Unveiling Divergent Inductive Biases of LLMs on Temporal Data

Sindhu Kishore
University of Rochester
skishor2@ur.rochester.edu

Hangfeng He
University of Rochester
hangfeng.he@rochester.edu

Abstract

Unraveling the intricate details of events in natural language necessitates a subtle understanding of temporal dynamics. Despite the adeptness of Large Language Models (LLMs) in discerning patterns and relationships from data, their inherent comprehension of temporal dynamics remains a formidable challenge. This research meticulously explores these intrinsic challenges within LLMs, with a specific emphasis on evaluating the performance of GPT-3.5 and GPT-4 models in the analysis of temporal data. Employing two distinct prompt types, namely Question Answering (QA) format and Textual Entailment (TE) format, our analysis probes into both implicit and explicit events. The findings underscore noteworthy trends, revealing disparities in the performance of GPT-3.5 and GPT-4. Notably, biases toward specific temporal relationships come to light, with GPT-3.5 demonstrating a preference for “AFTER” in the QA format for both implicit and explicit events, while GPT-4 leans towards “BEFORE”. Furthermore, a consistent pattern surfaces wherein GPT-3.5 tends towards “TRUE”, and GPT-4 exhibits a preference for “FALSE” in the TE format for both implicit and explicit events. This persistent discrepancy between GPT-3.5 and GPT-4 in handling temporal data highlights the intricate nature of inductive bias in LLMs, suggesting that the evolution of these models may not merely mitigate bias but may introduce new layers of complexity.[1].

1 Introduction

Temporal relations play a crucial role across diverse applications, including event summarization (Wang et al., 2018; Keith Norambuena et al., 2023), predicting future events (Lin et al., 2022), and medical information processing (Jung et al., 2011; Alfattni et al., 2020). Despite their importance, LLMs, especially those with limited context windows, face challenges in accurately sequencing events due to intricate temporal dependencies. Efforts have been devoted to developing methodologies for effective temporal relation extraction (Choubey and Huang, 2017; Ning et al., 2018b, 2019; Wang et al., 2020; Zhang et al., 2022), along with initiatives to create benchmark datasets with a temporal focus (Pustejovsky et al., 2003; Verhagen et al., 2010; Ning et al., 2018a; Zhou et al., 2021; Gantt et al., 2022). However, discerning causal relationships between events adds complexity and can lead to misunderstandings. This complexity is strengthened by the absence of explicit temporal reasoning mechanisms, introducing biases in models' predictions and preferences for specific temporal relations.

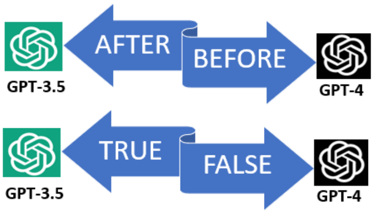


Figure 1: Inductive bias in OpenAI LLMs: GPT-4 exhibits a preference for BEFORE and FALSE, while GPT-3.5 tends to favor AFTER and TRUE.

Surprisingly, a notable gap exists in research exploring inductive bias in LLMs when discerning temporal relations. Our study investigates the temporal comprehension abilities of GPT-3.5 and GPT-4 (OpenAI, 2023), aiming to understand their grasp of temporal relationships. Despite frequent model updates, significant biases were unveiled. Using Question Answering (QA) and Textual Entailment (TE) prompts, we queried both models to determine temporal relations. Illustrated in Figure 1, the results expose variations in GPT-3.5 and GPT-4 performance, revealing biases towards specific temporal relationships. GPT-3.5 favors "AFTER" in QA for implicit and explicit events, while GPT-4 leans towards "BEFORE." In TE, GPT-3.5 tends

[1] Our code is publicly available at https://github.com/SindhuKRao/LLM_temporal_Bias.

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 220–228
June 16-21, 2024 ©2024 Association for Computational Linguistics

towards "TRUE," and GPT-4 prefers "FALSE" for both implicit and explicit events.

2 Methodology

Our analysis involved two types of temporal data: one focusing on implicit events, actions, or situations not directly articulated in the text but inferred from context, while the other centered on explicit events explicitly mentioned in the context. Furthermore, we delved into two prompt formats to gauge their influence on aiding LLM in generating responses. These formats comprised the QA format, where questions prompt the model, and the TE format, tailored to assess the logical relationships within sentences specifying temporal relations.

Question answering format. We initially conducted experiments using the QA format, focusing on explicit events. In this configuration, we tasked both models with determining the temporal relation ("BEFORE" or "AFTER") between two provided events within the given context. The same approach was applied to implicit events. Figure 2 provides the template and examples illustrating this format.

Textual Entailment Format. Subsequently, we employed the textual entailment format as the next prompt type. In this format, we presented the model with a context along with a sentence declaring the temporal relation between two events, and then tasked the model with assessing its truthfulness. For every pair, there exists one TRUE and one FALSE label, as one corresponds to the gold label, and the other represents an incorrect label.Examples illustrating this format are provided in Figure 2.

