations with CoT to guide the reasoning process. The prompt formulations are as follows

$$\mathrm{P}_{\mathrm{CoT}} = \{\mathrm{INST}\} \oplus \{Q\} \oplus \{\mathrm{TRIG}\} \quad (3)$$

$$\mathrm{P}_{\mathrm{CoT \cdot FS}} = \{\mathrm{INST}\} \bigoplus_{i=1}^{n} (\{Q_i\} \oplus \{R_i\} \oplus \{A_i\}) \oplus \{Q\} \quad (4)$$

where TRIG denotes the reasoning trigger and R represents the reasoning examples.

4 Experimental Results

4.1 Zero-shot Results

Our evaluation of LLMs on the four dimensions of abstract reasoning, covering time, space, sentiment, and causality, revealed significant performance differences (Table 1). In the zero-shot setting, GPT-4o achieved the highest overall average score (63.8%), outperforming all open-source models across every dimension. This superior performance is likely due to its training on large-scale data, which enables it to capture complex patterns and implicit structures across diverse domains. However, in causal reasoning, GPT-4o underperformed relative to most models in the Llama series. This is possibly because of its focus on lexical co-occurrence and syntactic structures, rather than understanding the causal nature of events.

Open-source models generally excelled in sentimental and causal reasoning tasks but struggled with temporal and spatial inference. Spatial reasoning showed the greatest variability among models, with GPT-4o averaging 68.5% versus Llama-2-13b's 30.3%. This disparity likely reflects the advantage of more advanced models that benefit from larger, more diverse training sets, which facilitate the learning of finer, more abstract spatiotemporal relationships.

4.2 The Impact of CoT Prompting

CoT prompting yields performance improvements. However, it is highly dependent on both the specific model and the type of reasoning task. In temporal reasoning, CoT prompting significantly boosts the performance of larger, more advanced models like GPT-4o (6.5%↑) and particularly Llama-3.3-70b (12.5%↑). Conversely, older or smaller models such as Llama-2 and Llama-3 showed minimal (1.5%↑) or even detrimental effects, suggesting they may not possess adequate autonomous reasoning capabilities. For spatial reasoning, Llama models generally benefited from CoT, with Llama-3.3 showing a notable 12.3% improvement, especially in multi-hop tasks where step-by-step reasoning proved advantageous. Sentimental reasoning and spatial reasoning exhibited mixed trends, with GPT-4o and Llama-3.1 showing performance declines in sentimental reasoning but improvements in spatial reasoning, underscoring the task-specific property of CoT's benefits.

4.3 Few-shot Setting and CoT Prompting

The utilization of few-shot has consistently enhanced performance. The average score of GPT-4o increases from 63.8% to 70.4%, while Llama-3.1-70b rises by 5.8%, and only the Llama-3-8b model shows a slight performance decline. For these abstract dimensions, the temporal, spatial, and sentimental reasoning capabilities of the LLMs are improved to varying degrees. Causal reasoning improvements are more pronounced in GPT-4o, but remains limitation across most Llama models. It suggests that GPT-4o shows exceptional potential in learning causal inference from instances in the few-shot scenario, whereas most Llama models still struggle to extract patterns of causal reasoning from examples.

Examples can strengthen and stabilize CoT reasoning. Combining few-shot with CoT yields the highest benefits, with the causal reasoning of GPT-4o jumping by 21.3%, and the sentimental reasoning of Llama-2-13B improving by 21.4%. Notably, few-shot CoT prompting mitigated the decline in reasoning capabilities caused by CoT in some models. This suggests that relying solely on CoT may lead to misleading results when the model lacks sufficient context. The addition of few-shot prompting provides more task-relevant information and guidance, helping the model understand diverse reasoning steps, avoiding over-reliance on single reasoning path, and thus enhancing the accuracy of causal reasoning.

5 Analysis and Discussion

We conduct a further analysis of the capacities of various LLMs to comprehend the world primarily through the lens of data distribution.

5.1 Evaluation on Temporal Inference

LLMs underperform in large temporal granularities, with the performance worsening even more at mixed granularities. As illustrated in Figure 1, LLMs generally show higher performance on small time scales (e.g., 9 a.m.) than on large time scales (e.g., after May 1939). This trend is attributed to the fact that the greater symbolic complexity involved in large time scales expressing introduces ambiguity and require more context to understand.

The capacity varies in different LLMs when dealing with different language complexities. Notably, GPT-4o, Llama-3.3, and Llama-3.1 exhibit superior performance on simple time expression tasks, whereas Llama-3 and Llama-2 demonstrate greater proficiency on compound or multiple time expression tasks. The observed performance disparity can arise from differences in the models' pre-training corpora, particularly in terms of their exposure to temporal expressions (Zhao et al., 2024b). Additionally, variations in model architecture, including the design of attention mechanisms that capture relationships across different positions within the input sequence, may also contribute to this discrepancy. Appendix D provides further experimental exploration based on this speculation.

Iterations have made the models show a steady improvement in handling event ordering issues. From llama2 to llama3.3, the model performance has continued to rise, which is exhibited in Figure 2. This is due to the inclusion of more diverse and complex data, along with optimized attention mechanisms and the resulting better contextual understanding (Harsha et al., 2024).

The model is limited in its ability to make autonomous choices, but few-shot and CoT can bring significant improvements. Due to the characteristics of typical time tasks, the model needs to autonomously select possible time nodes as the correct answer. In the zero-shot scenario, the performance of the LLMs is limited. Few-shot and CoT bring more examples or structured contexts to the models, which opens the models' ability to make autonomous choices.

5.2 Evaluation on Spatial Reasoning

Most models are not yet adequate for multi-hop spatial reasoning tasks involving complex relationships between multiple objects. In n-hop tasks (Figure 3), when $n > 4$, the average accuracy of LLMs is always below 30% under all methods. Although methods such as few-shot or CoT will bring some performance improvements when n is small, this improvement disappears when $n >= 6$. In addition, in 10-hop tasks, few-shot and CoT even become introduced noise and can no longer help LLMs summarize and process more complex spatial relationships.

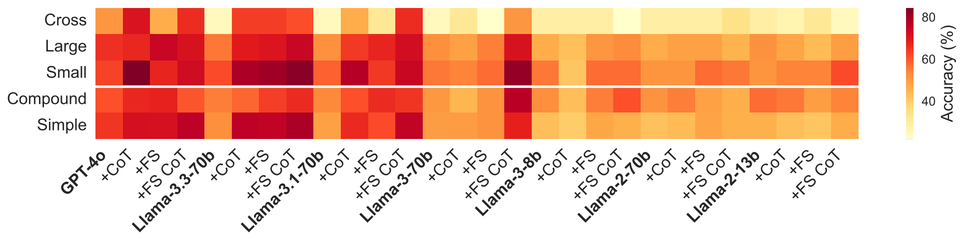


Figure 1: Performance of the LLMs on TempNLI. The dataset is divided into Large, Small and Cross-granularity according to the time granularity, and clasified into Simple and Compound based on the language complexity.

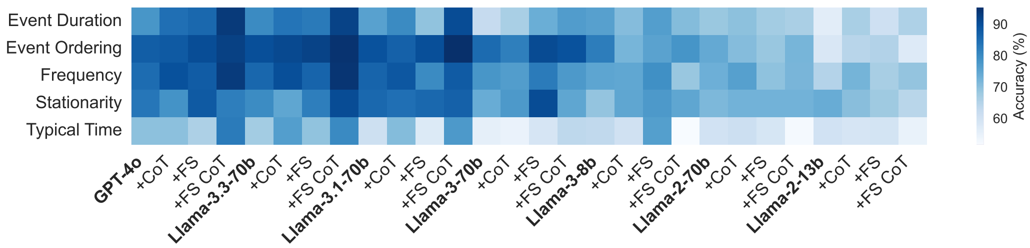


Figure 2: Performance of the LLMs on MCTACO. This dataset is grouped into Event Duration, Event Duration, Frequency, Stationarity and Typical Time.

