Conservative Bias in Large Language Models: Measuring Relation Predictions

Anonymous ACL submission

Abstract

Large language models (LLMs) exhibit pronounced *conservative bias* in relation extraction tasks, frequently defaulting to NO_RELATION label when an appropriate option is unavailable. While this behavior helps prevent incorrect relation assignments, our analysis reveals that it also leads to significant information loss when reasoning is not explicitly included in the output. We systematically evaluate this trade-off across multiple prompts, datasets, and relation types, introducing the concept of Hobson's choice to capture scenarios where models opt for safe but uninformative labels over hallucinated ones. Our findings suggest that conservative bias occurs twice as often as hallucination. To quantify this effect, we use SBERT and LLM prompts to capture the semantic similarity between conservative bias behaviors in constrained prompts and labels generated from semi-constrained and open-ended prompts.

1 Introduction

Recent advancements in LLMs have shown impressive ability to capture rich semantic knowledge and excel in tasks like text generation and question answering (Wadhwa et al., 2023a). As these models are increasingly deployed for complex natural language processing tasks, including relation extraction (Wadhwa et al., 2023b), distinct behavioral patterns have emerged that warrant careful examination.

One such pattern is hallucination, where LLMs generate content (or relations) beyond the provided context (or available options). This phenomenon has attracted enormous attention within the LLM community (Sriramanan et al., 2024; Zhang et al., 2024), as it is often perceived as a limitation in most applications. However, hallucination also presents opportunities for innovation, particularly in domains that benefit from creative generation such as image synthesis and other generative AI applications (Jiang et al., 2024).

Given the substantial research on hallucination detection (Yehuda et al., 2024; Li et al., 2024), we have observed a reduction in hallucination rates. This reduction has led us to explore other emergent behaviors of LLMs that may have significant downstream effects. We focused on relation extraction tasks using LLMs, where we initially anticipated some degree of hallucination. However, our findings revealed minimal occurrences of such behavior. Instead, we observed a distinct pattern where LLMs consistently exhibit a systematic bias towards classifying instances as NO_RELATION even when a more appropriate relation is available, presenting a "Hobson's choice" scenario. We attribute this behavior to alignment strategies designed to reduce hallucinations by reinforcing contextual adherence while suggesting external alternatives. We define this preference for overly cautious responses as ***Conservative Bias (CB)***.

In relation extraction tasks, LLMs exhibit a distinct CB, defaulting to the least incorrect classification when faced with uncertainty. Unlike hallucination, this bias leads to a unique form of information loss by creating "Hobson's choice" scenarios, where models favor safe but uninformative labels even when more suitable alternatives exist. This work addresses three key research questions: (1) How can we capture and quantify this CB? (2) What is its relationship to hallucination prevention? (3) How can we leverage this behavior to improve relation extraction tasks?

2 Related Work

Recent studies highlight emergent LLM behaviors like sycophancy and hallucination, which impact reliability and trustworthiness in downstream applications (Rrv et al., 2024). Sycophancy, where models tend to align responses with user views, ir-

LLM Output Analysis: The sentence states that *"Mr. Billingsley owned approximately 22.2% of the outstanding shares of Mentor Capital, Inc.'s Common Stock"*. This indicates that Mr. Billingsley has a stake or ownership in the company, but it does not necessarily mean he is an employee, founder, or member of the company. Therefore, none of the options 'founder_of', 'employee_of', 'member_of' accurately describe the relationship between Mr. Billingsley and Mentor Capital, Inc. However, since there is a clear relationship of ownership, the option 'no_relation' would also be incorrect. Unfortunately, the provided options do not include a relation class that accurately reflects this relationship. Given the provided options, the best choice would be 'no_relation' as it is the least incorrect. But please note, a more accurate relation class would be 'owner_of' or 'shareholder_of', if it was available in the options.

Relation type: Person–Organization

LLM Conclusion: NO_RELATION
LLM Suggestion: OWNER_OF OR SHAREHOLDER_OF

Figure 1: Example: LLM Output from REFinD dataset.

respective of objective correctness. (Sharma et al., 2023). This behavior is mostly prevalent in models whose fine-tuning employs human feedback and which can be mitigated using synthetic data (Wei et al., 2024).

Hallucinations have received much attention from the research community (Huang et al., 2025a; Sahoo et al., 2024; Huang et al., 2025b). Among proposed hallucination mitigation methods, Su et al. (2024) investigated LLM hallucinations in entity/relation extraction tasks proposing mitigating techniques. Advances in prompt engineering (Wadhwa et al., 2023a) have also mitigated hallucinations by constraining responses to given contexts (Sadat et al., 2023).