Inductive Bias Measurement: In our evaluation, we focused on probing the model's inductive biases related to temporal relations. To quantify the inductive bias, we examined the model's preference for "BEFORE" and "AFTER" relations in the QA format and assessed its tendencies toward "TRUE" and "FALSE" in the Textual Entailment format.

3 Experimental Settings

Dataset For our experimentation, we employed datasets such as TimeBank (Pustejovsky et al., 2003), Tempeval (Verhagen et al., 2010), AQUAINT, and TRACIE (Zhou et al., 2021). The analysis of implicit and explicit events was conducted separately. We extracted "BEFORE"/"IBEFORE" as "BEFORE" and "AFTER," /"IAFTER" as "AFTER" from TimeBank, events from Task C in TempEval featuring "BEFORE" or "AFTER" relations. This yielded 1576 explicit events from TimeBank, TempEval, and AQUAINT datasets, comprising 815 "AFTER" and 761 "BEFORE" events. The dataset was duplicated for the Entailment format, creating inverse relations with gold as "TRUE" and inverse as "FALSE," expanding the dataset to 3150 events with 1575 "TRUE" and "FALSE" values.

For Implicit events, the TRACIE dataset (Zhou et al., 2021) was used. Transforming "starts after/ends after" into "AFTER" and "starts before/ends before" into "BEFORE",the dataset included a total of 22,050 events evenly distributed between "TRUE" and "FALSE" labels, representing gold and inverse relations. Among the 11,025 gold relations, 4,659 were identified as "AFTER," while 6,366 were classified as "BEFORE".

Large language models. The GPT series, renowned as the leading range of Large Language Models (LLMs), holds widespread popularity. Our analysis began with these models due to their extensive usage, leaving the investigation of biases in other LLMs for potential future research. We conducted our analysis using OpenAI's GPT-3.5 and GPT-4 models, specifically employing the latest stable versions: gpt-3.5-turbo-1106 and gpt-4-1106-preview. The gpt-3.5-turbo-1106 model has a default context window of 16k tokens, while GPT-4 features a context window of 128k tokens.

Performance Measurement We assessed bias by examining patterns in prediction preferences, aiming to determine if the models consistently favored or exhibited imbalances in predicting specific temporal relations. We scrutinized tendencies towards "BEFORE" and "AFTER" relations in the QA format, and in the TE format, we analyzed biases towards "TRUE" and "FALSE". Furthermore, we tested the models for consistency by presenting identical events and contexts with reversed temporal relations.

4 Results and Analysis

BEFORE/AFTER bias in QA: We evaluated the models' performance in the QA format for explicit events. Among 1576 instances, comprising 914 with a "BEFORE" relation and 662 with an "AFTER" relation, GPT-3.5 demonstrated a bias towards 815 prompts as "AFTER" and 761 as "BE-

QA:Implicit Event

Template: *context:<context>.E1:<event1>.E2:<event2>. For the given context, give the temporal relation between E1 and E2 as before or after. Give your answer in the format, Relation:<ans>.*

Prompt:Context:'Trisha was a manager of a local diner. It was a slow night. The diner hadn't been making any money. Trisha didn't want the diner to lose any more money. Trisha closed the diner an hour early.'E1:'Trisha got a job at a diner '. E2:'Trisha closed the diner'. For the given context, give the temporal relation between E1 and E2 as before or after. Give your answer in the format, Relation:<ans>.

Response:Relation:AFTER.

QA:Explicit Event

Template: *Context:<context>.For the given context,give the temporal relation between (eid:<event>) and (eid:<event>) as BEFORE or AFTER. Give your answer in the format, Relation:<ans>.*

Prompt:context:'The American ambassador to Kenya was among hundreds (e12:injured), a local TV (e4:said)'.For the given context, give the temporal relation between (e12:injured) and (e4:said) as before or after. Give your answer in the format, Relation:<ans>.

Response:Relation:BEFORE.

TE:Implicit Event

Template: *context:<context>.For the given context,state if the statement event:<event> is True or False. Give your answer in the format , Ans:<ans>.*

Prompt:context:'Trisha was a manager of a local diner. It was a slow night. The diner hadn't been making any money. Trisha didn't want the diner to lose any more money. Trisha closed the diner an hour early.'For the given context,state if the statement event:' Trisha got a job at a diner starts before Trisha closed the diner ' is True or False. Give your answer in the format , Ans:<ans>.

Response: Ans:True.

Prompt:context:'Trisha was a manager of a local diner. It was a slow night. The diner hadn't been making any money. Trisha didn't want the diner to lose any more money. Trisha closed the diner an hour early.'.For the given context,state if the statement event:' Trisha got a job at a diner starts after Trisha closed the diner ' is True or False. Give your answer in the format , Ans:<ans>.

Response: Ans:False.