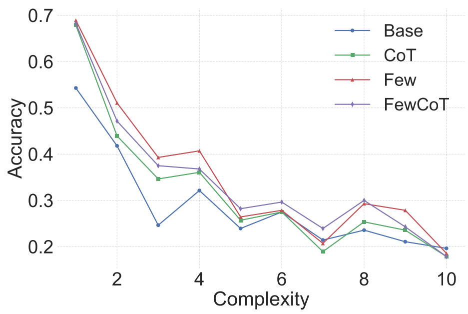


Figure 3: Average performance of all the LLMs on Multi-hop Space, ranging from 1-hop to 10-hop.

Metaphorical relations make it difficult for models to maintain consistent performance. Within the SpaceTrans task (Figure 4), LLMs generally perform well on physical spatial relations, achieving high accuracy in all prompting strategies. However, when it comes to metaphorical spatial prepositions, LLMs perform poorly. And the improvement brought by few-shot or CoT does not catch up with the former. On physical-metaphorical composite spatial relations, models like Llama-2-13b and Llama-2-70b show lower accuracy, indicating that the mixture of different types of semantic relations may confuse the model and negatively affect its performance.

Few-shot CoT prompting can significantly improve the performance of LLMs in processing composite spatial semantic relations. Although LLMs are not satisfactory in processing metaphors or physical-metaphor compound relations, the performance of LLMs can be greatly improved when using Few-shot CoT prompting. In particular, the improvement in physical-metaphor compound relations exceeds that of pure metaphorical relations. The phenomenon shows that although the complexity of the task increases with mixed relations, the models benefit from the additional context provided by the few-shot examples and their thought chains. This helps them improve the ability to distinguish between both physical and metaphorical relations, thereby better handling the related tasks.

5.3 Evaluation on Sentimental Reasoning

LLMs have the ability to judge the polarity of sentiment, but they are often erratic at a fine granularity. For most models, the dark colors of the confusion matrix are mainly on the diagonal, and confusion mainly occurs on adjacent grids. This demonstrates that LLMs can effectively judge the sentiment tendency of the text but will bring deviation to refined scoring. And CoT Few-shot (Figure 5) will even deepen the confusion in most models, indicating that LLMs still have difficulty learning fine-grained scoring criteria from examples.

LLMs encounter notable difficulties in detecting subtle implicit irony. As shown in Figure 6, the performance of LLMs on the explicit and implicit irony datasets reveals significant variations,

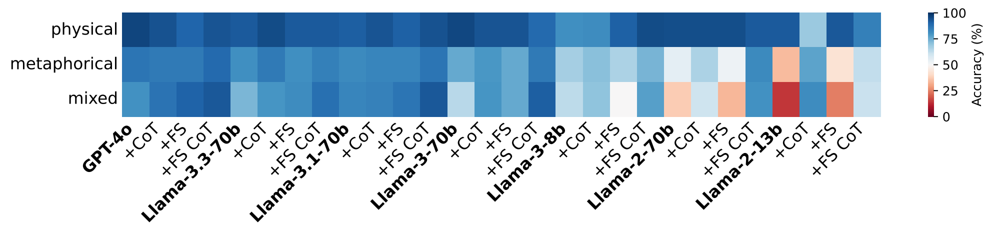


Figure 4: Performance of the LLMs on SpaceTrans, which is segmented into physical, metaphorical, and mixed scenarios.

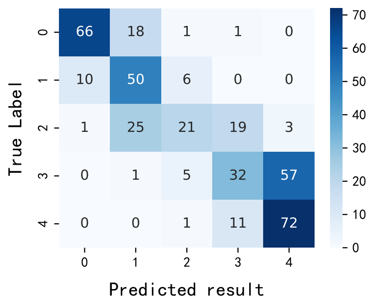


Figure 5: Confusion Matrix of GPT-4o in Yelp-5 utilizing CoT Few-shot prompting. The confusion matrices for all the models are demonstrated in Appendix C.

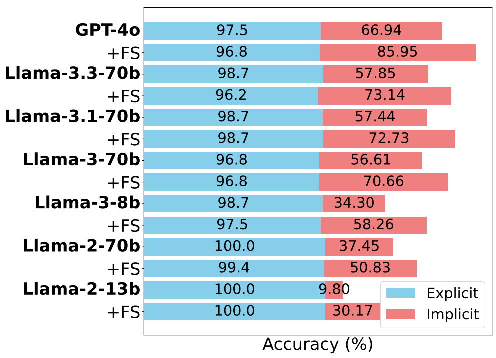


Figure 6: Performance on IronyEval, which is devided into explicit and implicit expressions.

Model	Event Type		Text Distance	
	Natural	Man-made	close	Far
GPT-4o	65.45	**66.67**	62.11	**72.25**
Llama-3.3-70b	**61.82**	59.13	56.83	**63.01**
Llama-3.1-70b	**70.91**	67.54	66.52	**69.94**
Llama-3-70b	49.09	**57.68**	52.86	**61.27**
Llama-3-8b	**50.91**	46.09	44.49	**49.71**
Llama-2-70b	38.18	**38.84**	36.56	**41.62**
Llama-2-13b	**40.00**	37.39	**39.21**	35.84

Table 2: Performance comparison of different models on ECI with few-shot and CoT setting.

with most models performing better on explicit irony, where clear markers are present. For instance, GPT-4o achieved 97.5% accuracy in detecting explicit irony, but the performance dropped to 66.9% for implicit irony. This performance gap suggest that while large language models are effective at identifying clear markers of irony, they struggle to discern more subtle, context-dependent instances of implicit irony.

5.4 Evaluation on Causal Reasoning

The LLMs have roughly equivalent causal identification ability for two categories of events. Table 2 suggests that GPT-4o and Llama demonstrate a similar level of accuracy in identifying causal relationships across different event categories, whether "natural" or "man-made." This indicates that the models can recognize and process causal events in both contexts without significant bias.

Current LLMs exhibit notable limitations in identifying causal relations within close textual distance. It is attributable to rapid context shifts and token proximity. This emphasizes the need for enhanced contextual awareness and improved disambiguation of closely related events (Joshi et al., 2024).

Most models can make accurate inferences in counterintuitive scenes. However, this doesn't conclude that the model is capable of human-like thinking, because the model may just replace the subjects or concepts based on the shortcut reasoning paradigms learnt (Du et al., 2023). Just as although few-shot CoT can bring an 11.5% improvement to GPT-4o, CoT and few-shot can only bring a 1% improvement when acting alone.

CoT and Few-shot have shown significant promise in eliminating the deviation of the model's causal reasoning ability between explicit and implicit data. From Llama-2 to Llama-3, CoT and few-shot settings each demonstrates different debiasing effects (Table 3). These approaches together contribute to a more balanced reasoning way, enabling the models to perform consistently across distinct causal reasoning tasks, thus reducing the performance discrepancies.