As prompt engineering advances, new emergent behaviors in LLMs may arise. To the best of our knowledge, Conservative Bias behavior has not been explored in existing literature.

3 Method

Our research aims to analyze CB in LLMs. We investigate the frequency with which LLMs default to the least incorrect labels from a list of options, as opposed to generating hallucinated relations. We analyze the rates of hallucination and CB across multiple prompt iterations. We also explore practical applications where CB can be utilized to refine relation classification, potentially expanding existing relations.

Formally, CB is detected in an output when: **(i.)** the model recognizes that a valid relation exists. **(ii.)** the correct relation type is not available in the option set. **(iii.)** the model chooses to default to NO_RELATION or selects the least incorrect (suboptimal) option. **(iv.)** the model demonstrates awareness of the correct relation through reasoning, suggesting it when appropriate to preserve the integrity of extracted relations. See example in Figure 1.

For evaluation purposes, we designed three types of prompts: Constrained Prompt, Semi-constrained Prompt, and Open-ended Prompt and assessed performance using four measures: Hobson's Choice Rate (**HCR**), Conservative Bias Rate (**CBR**), Hallucination Rate (**HR**) and New Relation Rate (**NRR**).

3.1 Prompting Design

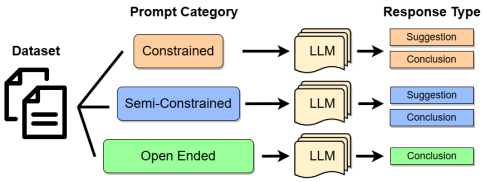

Figure 2: Process Workflow.

We adopted a multi-tiered approach to prompt design, where each level offers varying degrees of specificity to the LLMs. This approach explores how different levels of constraint affect the LLMs' ability to generate and select appropriate relations. The prompt categories are defined as follows:

Open-ended Prompts: represent the least constrained interaction with LLMs. In this setup, no predefined list of relation classes is provided. Instead, the LLMs are tasked with generating the most suitable relation between subject and object based on the input data.

Semi-Constrained Prompts: offer a moderate level of guidance. Here, LLMs are provided with a list of relations to choose from, which varies based on entity-pair type. However, the models retain the flexibility to propose a relation if none of the provided options are deemed most appropriate.

Constrained Prompts: are the most restrictive, requiring LLMs to select the best relation from a predefined list of options (relation classes). These prompts are designed to assess the LLMs' judgment and decision-making capabilities when faced with a limited set of possibilities.

By employing this tiered prompting strategy, as seen in Figure 2, we provide the LLMs with multiple perspectives before prompting them to select

a final label, aiming to provide enough context for detecting relations between subject and object that might be missed by a human labeler.

3.2 Metrics

HCR represents a scenario where a model selects the least incorrect option due to absence of a truly correct one, with CBR measuring how often LLM defaults to a conservative choice (e.g., NO_RELATION) despite recognizing a more appropriate but unavailable relation.

$$HCR = \frac{N^{HC}}{N^{total}},\ CBR = \frac{N^{CB}}{N^{total}} \quad (1)$$

where N_{HC} = Number of instances where model selects NO_RELATION (or suboptimal option) despite recognizing a valid relation, N_{CB} = Number of instances where model exhibits CB, meaning it chooses NO_RELATION (or least incorrect option) despite recognizing a valid relation.

The HR quantifies how often an LLM generates an unsupported or non-existent relation, while the NRR measures how often the model proposes a valid relation not present in the provided options, helping detect meaningful relations and justifying the correctness of CB.

$$HR = \frac{N^{H}}{N^{total}},\ NRR = \frac{N^{NR}}{N^{total}} \quad (2)$$

where N_H = Number of instances where the model hallucinates (i.e., generates a relation that is factually incorrect or not supported by the input data), where N_{NR} = Number of instances where the model suggests a valid relation that is not present in the predefined option set. N_{total} = Total number of relation extraction cases evaluated.

4 Experiments and Results

Data For our experiment, we focus on two datasets: REFinD (Kaur et al., 2023) (financial domain) and TACRED (Zhang et al., 2017) (general domain).[1] For our analysis, we focus on subset of data where gold_relation is labeled as 'no/other relation' or 'no_relation' and constitutes 45% of the REFinD and 79.5% of the TACRED dataset (statistics shown in App A.2).