TE:Explicit Event

Prompt:context:'The American ambassador to Kenya was among hundreds (e12:injured), a local TV (e4:said)'.For the given context, state if the statement: 'event (e12:injured) is BEFORE (e4:said)' is True or False. Give your answer in the format , Ans:<ans>.

Response: Ans:True.

Prompt:context:'The American ambassador to Kenya was among hundreds (e12:injured), a local TV (e4:said)'.For the given context,state if the statement: 'event (e12:injured) is AFTER (e4:said)' is True or False. Give your answer in the format , Ans:<ans>.

Response: Ans:False.

Figure 2: Template and Examples of QA and TE prompts for implicit & explicit events.

FORE", indicating a preference for AFTER, as shown in Figure 3. In contrast, GPT-4 exhibited a preference for "BEFORE", leaning towards 1057 prompts as "BEFORE" and 519 as "AFTER", revealing a divergent pattern between the two models.

For implicit events, totaling approximately 11,652 entries, with 6,735 indicating a "BEFORE" relation and 4,917 an "AFTER" relation, both models displayed patterns resembling those observed in explicit events. GPT-3.5 predominantly favored the "AFTER" relation, identifying 6,232 instances as "AFTER", 5,329 as "BEFORE", and approximately 91 as indeterminable. Conversely, GPT-4 leaned towards "BEFORE", marking 6,811 instances as "BEFORE", 4,594 as "AFTER", and 247 as indeterminable. The contrasting outcomes between the models in both explicit and implicit events add an intriguing and contradictory dimension to their assessments.

Consistency in TE. We now analyzed the results of TE format. We encountered an unexpected pattern in both the implicit and explicit events. We had anticipated an even distribution of 'True' and 'False' responses due to the contradictory pairs as discussed in Section 2. However, we found inconsistencies where the model consistently produced matching responses—yielding ("True", "True") or ("False", "False") rather than the expected mix of ("True", "False") or ("False", "True") in numerous instances. To delve deeper, we categorized our findings into consistent and inconsistent pairs for further examination.

TRUE/FALSE bias in TE-Inconsistent Pair These pairs contain actual values of "True" and "False", yet the predicted values consistently align as either ("True", "True") or ("False", "False"). For implicit events, GPT-3.5 exhibited approximately 83.3% inconsistency, while GPT-4 showed 67.1% inconsistency. In explicit events, GPT-3.5 demon-

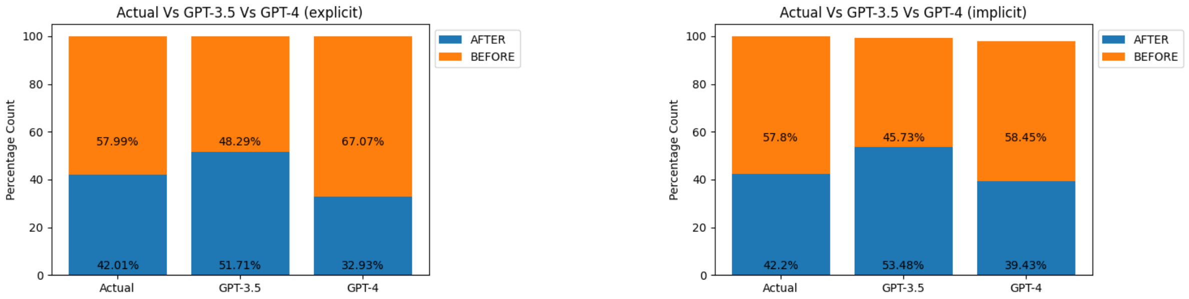


Figure 3: GPT-3.5 biased towards AFTER and GPT-4 biased towards BEFORE in QA.

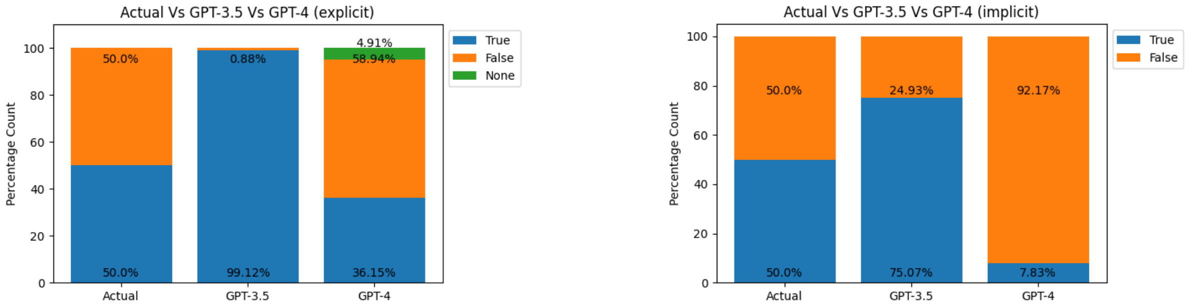


Figure 4: GPT-3.5 biased towards TRUE and GPT-4 biased towards FALSE in TE -Inconsistent pair.

strated 94.6% inconsistency, whereas GPT-4 presented only 32.4% inconsistent results. Comparing both models, it's evident that GPT-4 displays greater consistency compared to GPT-3.5 based on these findings. Upon further analysis of these inconsistent results to check for biases, we made a surprising discovery. GPT-3.5 tends to show a bias towards "True", while GPT-4 leans towards "False" as shown in Figure 4. This bias was consistently observed in both implicit and explicit events, revealing a contradicting bias between the models.