Method	GPT-4o	Llama-3.3-70b	Llama-3.1-70b	Llama-3-70b	Llama-3-8b	Llama-2-70b	Llama-2-13b
basic	8.79	-6.92	**-4.51**	-10.77	-8.02	-8.68	-3.19
CoT	3.74	-6.59	4.84	**-3.19**	-3.63	3.30	**2.53**
FS	-0.66	-2.97	-4.84	4.84	**-2.09**	**1.76**	-3.19
FS CoT	**0.11**	**0.22**	-5.71	-4.62	-11.43	-1.98	7.36

Table 3: The difference in model accuracy between the explicit and implicit data. Applying different prompting methods has a significant effect in helping the model eliminate explicit and implicit biases in FantasyR. The smallest absolute value of the bias for each model is marked in bold.

5.5 Summary of Findings

LLMs exhibit glaring deficiencies in processing large and mixed temporal granularities, complex linguistic phenomena, and metaphorical relations, exposing critical limitations in current generative models. While iterative improvements enhance event ordering and causal reasoning, many models still falter in multi-hop spatial reasoning, detecting subtle irony, and fine-grained sentiment analysis. Few-shot and chain-of-thought prompting significantly boost performance in autonomous decision-making, mixed spatial semantic processing, and aligning explicit and implicit causal reasoning, highlighting promising directions for future development.

6 Related Work

Recent research has increasingly focused on exploring the intersections between LLMs and human cognitive processes. Cognitive psychology techniques reveal that, although task-specific estimates from LLMs can sometimes align with human behavior, these models exhibit substantial variability across tasks (Niu et al., 2024; Chu et al., 2024b; Suresh et al., 2023), and their inductive reasoning—exemplified by GPT-3 and ChatGPT—differs markedly from human patterns (Lamprinidis, 2024). These findings highlight both the promise and limitations of LLMs as cognitive models, indicating a need for further research.

Temporal reasoning has been explored via graph-based paradigms that use synthetic datasets and CoT symbolic reasoning (Xiong et al., 2024; Yuan et al., 2024), as well as through synthetic and hierarchical benchmarks that reveal performance gaps between LLMs and human (Fatemi et al., 2024; Chu et al., 2024b). Moreover, knowledge induction frameworks have been applied to improve temporal QA, with dedicated QA datasets and prompt engineering strategies addressing specific vulnerabilities (Wei et al., 2023; Chen et al., 2024).

Spatial reasoning investigations have shown that prefix-based prompts can enhance zero-shot performance on 3D trajectory tasks (Sharma, 2023), while studies in visual question answering and navigation highlight performance variability and ethical concerns (Dugar and Asesh, 2023; Yamada et al., 2024). Qualitative assessments in commonsense spatial tasks and tic-tac-toe reveal further limitations, with chain-of-symbol prompting notably improving spatial planning (Cohn, 2023; Liga and Pasetto, 2023; Cohn and Hernandez-Orallo, 2023). Evaluations of sentimental understanding (Lei et al., 2024; Sun et al., 2023; Fei et al., 2023) indicate that LLMs generate appropriate yet not fully human-aligned responses (Huang et al., 2024; Wang et al., 2023; Li et al., 2023a; Balamurali et al., 2023), while studies in causal reasoning demonstrate accurate causal argument generation alongside persistent failure modes (Kıcıman et al., 2024; Jin et al., 2024; Vashishtha et al., 2023; Cai et al., 2024; Li et al., 2023b).

Distinguished from other works, our study examines the capacity of LLMs to comprehend the world from the perspective of data distribution, leveraging secondary annotations of comprehensive data.

7 Conclusion

Although large language models demonstrate exceptional language processing capabilities, they continue to face significant challenges in capturing complex human experiences. Variability in performance across time, space, sentiment, and causality indicates that even advanced models have limitations. Enhanced prompting methods, such as chain-of-thought and few-shot approaches, provide improvements but do not fully resolve these issues. These insights offer a clear direction for future research focused on strengthening abstract reasoning in language models.

Limitations

This work evaluates LLMs from multiple abstract perspectives of human perception of the world, relying on the selected datasets, which may not fully reflect the diversity of human perceptions of the world. Although prompting strategies can enhance performance, they do not address the inherent gaps in the model architecture and training data. Future research should investigate more diverse datasets and more comprehensive evaluation methods to gain deeper insights into how to strengthen the abstract reasoning capabilities of the models.

Ethics Statement

We do not foresee any immediate negative ethical consequences of our research.

References

Ommi Balamurali, A.M. Abhishek Sai, Moturi Karthikeya, and Sruthy Anand. 2023. Sentiment analysis for better user experience in tourism chatbot using LSTM and LLM. In *2023 9th International Conference on Signal Processing and Communication (ICSC)*, pages 456–462.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Hengrui Cai, Shengjie Liu, and Rui Song. 2024. Is knowledge all large language models needed for causal reasoning? *arXiv preprint*. ArXiv:2401.00139.

Ziyang Chen, Dongfang Li, Xiang Zhao, Baotian Hu, and Min Zhang. 2024. Temporal knowledge question answering via abstract reasoning induction. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 4872–4889, Bangkok, Thailand. Association for Computational Linguistics.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2024a. Navigate through enigmatic labyrinth a survey of chain of thought reasoning: Advances, frontiers and future. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 1173–1203, Bangkok, Thailand. Association for Computational Linguistics.

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Haotian Wang, Ming Liu, and Bing Qin. 2024b. TimeBench: A comprehensive evaluation of temporal reasoning abilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 1204–1228, Bangkok, Thailand. Association for Computational Linguistics.

Anthony G. Cohn. 2023. An evaluation of ChatGPT-4's qualitative spatial reasoning capabilities in RCC-8. *arXiv preprint*. ArXiv:2309.15577.

Anthony G. Cohn and Jose Hernandez-Orallo. 2023. Dialectical language model evaluation: An initial appraisal of the commonsense spatial reasoning abilities of LLMs. *arXiv preprint*. ArXiv:2304.11164.

Iulia Comsa and Srini Narayanan. 2023. A benchmark for reasoning with spatial prepositions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16328–16335, Singapore. Association for Computational Linguistics.

Zhuyun Dai, Vincent Y. Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B. Hall, and Ming-Wei Chang. 2022. Promptagator: Few-shot dense retrieval from 8 examples. *arXiv preprint*. ArXiv:2209.11755.

Xiangjue Dong, Maria Teleki, and James Caverlee. 2024. A survey on LLM inference-time self-improvement. *arXiv preprint*. ArXiv:2412.14352.

Mengnan Du, Fengxiang He, Na Zou, Dacheng Tao, and Xia Hu. 2023. Shortcut learning of large language models in natural language understanding. *arXiv preprint*. ArXiv:2208.11857.

Meenal Dugar and Aishwarya Asesh. 2023. Spatial interpretation and LLMs. In *2023 IEEE International Conference on Machine Learning and Applied Network Technologies (ICMLANT)*, pages 1–6.

Bahare Fatemi, Mehran Kazemi, Anton Tsitsulin, Karishma Malkan, Jinyeong Yim, John Palowitch, Sungyong Seo, Jonathan Halcrow, and Bryan Perozzi. 2024. Test of time: A benchmark for evaluating LLMs on temporal reasoning. *arXiv preprint*. ArXiv:2406.09170.

Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. 2023. Reasoning implicit sentiment with chain-of-thought prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 1171–1182, Toronto, Canada. Association for Computational Linguistics.

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

Mingqi Gao, Jie Ruan, Renliang Sun, Xunjian Yin, Shiping Yang, and Xiaojun Wan. 2023b. Human-like summarization evaluation with ChatGPT. *arXiv preprint*. ArXiv:2304.02554.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *arXiv preprint*. ArXiv:2407.21783.

Koppuravuri Harsha, Kanakam Tarun Kumar, D. Sumathi, and E. Ajith Jubilson. 2024. A survey on LLMs: Evolution, applications, and future frontiers. In Khalid Raza, Naeem Ahmad, and Deepak Singh, editors, *Generative AI: Current Trends and Applications*, pages 289–327. Springer Nature, Singapore.