Models We leveraged GPT-4 as our main LLM. We utilized two reduced temperature settings, specifically 0.2 and 0.5, and captured models' output consistency by conducting multiple iterations of each prompt by temperature settings. We also performed further analysis to see the behaviors in other LLMs, specifically Llama3.1-8B-Instruct.

Prompt Setup Prompts are structured in a hierarchical manner, allowing us to evaluate how varying level of constraints can affect LLMs' responses. While all prompts share the same basic structure, they differ in their option list setup. Among these, only the *constrained* prompt is prone to hallucination. Outputs from *semi-constrained* and *open-ended* prompts will be used to validate the CB behavior in the *constrained* prompt.

4.1 Results

4.1.1 Model Performance

Across different temperature settings and prompt configurations, we observe a range of outcomes when using the "step-by-step" instruction (Lightman et al., 2023). We analyze the outputs and categorize responses into 'conclusions' and 'suggestions' for both constrained and semi-constrained prompt responses. The results reveal distinct patterns in hallucination mitigation and the manifestation of CB.

On the REFinD dataset, GPT-4 outperformed Llama3.1, exhibiting a notably low HR of 0.02-0.04% and CBR of 1-1.33% for the constrained prompt. This pattern persists with the semi-constrained prompt, where we observe NRR of 7-10% and CBR of 37-41%.

Our analysis, summarized in Table 1, shows that the CBR can be more prevalent than the HR in relation extraction tasks. While GPT-4 demonstrated strong hallucination resistance under constrained prompting, the transition to semi-constrained prompt yielded interesting dynamics: although models showed an increased tendency to suggest novel relations when explicitly allowed, we observed a concurrent quadrupling (4x) of the CBR compared to NRR. In the semi-constrained scenarios, GPT-4 frequently generated novel relation suggestions but exhibited reluctance in conclusively asserting them (avoiding hallucinations), often defaulting to NO_RELATION or selecting from other predefined options.

To assess the semantic validity of CB labels identified in the constrained prompt, we conducted a semantic similarity analysis using outputs from semi-constrained and open-ended prompts. Focusing on instances flagged for CB, we found that over 57% of CB-flagged instances in the REFinD

[1] Dataset statistics can be obtained in their original papers.

dataset defaulted to Hobson's choice, as detailed in Table 1. These findings provide quantitative insights into the detection and measurement of CB in LLMs, addressing our primary research question regarding the characterization and quantification of this CB phenomenon.

Prompt	Dataset	Temp	CBR%	HR%	NRR%	HCR%
GPT-4						
Const.	REFinD	0.2	1.14	0.04	-	57.72
		0.5	1.33	0.06	-	64.37
	TACRED	0.2	7.99	15.47	-	1.23
		0.5	7.11	13.87	-	4.86
Semi	REFinD	0.2	37.67	-	9.75	69.15
		0.5	40.68	-	7.27	67.29
	TACRED	0.2	9.70	-	28.08	2.80
		0.5	9.20	-	27.04	10.46
Open	REFinD	0.2	-	-	81.66	-
		0.5	-	-	81.78	-
	TACRED	0.2	-	-	82.46	-
		0.5	-	-	76.96	-
Llama3.1						
Const.	REFinD	0.2	0.29	8.18	-	2.63
		0.5	1.07	4.67	-	1.44
Semi	REFinD	0.2	0.61	-	7.89	5.06
		0.5	3.78	-	10.83	4.26
Open	REFinD	0.2	-	-	66.19	-
		0.5	-	-	76.81	-

Table 1: Results for Constrained and Semi-Constrained prompt types on GPT-4 and Llama3.1-8B-Instruct.

4.1.2 Quality of LLM-Generated Relations

To evaluate the semantic quality of LLM-generated relations, we employ two methods: SBERT and GPT-4, employing a prompt instruction for LLM-based similarity assessment. All semantic similarity scores range from 0 to 1. We set our similarity threshold to 0.7 to align with established benchmarks (Okazaki and Tsujii, 2010)

Model	Dataset	κ	ρ
GPT-4	REFinD	0.65-0.77	0.66-0.79
	TACRED	0.30-0.53	0.33-0.54
Llama3.1	REFinD	0.31-1.0	0.32-1.0

Table 2: Inter Annotator Agreement Scores. Metrics: Cohen's Kappa (κ) and Spearman's Correlation Coefficient (ρ) on dataset per model for multiple runs.