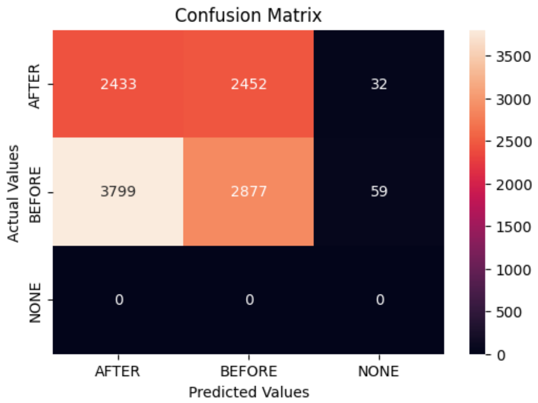
Model	Event	Relation	Actual	Prediction
GPT-3.5	Implicit	BEFORE AFTER	48.02% 51.98%	26.48% 73.52%
	Explicit	BEFORE AFTER	50.00% 50.00%	50.00% 50.00%
GPT-4	Implicit	BEFORE AFTER	62.82% 37.18%	70.42% 29.58%
	Explicit	BEFORE AFTER	50.00% 50.00%	50.00% 50.00%

Table 1: Actual vs Predicted distribution in consistent TE.

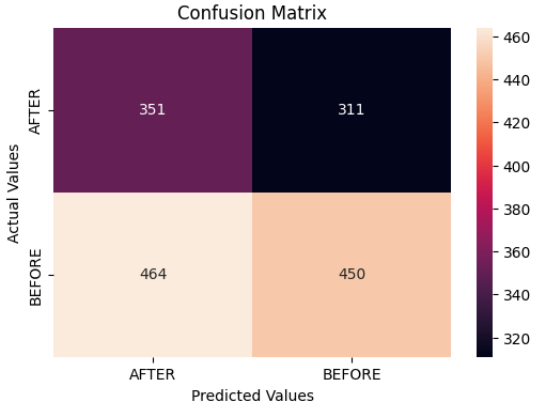
BEFORE/AFTER bias in TE-Consistent Pair
These pairs encompass actual values of "True" and "False", with predicted values aligning as either ("True", "False") or ("False", "True"). For implicit events, roughly 16.7% of GPT-3.5's total results were consistent, while GPT-4 showed 32.9% consistency, while explicit events had, approximately 5.4% consistency for GPT-3.5 and 67.6% consistency in GPT-4. Upon examining these consistent pairs to detect bias toward "BEFORE" and "AFTER", we noted a familiar pattern. For implicit events GPT-3.5 displayed a bias toward "AFTER", whereas GPT-4 leaned toward "BEFORE". However, the results were notably consistent and unbiased for explicit events as shown in Table1. This discrepancy might arise since the model is able to comprehend context more effectively and provide unbiased predictions. In contrast, the implicit events poses challenges for the model to assess accurately, potentially leading to biased results. Additional information is available in Appendix A.

5 Conclusion and Future Work

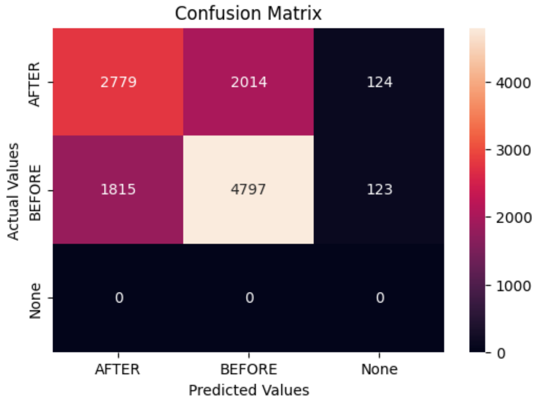
Our study identified performance disparities between GPT-3.5 and GPT-4, with the latter showing more consistency. Notably, biases were observed, as GPT-3.5 favored "AFTER", while GPT-4 favored "BEFORE", and GPT-3.5 tended towards "TRUE", while GPT-4 favored "FALSE". This consistent yet contradictory pattern raises questions about whether new model releases might unintentionally introduce new biases. The observed biases across multiple datasets and prompt formats warrant a deeper exploration of the models' understanding of temporal data. Future investigations should prioritize tasks involving temporal reasoning to address biases in GPT-3.5 and GPT-4, considering diverse datasets and prompt structures.