Hanxu Hu, Hongyuan Lu, Huajian Zhang, Yun-Ze Song, Wai Lam, and Yue Zhang. 2024. Chain-of-symbol prompting elicits planning in large langauge models. *arXiv preprint*. ArXiv:2305.10276.

Jen-tse Huang, Man Ho Lam, Eric John Li, Shujie Ren, Wenxuan Wang, Wenxiang Jiao, Zhaopeng Tu, and Michael R. Lyu. 2024. Emotionally numb or empathetic? evaluating how LLMs feel using EmotionBench. *arXiv preprint*. ArXiv:2308.03656.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. 2024. CLadder: Assessing causal reasoning in language models. *arXiv preprint*. ArXiv:2312.04350.

Nitish Joshi, Abulhair Saparov, Yixin Wang, and He He. 2024. LLMs are prone to fallacies in causal inference. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10553–10569, Miami, Florida, USA. Association for Computational Linguistics.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, pages 22199–22213, Red Hook, NY, USA. Curran Associates Inc.

Emre Kıcıman, Robert Ness, Amit Sharma, and Chenhao Tan. 2024. Causal reasoning and large language models: Opening a new frontier for causality. *arXiv preprint*. ArXiv:2305.00050.

Sotiris Lamprinidis. 2024. LLM cognitive judgements differ from human. In *Frontiers of Artificial Intelligence, Ethics, and Multidisciplinary Applications*, pages 17–23, Singapore. Springer Nature.

Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, Runqi Qiao, and Sirui Wang. 2024. InstructERC: Reforming emotion recognition in conversation with multi-task retrieval-augmented large language models. *arXiv preprint*. ArXiv:2309.11911.

Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. 2023a. Large language models understand and can be enhanced by emotional stimuli. *arXiv preprint*. ArXiv:2307.11760.

Fangjun Li, David C. Hogg, and Anthony G. Cohn. 2024. Advancing spatial reasoning in large language models: An in-depth evaluation and enhancement using the StepGame benchmark. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):18500–18507. Number: 17.

Shun-Hang Li, Gang Zhou, Zhi-Bo Li, Ji-Cang Lu, and Ning-Bo Huang. 2023b. The causal reasoning ability of open large language model: A comprehensive and exemplary functional testing. In *2023 IEEE 23rd International Conference on Software Quality, Reliability, and Security (QRS)*, pages 240–249.

Davide Liga and Luca Pasetto. 2023. Testing spatial reasoning of large language models: The case of tic-tac-toe. In *AIxPAC*.

Xiaoyu Liu, Paiheng Xu, Junda Wu, Jiaxin Yuan, Yifan Yang, Yuhang Zhou, Fuxiao Liu, Tianrui Guan, Haoliang Wang, Tong Yu, Julian McAuley, Wei Ai, and Furong Huang. 2025. Large Language Models and Causal Inference in Collaboration: A Comprehensive Survey. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7668–7684, Albuquerque, New Mexico. Association for Computational Linguistics.

Zhiheng Lyu, Zhijing Jin, Rada Mihalcea, Mrinmaya Sachan, and Bernhard Schölkopf. 2022. Can large language models distinguish cause from effect?

Maurice Merleau-Ponty, Donald Landes, Taylor Carman, and Claude Lefort. 2013. *Phenomenology of perception*. Routledge.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *arXiv preprint*. ArXiv:2402.06196.

Qian Niu, Junyu Liu, Ziqian Bi, Pohsun Feng, Benji Peng, Keyu Chen, Ming Li, Lawrence KQ Yan, Yichao Zhang, Caitlyn Heqi Yin, Cheng Fei, Tianyang Wang, Yunze Wang, Silin Chen, and Ming Liu. 2024. Large language models and cognitive science: A comprehensive review of similarities, differences, and challenges. *arXiv preprint*. ArXiv:2409.02387.

Sumanth Prabhu. 2024. PEDAL: Enhancing greedy decoding with large language models using diverse exemplars. *arXiv preprint*. ArXiv:2408.08869.

Manasi Sharma. 2023. Exploring and improving the spatial reasoning abilities of large language models. *arXiv preprint*. ArXiv:2312.01054.

Anthony Sicilia and Malihe Alikhani. 2022. LEATHER: A framework for learning to generate human-like text in dialogue. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 30–53, Online only. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, and 431 others. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*, 2023(5):1–95.

Xiaofei Sun, Xiaoya Li, Shengyu Zhang, Shuhe Wang, Fei Wu, Jiwei Li, Tianwei Zhang, and Guoyin Wang. 2023. Sentiment analysis through LLM negotiations. *arXiv preprint*. ArXiv:2311.01876.

Siddharth Suresh, Kushin Mukherjee, Xizheng Yu, Wei-Chun Huang, Lisa Padua, and Timothy Rogers. 2023. Conceptual structure coheres in human cognition but not in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 722–738, Singapore. Association for Computational Linguistics.

Shivin Thukral, Kunal Kukreja, and Christian Kavouras. 2021. Probing language models for understanding of temporal expressions. In *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 396–406, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint*. ArXiv:2307.09288.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. SemEval-2018 task 3: Irony detection in english tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 39–50, New Orleans, Louisiana. Association for Computational Linguistics.

Aniket Vashishtha, Abbavaram Gowtham Reddy, Abhinav Kumar, Saketh Bachu, Vineeth N. Balasubramanian, and Amit Sharma. 2023. Causal inference using LLM-guided discovery. *arXiv preprint*. ArXiv:2310.15117.

Xuena Wang, Xueting Li, Zi Yin, Yue Wu, and Jia Liu. 2023. Emotional intelligence of large language models. *Journal of Pacific Rim Psychology*, 17:18344909231213958. Publisher: SAGE Publications.

Yajing Wang and Zongwei Luo. 2023. Enhance multi-domain sentiment analysis of review texts through prompting strategies. In *2023 International Conference on High Performance Big Data and Intelligent Systems (HDIS)*, pages 1–7.

Yiqi Wang, Wentao Chen, Xiaotian Han, Xudong Lin, Haiteng Zhao, Yongfei Liu, Bohan Zhai, Jianbo Yuan, Quanzeng You, and Hongxia Yang. 2024. Exploring the reasoning abilities of multimodal large language models (MLLMs): A comprehensive survey on emerging trends in multimodal reasoning. *arXiv preprint*. ArXiv:2401.06805.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, pages 24824–24837, Red Hook, NY, USA. Curran Associates Inc.

Yifan Wei, Yisong Su, Huanhuan Ma, Xiaoyan Yu, Fangyu Lei, Yuanzhe Zhang, Jun Zhao, and Kang Liu. 2023. MenatQA: A new dataset for testing the temporal comprehension and reasoning abilities of large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1434–1447, Singapore. Association for Computational Linguistics.

Wei Xiang and Bang Wang. 2022. A survey of implicit discourse relation recognition. *arXiv preprint*. ArXiv:2203.02982.

Siheng Xiong, Ali Payani, Ramana Kompella, and Faramarz Fekri. 2024. Large language models can learn temporal reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 10452–10470, Bangkok, Thailand. Association for Computational Linguistics.

Yutaro Yamada, Yihan Bao, Andrew K. Lampinen, Jungo Kasai, and Ilker Yildirim. 2024. Evaluating spatial understanding of large language models. *arXiv preprint*. ArXiv:2310.14540.

Nutchanon Yongsatianchot, Parisa Ghanad Torshizi, and Stacy Marsella. 2023. Investigating large language models' perception of emotion using appraisal theory. In *2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 1–8.

Chenhan Yuan, Qianqian Xie, Jimin Huang, and Sophia Ananiadou. 2024. Back to the future: Towards explainable temporal reasoning with large language models. In *Proceedings of the ACM Web Conference 2024*, WWW '24, pages 1963–1974, New York, NY, USA. Association for Computing Machinery.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. 2024. Sentiment analysis in the era of large language models: A reality check. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3881–3906, Mexico City, Mexico. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, and 3 others. 2024a. A survey of large language models. *arXiv preprint*. ArXiv:2303.18223.

Yang Zhao, Li Du, Xiao Ding, Kai Xiong, Zhouhao Sun, Shi Jun, Ting Liu, and Bing Qin. 2024b. Deciphering the impact of pretraining data on large language models through machine unlearning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9386–9406, Bangkok, Thailand. Association for Computational Linguistics.

Ben Zhou, Daniel Khashabi, Qiang Ning, and Dan Roth. 2019. "going on a vacation" takes longer than "going for a walk": A study of temporal commonsense understanding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3363–3369, Hong Kong, China. Association for Computational Linguistics.

Yu Zhou, Xingyu Wu, Beicheng Huang, Jibin Wu, Liang Feng, and Kay Chen Tan. 2024. CausalBench: A comprehensive benchmark for causal learning capability of LLMs. *arXiv preprint*. ArXiv:2404.06349.

A Dataset Instances

Examples from the datasets employed in this study are presented in Figure 7.

B Full Results

This study evaluates model performance across eight datasets, each using specific scoring metrics to assess different aspects of effectiveness. For the TempNLI, SpaceTrans, and IronyEval datasets, accuracy (Acc) is used. The MCTACO, Yelp-5, and ECI datasets are evaluated with exact match (EM), F1 score, and tolerant accuracy (ToAcc). The FantasyR dataset includes Acc along with implicit (Acc-i) and explicit (Acc-e) accuracy variants to capture nuanced performance. The full experimental results can be found in Table 4.

Here we explain the evaluation index ToAcc. For the MCTACO dataset, the default evaluation metrics employ a strict matching criterion, awarding a score of 1 for an exact correspondence between the prediction and the ground truth label, and 0 otherwise. To accommodate instances of partial correctness, we introduce a tolerant scoring mechanism. For example, a prediction of "right" or "below" would receive a predefined partial score when the ground truth label is "lower-right". This is achieved through a scoring matrix M, where scoring coefficients are explicitly defined for each prediction-label pair.

The tolerant score ToAcc, denoted as $S(l_{true}, l_{pred})$, for a true label l_{true} and predicted label l_{pred} is given by

$$S(l_{true}, l_{pred}) = M_{ij} \tag{5}$$

where i and j are the indices of l_{true} and l_{pred} in M, respectively. The scoring matrix M (A: above, B: below, L: left, LL: lower-left, LR: lower-right, O: overlap, R: right, UL: upper-left, UR: upper-right) for metric ToAcc-l is

$$M = \begin{array}{cc} & \begin{array}{ccccccccc} \text{A} & \text{B} & \text{L} & \text{LL} & \text{LR} & \text{O} & \text{R} & \text{UL} & \text{UR} \end{array} \\ \begin{array}{c} \text{A} \\ \text{B} \\ \text{L} \\ \text{LL} \\ \text{LR} \\ \text{O} \\ \text{R} \\ \text{UL} \\ \text{UR} \end{array} & \begin{pmatrix} 1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.3 & 0.3 \\ 0.0 & 1.0 & 0.0 & 0.3 & 0.3 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 1.0 & 0.6 & 0.0 & 0.0 & 0.0 & 0.6 & 0.0 \\ 0.0 & 0.3 & 0.6 & 1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.3 & 0.0 & 0.0 & 1.0 & 0.0 & 0.6 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.6 & 0.0 & 1.0 & 0.0 & 0.6 \\ 0.3 & 0.0 & 0.6 & 0.0 & 0.0 & 0.0 & 0.0 & 1.0 & 0.0 \\ 0.3 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.6 & 0.0 & 1.0 \end{pmatrix} \end{array} \tag{6}$$

And the scoring matrix M for metric ToAcc-a is

$$M = \begin{array}{cc} & \begin{array}{ccccccccc} \text{A} & \text{B} & \text{L} & \text{LL} & \text{LR} & \text{O} & \text{R} & \text{UL} & \text{UR} \end{array} \\ \begin{array}{c} \text{A} \\ \text{B} \\ \text{L} \\ \text{LL} \\ \text{LR} \\ \text{O} \\ \text{R} \\ \text{UL} \\ \text{UR} \end{array} & \begin{pmatrix} 1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.6 & 0.6 \\ 0.0 & 1.0 & 0.0 & 0.6 & 0.6 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 1.0 & 0.3 & 0.0 & 0.0 & 0.0 & 0.3 & 0.0 \\ 0.0 & 0.6 & 0.3 & 1.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.6 & 0.0 & 0.0 & 1.0 & 0.0 & 0.3 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 1.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.3 & 0.0 & 1.0 & 0.0 & 0.3 \\ 0.6 & 0.0 & 0.3 & 0.0 & 0.0 & 0.0 & 0.0 & 1.0 & 0.0 \\ 0.6 & 0.0 & 0.0 & 0.0 & 0.0 & 0.0 & 0.3 & 0.0 & 1.0 \end{pmatrix} \end{array} \tag{7}$$

For the Yelp-5 dataset, the tolerant score ToAcc is also follows equation 5, where the scoring matrix M is

$$M = \begin{array}{cc} & \begin{array}{ccccc} 0 & 1 & 2 & 3 & 4 \end{array} \\ \begin{array}{c} 0 \\ 1 \\ 2 \\ 3 \\ 4 \end{array} & \begin{pmatrix} 0.0 & 0.5 & 0.0 & 0.0 & 0.0 \\ 0.5 & 1.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.5 & 1.0 & 0.5 & 0.0 \\ 0.0 & 0.0 & 0.0 & 1.0 & 0.5 \\ 0.0 & 0.0 & 0.0 & 0.5 & 1.0 \end{pmatrix} \end{array} \tag{8}$$

C Confusion Matrices on Yelp-5

The confusion matrices for all the LLMs on Yelp-5 are illustrated in Figure 8. For most models, the dark part of the confusion matrix appears mainly on the diagonal, but there is still confusion on nearby prediction-label pairs (such as 1-2, 2-3). The Llama-2 models show a non-diagonal distribution and confusion on prediction-label pairs at longer distances.

D Further exploration on the attention mechanism

To understand how different components of the model handle positional information in text, we perform a quantitative analysis of the functional characteristics of the attention heads in the open-source Llama models.

After extracting the attention weights from all layers of models, we calculate the positional sensitivity of each attention head in every layer for each model. Specifically, for the attention matrix of a given head in a particular layer, we identify all token pairs that are separated by a distance d and compute the average attention for these pairs. Then, we fit a linear regression between attention and distance to obtain the slope of the regression line. If the slope is negative, i.e. attention decreases as the distance increases, the attention head is considered to exhibit positional sensitivity. The larger the absolute value of the slope, the faster the decay in attention and the stronger the positional sensitivity. If the slope is positive or zero, the positional sensitivity is set to 0, indicating that the head does not focus on positional information.

The heatmaps of the positional sensitivity matrices for different models are demonstrated in figure 9, with the horizontal and vertical axes representing the attention heads and layers, respectively. Notably, the positional sensitivity matrices of Llama-3.3-70b and Llama-3.1-70b are highly similar, while Llama-3-70b shows a related distribution but with some numerical differences. The matrices for Llama-3-8b, Llama-2-70b, and Llama-2-13b are all distinct.