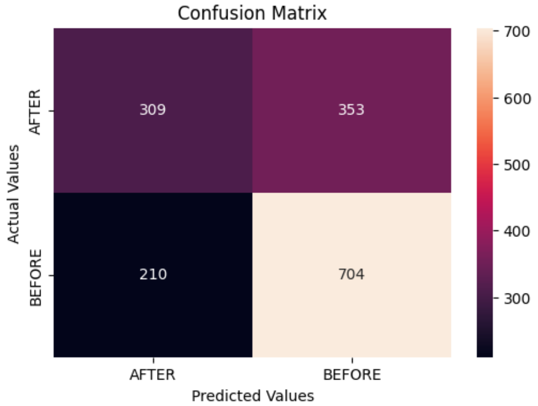
(a) GPT-3.5 : QA implicit events.



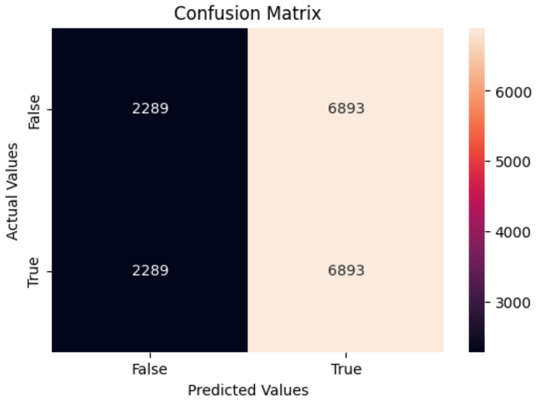
(b) GPT-3.5 : QA explicit events.



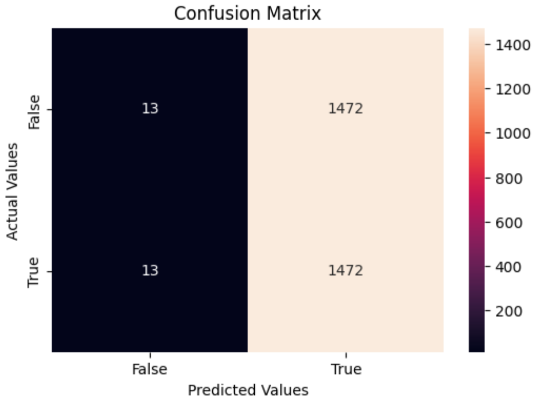
(c) GPT-4 : QA implicit events.



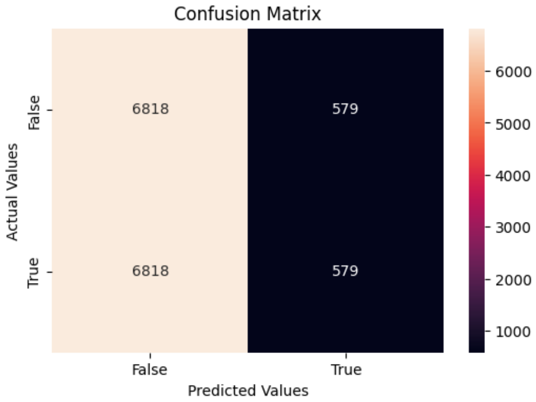
(d) GPT-4 : QA explicit events.



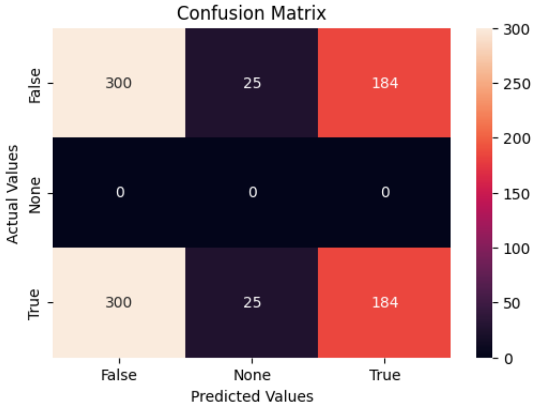
(a) GPT-3.5 : inconsistent TE implicit events.



(b) GPT-3.5 : inconsistent TE explicit events.



(c) GPT-4 : inconsistent TE implicit events.



(d) GPT-4 : inconsistent TE explicit events.

6 Limitations

Our study's findings, drawn from the analysis of GPT-3.5 and GPT-4, suggest that the identified patterns may be specific to these models and not universally applicable to language models with different architectures or training methodologies. Given the continuous development of language models and the potential for new versions with updates, the biases observed in GPT-3.5 and GPT-4 may not persist in future releases. Recognizing the impact of prompt types on model performance, our study emphasizes the ongoing need for exploration to determine the most effective prompt types across different contexts. While the QA prompt showed improved predictions in some cases, the Textual Entailment format proved beneficial in others, underscoring the importance of selecting appropriate prompt types for comprehensive analyses. Interestingly, the "BEFORE"/"AFTER" bias observed in the QA format and TE consistent pair implicit events did not reappear in TE consistent pair explicit events, potentially influenced by the lower percentage of data in this category.