Additionally, we present visualizations of attention matrices for some input instances at specific layers and attention heads in figure 10, illustrating the distribution of attention weights between words. The lower and upper layers tend to attend more broadly to contextual information, while the middle layers focus more on transforming local patterns.

TempNLI
Premise: Before 3 days, the grocery store will close.
Hypothesis: The grocery store will close after 54 hours.
Label: Neutral

MCTACO
C: It seemed strange to him, but not as strange as it was to see Linda the brown chicken in the living room last spring.
Q: How often does he find a wild animal in his house?
Options: he sees a wild animal in his house once every five years; he finds a wild animal in his house once a day; he finds a wild animal in his house once every five years; he finds a wild animal in his house once every five seconds.
Label: yes; no; yes; no

(a) Data instances in temporal datasets.

Multi-hop Space
C1: D presents left to N.
C2: D is at P's 3 o'clock.
C3: S and P are parallel, and S is on top of P.
C4: S is positioned in the front right corner of M.
Q: What is the relation of the agent S to the agent N?
Label: upper-left

SpaceTrans
Premise: The painting is above the garden.
The garden is behind my need for a hobby.
Statement: The painting is behind my need for a hobby.
Label: no

(b) Data instances in spatial datasets.

ECI
C: The Third Cod War concluded in 1976 , with a highly favourable agreement for Iceland ; the United Kingdom conceded to a Icelandic exclusive fishery zone after threats that Iceland would withdraw from NATO , which would have forfeited NATO 's access to most of the GIUK gap , a critical anti-submarine warfare during the Cold War .
Events: threats, conceded
Label: 1

FantasyR
C: In a world filled with magic, your family is scorned for generations for wasting time with science. Your mother was a botanist. Your father, a biologist. Mages can heal by touching. You developed steam locomotion when mages teleport. Your family has never trusted magic. One day, also known as the Fateful Day, the magic stops working. A mage is suspended in the air by magic when the Fateful Day arrives.
Q: Can the mage touch the ground anymore?
Label: yes

(c) Data instances in sentimental datasets.

Yelp-5
C: Arriba's was not as good as they used to be, apparently the original owner passed away and its under new ownership. Won't be coming back here again.
Label: 1 (0~4)

IronyEval
C: Waking up with a pounding headache is just what I need for this final.
Label: 1

(d) Data instances in causal datasets.

Figure 7: Data instances of the WorldInsight BENCH.

Method	Temporal			Spatial					Sentimental			Causal				
	TempNLI	MCTACO		M-h Space				SpaceT	Yelp-5		IronyEval	ECI		FantasyR		
	acc	EM	F1	Acc	Macro F1	ToAcc-l	ToAcc-a	Acc	Acc	ToAcc	Acc	Acc	F1	Acc	Acc-i	Acc-e
GPT-4o	63.50	53.75	77.08	48.75	44.06	**66.00**	**66.00**	88.25	61.50	78.50	79.00	35.25	35.25	80.00	76.92	85.71
+COT	70.25	60.00	80.65	42.50	37.22	52.32	51.35	89.50	59.25	77.62	77.50	59.00	54.95	81.00	82.31	78.57
+FS	70.25	57.25	80.80	46.75	44.98	55.53	54.85	89.25	**63.50**	**79.12**	**90.25**	64.75	58.14	81.00	80.77	81.43
+FS CoT	**70.75**	**74.50**	**89.42**	**52.75**	**50.63**	61.30	63.55	**92.00**	60.25	77.75	81.75	**66.50**	**59.12**	**91.50**	**91.54**	**91.43**
Llama-3.3-70b-instruct	53.50	54.75	78.26	36.00	32.25	46.95	49.57	82.50	57.75	76.38	74.00	58.50	54.97	75.50	73.08	80.00
+COT	70.00	63.25	84.50	48.25	43.34	58.23	59.87	87.25	**58.00**	**77.12**	76.25	54.25	52.58	80.00	77.69	**84.29**
+FS	71.25	58.50	81.50	**54.75**	**50.71**	**63.90**	**66.53**	85.75	57.50	76.75	**82.25**	31.75	31.48	79.50	78.46	81.43
+FS CoT	**74.50**	**72.75**	**90.45**	45.00	41.64	54.98	57.30	**88.75**	55.75	75.00	78.50	**59.50**	**56.32**	**83.00**	**83.08**	82.86
Llama-3.1-70b-instruct	50.50	49.25	73.96	38.00	33.44	46.85	48.05	86.25	**58.25**	**76.25**	73.75	43.75	43.26	78.50	76.92	81.43
+COT	64.50	57.50	79.43	44.00	40.02	53.15	54.42	87.50	52.75	74.00	72.50	55.50	53.65	76.00	77.69	72.86
+FS	63.00	44.75	67.62	**50.00**	**46.42**	**59.15**	**59.30**	87.50	56.50	75.00	**83.00**	55.75	52.74	**84.00**	**82.31**	**87.14**
+FS CoT	**72.00**	**66.50**	**86.92**	44.00	40.29	51.88	53.67	**91.75**	53.50	73.38	78.50	**68.00**	**62.92**	82.00	80.00	85.71
Llama-3-70b-instruct	50.25	33.25	59.92	25.25	22.55	34.32	33.50	79.75	55.00	71.75	72.50	**70.25**	**55.75**	63.00	59.23	70.00
+COT	48.25	31.25	59.40	31.75	28.96	39.47	40.45	85.25	57.75	76.00	73.75	49.75	48.30	76.50	75.38	78.57
+FS	51.75	**48.75**	**72.63**	**40.25**	**36.82**	**48.43**	**49.40**	83.00	**59.50**	**77.75**	**81.00**	28.75	28.14	76.00	**77.69**	72.86
+FS CoT	**70.75**	47.00	71.43	28.25	23.62	34.47	34.62	**89.00**	56.25	74.75	79.50	56.50	54.07	**77.00**	75.38	**80.00**
Llama-3-8b-instruct	46.25	37.75	**71.07**	**23.25**	20.99	**30.97**	31.72	71.50	46.25	68.88	59.75	**71.00**	**59.67**	70.50	67.69	75.71
+COT	41.00	18.25	60.88	15.50	15.37	20.75	19.25	75.00	50.75	73.88	56.75	47.25	44.37	70.50	69.23	72.86
+FS	50.00	**41.50**	76.64	20.25	15.02	29.17	27.97	70.50	51.75	73.50	73.75	38.75	38.38	61.50	60.77	62.86
+FS CoT	**50.75**	28.50	57.31	22.75	**22.32**	30.47	**32.80**	**84.00**	**57.50**	**75.50**	**77.50**	46.75	46.00	**74.00**	**70.00**	**81.43**
Llama-2-70b-chat-hf	45.50	**24.50**	**63.08**	22.75	18.39	32.87	38.95	65.25	29.50	45.25	61.50	19.00	16.77	61.50	58.46	67.14
+COT	47.25	19.25	61.64	**25.25**	**21.83**	**36.80**	**44.75**	76.00	**59.50**	**75.25**	52.00	**45.75**	**43.88**	**75.00**	**76.15**	72.86
+FS	**48.50**	14.25	56.35	21.00	13.31	32.55	30.82	63.25	50.25	67.12	**70.00**	21.50	19.16	64.00	64.62	62.86
+FS CoT	45.75	23.00	55.53	24.25	21.51	33.10	35.65	**85.50**	58.50	73.38	69.50	38.75	38.69	73.00	72.31	**74.29**
Llama-2-13b-chat-hf	**49.50**	7.75	54.39	9.00	9.39	14.32	14.70	51.50	47.25	64.38	42.00	31.75	31.52	**66.50**	**65.38**	**68.57**
+COT	47.00	13.25	55.18	17.75	14.78	27.87	**31.47**	**75.00**	39.50	52.62	49.50	**38.75**	**38.22**	64.50	**65.38**	62.86
+FS	44.25	**15.50**	**57.21**	12.50	11.08	20.22	19.17	57.25	33.00	46.88	57.75	21.25	19.11	**66.50**	**65.38**	**68.57**
+FS CoT	49.00	15.00	47.38	**23.50**	**17.73**	**31.82**	31.37	71.25	**60.50**	**74.00**	**71.50**	37.75	37.55	60.50	63.08	55.71

Table 4: Full experimental results.

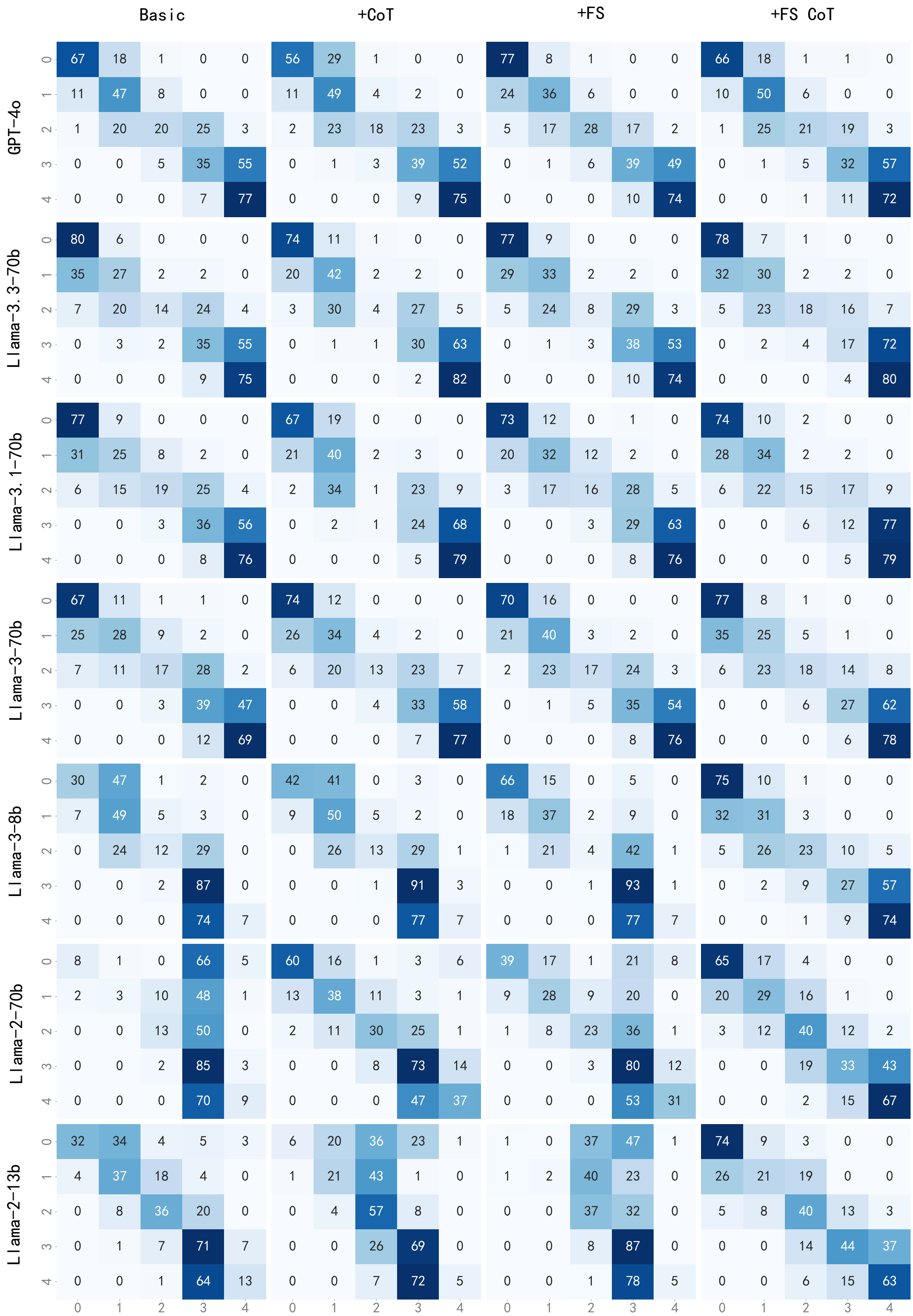
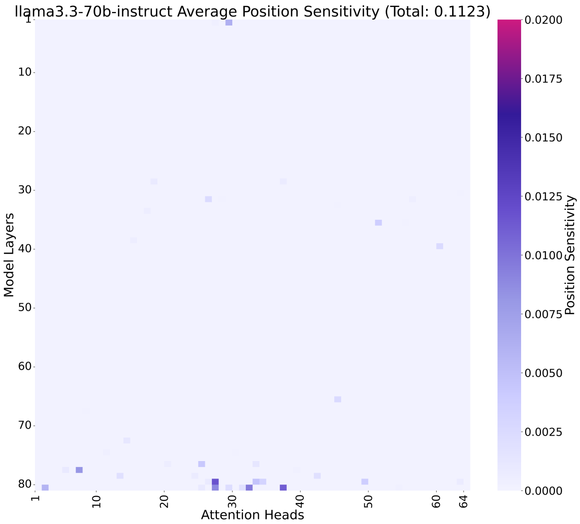
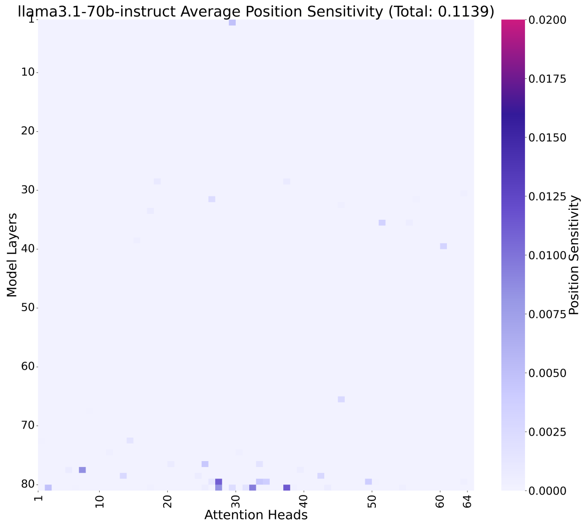


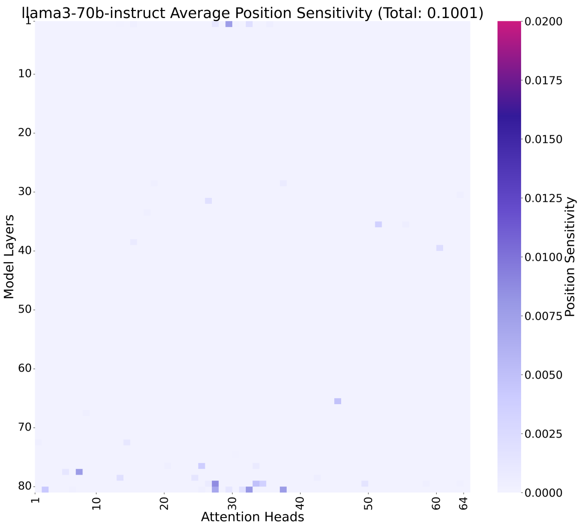
Figure 8: All the LLMs are assessed with confusion matrices on Yelp-5. The horizontal axis represents the predicted value, and the vertical axis represents the true value. The color depth on the diagonal determines the ability of models to explicit classify.