References

Ghada Alfattni, Niels Peek, and Goran Nenadic. 2020. Extraction of temporal relations from clinical free text: A systematic review of current approaches. *Journal of Biomedical Informatics*, 108:103488.

Prafulla Kumar Choubey and Ruihong Huang. 2017. A sequential model for classifying temporal relations between intra-sentence events. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1796–1802.

William Gantt, Lelia Glass, and Aaron Steven White. 2022. Decomposing and recomposing event structure. *Transactions of the Association for Computational Linguistics*, 10:17–34.

Hyuckchul Jung, James Allen, Nate Blaylock, William de Beaumont, Lucian Galescu, and Mary Swift. 2011. Building timelines from narrative clinical records: initial results based-on deep natural language understanding. In *Proceedings of BioNLP 2011 workshop*, pages 146–154.

Brian Felipe Keith Norambuena, Tanushree Mitra, and Chris North. 2023. A survey on event-based news narrative extraction. *ACM Computing Surveys*, 55(14s):1–39.

Li Lin, Yixin Cao, Lifu Huang, Shu'Ang Li, Xuming Hu, Lijie Wen, and Jianmin Wang. 2022. What makes the story forward? inferring commonsense explanations as prompts for future event generation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1098–1109.

Qiang Ning, Sanjay Subramanian, and Dan Roth. 2019. An improved neural baseline for temporal relation extraction. 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 6203–6209.

Qiang Ning, Hao Wu, and Dan Roth. 2018a. A multi-axis annotation scheme for event temporal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1318–1328.

Qiang Ning, Ben Zhou, Zhili Feng, Haoruo Peng, and Dan Roth. 2018b. Cogcomptime: A tool for understanding time in natural language. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 72–77.

OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, et al. 2003. The timebank corpus. In *Corpus linguistics*, volume 2003, page 40. Lancaster, UK.

Marc Verhagen, Roser Sauri, Tommaso Caselli, and James Pustejovsky. 2010. Semeval-2010 task 13: Tempeval-2. In *Proceedings of the 5th international workshop on semantic evaluation*, pages 57–62.

Chengyu Wang, Xiaofeng He, and Aoying Zhou. 2018. Event phase oriented news summarization. *World Wide Web*, 21:1069–1092.

Haoyu Wang, Muhao Chen, Hongming Zhang, and Dan Roth. 2020. Joint constrained learning for event-event relation extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 696–706.

Shuaicheng Zhang, Qiang Ning, and Lifu Huang. 2022. Extracting temporal event relation with syntax-guided graph transformer. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 379–390.

Ben Zhou, Kyle Richardson, Qiang Ning, Tushar Khot, Ashish Sabharwal, and Dan Roth. 2021. Temporal reasoning on implicit events from distant supervision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1361–1371.

A Appendix

A.1 Experimental Details

Below are few with the experimental results and other data gathered from our experiments in both implicit and explicit event for both Textual Entailment and Question Answering format.

A.1.1 Textual Entailment

Model's Consistency

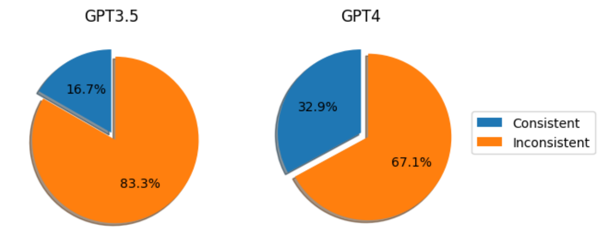


Figure 7: Consistency in Response for implicit events

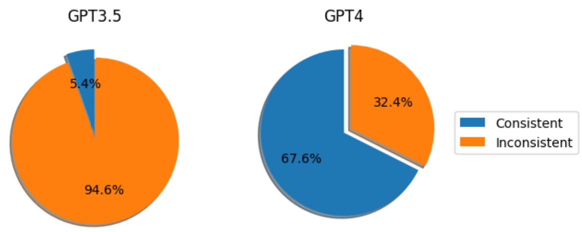


Figure 8: Consistency in Response for explicit events

Although our anticipation was for the model to provide one TRUE and one FALSE for each pair, we observed a discrepancy. The models did yield ("TRUE","FALSE") for certain instances, but surprisingly, they often produced ("TRUE","TRUE") or ("FALSE","FALSE"). Figures 7 and 8 visually depict the inconsistency discussed in Section 4. Notably, we observe that GPT-4 exhibits more consistency than GPT-3.5 for both implicit and explicit events when prompts are presented in the Textual Entailment format.

BEFORE/AFTER Bias in Consistent pair (Implicit Events)

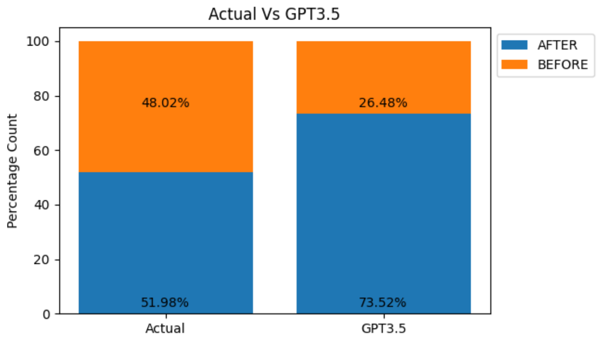


Figure 9: GPT-3.5 biased towards AFTER

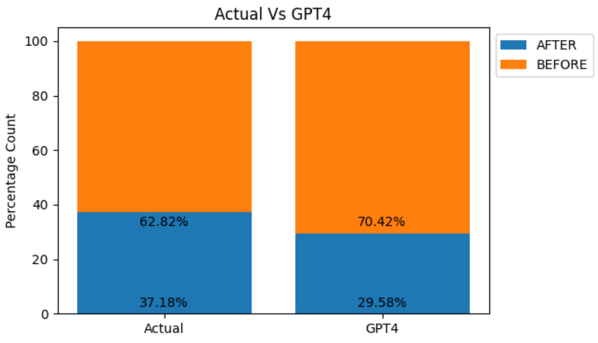


Figure 10: GPT-4 biased towards BEFORE

Figures 9 & 10 visually illustrate the observed bias in the Textual Entailment consistent pair. As outlined in the Section 4, we note GPT-3.5 demonstrating a bias towards "AFTER" and GPT-4 exhibiting a bias towards "BEFORE". While this behavior was previously observed in the Question Answering format for both implicit and explicit events, it is notable that in the Textual Entailment consistent format, this bias is observed only for implicit events.

Unbiased consistent pair (Explicit Events)

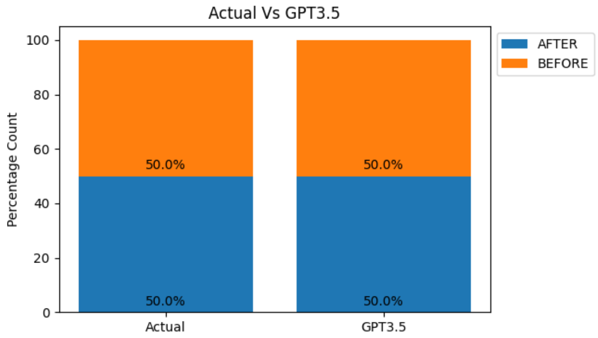


Figure 11: Unbiased GPT-3.5

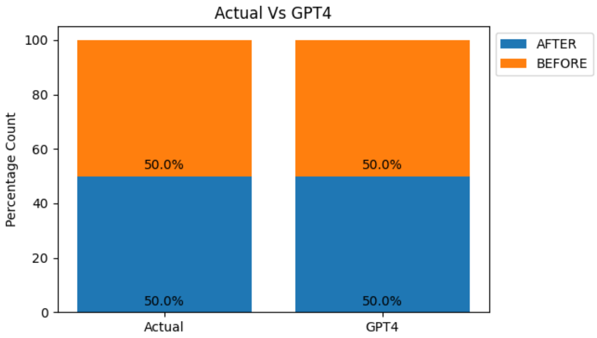


Figure 12: Unbiased GPT-4

In the preceding section for Consistent pair Textual Entailment, we noted that bias is observed in implicit events, attributed to their complexity. However, explicit events are better comprehended by the model, and the bias is absent here as shown in the Figure 11 & 12 .

Results of TE(Inconsistent pair)

Event	Relation	Actual	GPT-3.5	GPT-4
Implicit	TRUE	50.0%	75.07%	7.83%
	FALSE	50.0%	24.93%	92.17%
Explicit	TRUE	50.0%	99.12%	0.88%
	FALSE	50.0%	36.15%	58.94%

Table 2: Actual vs Predicted distribution of GPT-3.5 & GPT-4 in TE.

Table 2 clearly shows that GPT-3.5 tends to favor "FALSE" for both implicit and explicit events, whereas GPT-4 shows a preference for "TRUE" in both event types.

A.1.2 Question Answering

Results of QA

Event	Relation	Actual	GPT-3.5	GPT-4
Implicit	BEFORE	57.8%	45.73%	58.45%
	AFTER	42.2%	53.48%	39.43%
Explicit	BEFORE	57.99%	48.29%	67.07%
	AFTER	42.01%	51.71%	32.93%

Table 3: Actual vs Predicted distribution of GPT-3.5 & GPT-4 in QA

As previously discussed in Section 4, Table 3 shows GPT-3.5 demonstrates a bias toward the “BEFORE” relation for both implicit and explicit events. Conversely, GPT-4 exhibits a conflicting bias, showing a preference for the “AFTER” relation in both types of events.