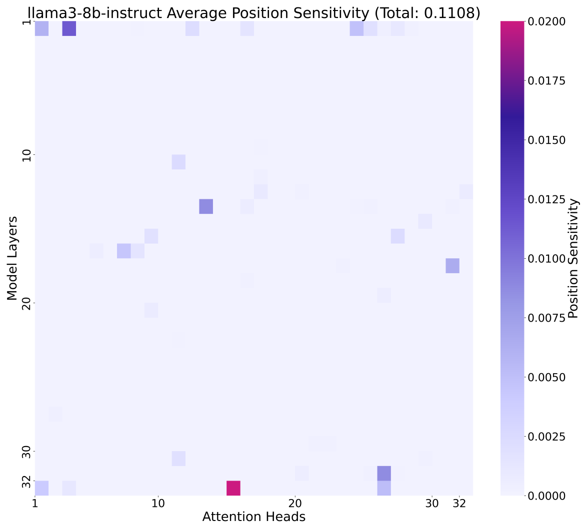
(a) Llama-3.3-70b



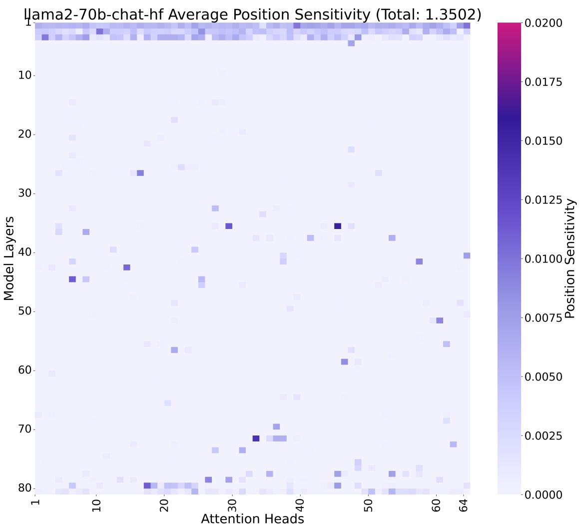
(b) Llama-3.1-70b



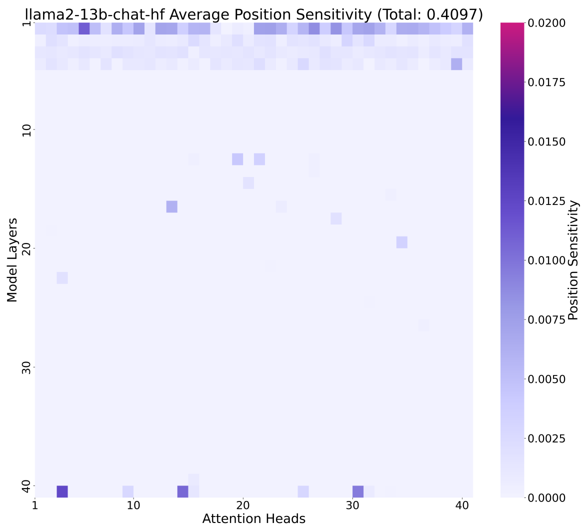
(c) Llama-3-70b



(d) Llama-3-8b

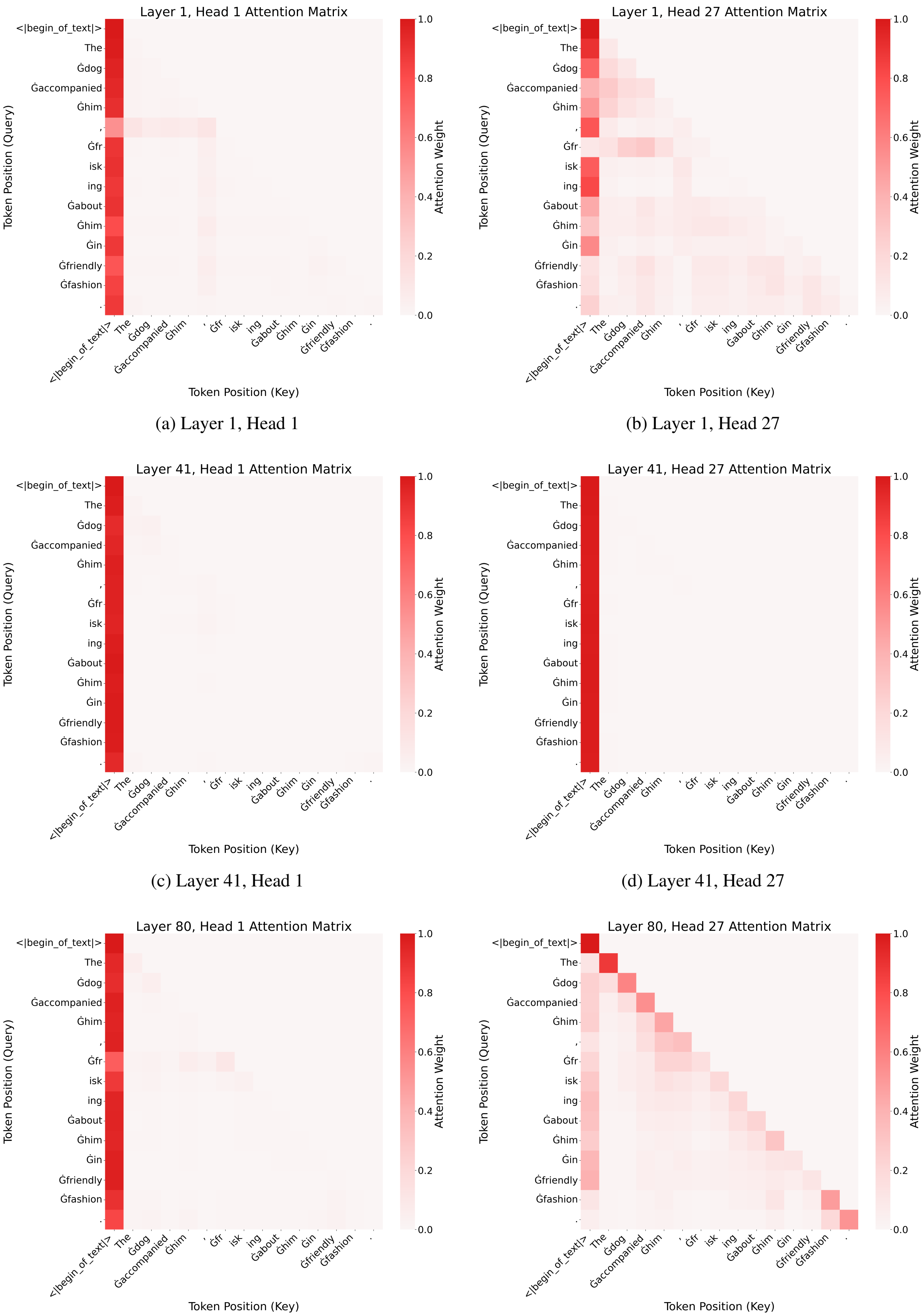


(e) Llama-2-70b



(f) Llama-2-13b

Figure 9: Positional sensitivity matrices for LLMs.



(a) Layer 1, Head 1

(b) Layer 1, Head 27

(c) Layer 41, Head 1

(d) Layer 41, Head 27

(e) Layer 80, Head 1

(f) Layer 80, Head 27

Figure 10: Attention matrices for Llama-3.3-70b.