A.2 Additional Results

Consistent TE GPT-3.5

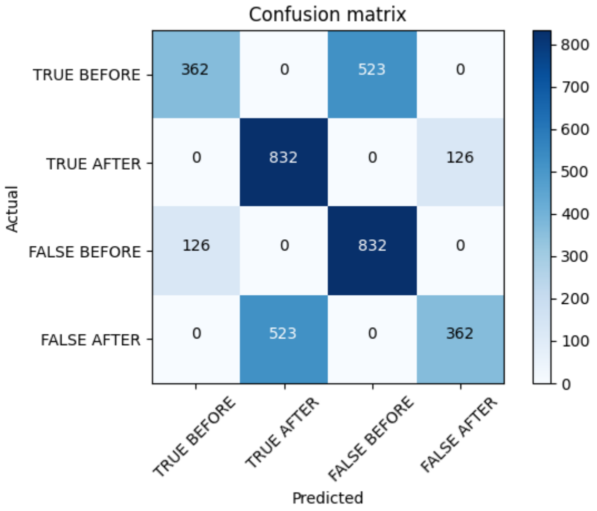


Figure 13: Confusion matrix of GPT-3.5 for implicit events

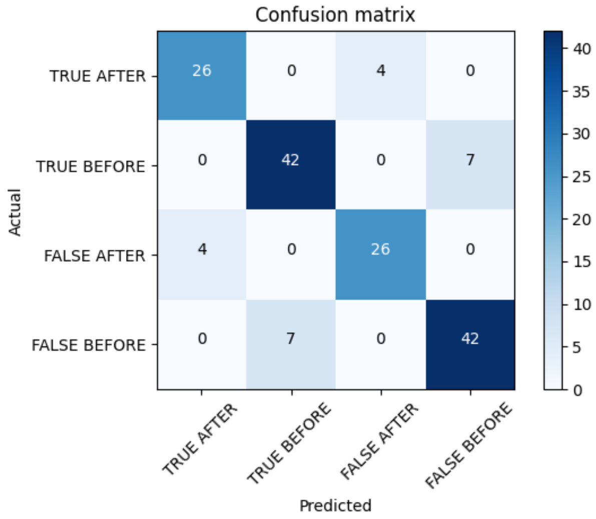


Figure 14: Confusion matrix of GPT-3.5 for explicit events

Figure 13 & 14 display the confusion matrix of GPT-3.5 for TE consistent pairs. These images vividly illustrate the breakdown of actual and predicted values for the BEFORE and AFTER relations. As mentioned earlier in Section 4, We see that GPT-3.5 tends to exhibit bias towards AFTER in implicit events, while remaining unbiased for explicit events.

Consistent TE GPT-4

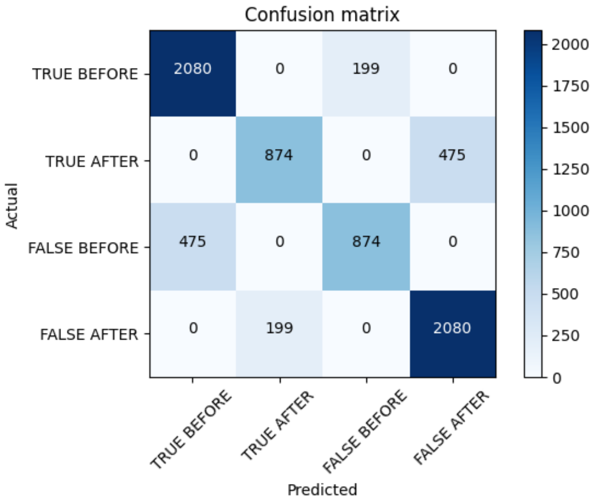


Figure 15: Confusion matrix of GPT-4 for implicit events

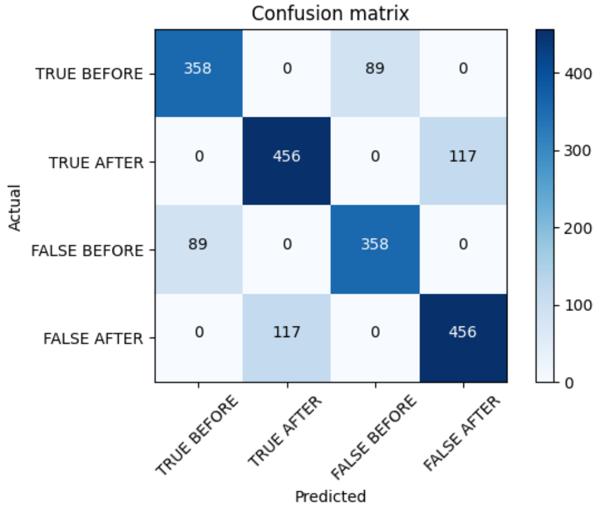


Figure 16: Confusion matrix of GPT-4 for explicit events.

Figures 15 and 16 present the confusion matrix of GPT-4 for TE consistent pairs. Once more, they depict the breakdown of actual and predicted values for the BEFORE and AFTER relations. As previously discussed in Section 4, we observe that GPT-4 demonstrates a tendency towards biasing predictions towards BEFORE in implicit events, which contrasts with the behavior of GPT-3.5. However, GPT-4 remains unbiased for explicit events.