Moreover, we calculated the inter-annotator agreement (IAA) to assess reliability, using Cohen's Kappa (κ) across multiple model runs at consistent temperature settings. Our results demonstrated substantial agreement (κ = 0.65-0.80), indicating significant reliability (McHugh, 2012) of our generated relations (particularly for GPT-4 relation extraction). Both Spearman's rank correlation and Cohen's Kappa lead to the same conclusion - higher reliability for GPT-4 relation extractions and lower for Llama3.1. When comparing both methods of similarity assessment in figure 2, the semantic similarity score from SBERT and GPT-4, the GPT-4 similarity scores appear to be higher.

5 Discussion

Our findings confirm the presence of CB tendencies in LLMsduring relation extraction. While GPT-4 demonstrates strong hallucination resistance under constrained conditions (0.02-0.04% HR), it also shows a much higher frequency of conservatism. This pattern persists in the semi-constrained design, suggesting a fundamental tension between innovation and accuracy in LLM behavior. In contrast, Llama3.1 shows less CB but a higher HR. This indicates that as models become more resistant to hallucinations, they tend to exhibit increased CB, presenting a crucial trade-off in model behavior that requires careful consideration in application design.

There are significant differences in output quality between GPT-4 and Llama3.1 when using identical prompts. Llama3.1 generated noisier outputs, often returning meta-responses such as *"Please specify title example"*, resulting in substantial data loss during the cleanup process. This disparity in output quality highlights the importance of model selection and prompt engineering in relation extraction tasks. To mitigate this limitation, our research indicates that detailed prompting strategies incorporating step-by-step reasoning are essential. This finding is particularly relevant in specialized professional contexts; for example - *A boutique law firm employing AI for litigation analysis*. Without structured reasoning steps in the prompting strategy, these systems risk returning conclusions that may be either overly conservative or inappropriately broad, potentially missing crucial legal nuances within the established constraints.

6 Conclusion & Future Work

This study explored the CB in LLMs during relation extraction, where models default to NO_RELATION when a correct option is unavailable. Our experiment confirmed an inverse relationship between CBR and HR, highlighting a trade-off between accuracy and innovation. Future research should focus on developing prompting strategies that balance CB with the need for novel relation identification, potentially by refining prompt designs and integrating external knowledge bases.

7 Limitations

While our work provides novel insights into CB detection in relation extraction tasks using LLMs, we acknowledge some limitations. We focused on two LLMs (GPT-4 and Llama3.1-8B-Instruct), limiting the generalizability of our findings across other models. Although we introduced metrics to quantify CB occurrences, there is a need for more robust evaluation frameworks to capture nuanced aspects of CB. Additionally, the quality of datasets used can significantly impact the results.

The study primarily relied on automated metrics for evaluation. Incorporating human evaluation could provide a more nuanced understanding of the quality and relevance of the extracted relations. Finally, as LLMs and their training data evolve, the behavior of models regarding CB and hallucination might change. The findings may need to be revisited with newer versions of models and updated datasets.

As this work represents one of the first systematic investigations of CB in relation extraction, our findings should be considered initial benchmarks rather than definitive measurements. We hope this paper will spur further research into CB detection and mitigation strategies in LLMs, extending beyond relation extraction tasks.

8 Ethics Statement

This research was conducted with a focus on ethical standards, particularly in addressing the CB in LLMs for relation extraction tasks. We used publicly available datasets, REFinD and TACRED, acknowledging potential biases inherent in them. Our study does not involve human subjects or personal data, minimizing privacy concerns. Our findings serve as initial benchmarks, and we encourage further research to explore ethical implications and enhance the social benefits of LLMs.

References

Rie Kubota Ando and Tong Zhang. 2005. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853.

Galen Andrew and Jianfeng Gao. 2007. Scalable training of L1-regularized log-linear models. In *Proceedings of the 24th International Conference on Machine Learning*, pages 33–40.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025a. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Trans. Inf. Syst.*, 43(2).

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025b. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Xuhui Jiang, Yuxing Tian, Fengrui Hua, Chengjin Xu, Yuanzhuo Wang, and Jian Guo. 2024. A survey on large language model hallucination via a creativity perspective. *Preprint*, arXiv:2402.06647.

Simerjot Kaur, Charese Smiley, Akshat Gupta, Joy Sain, Dongsheng Wang, Suchetha Siddagangappa, Toyin Aguda, and Sameena Shah. 2023. Refind: Relation extraction financial dataset. SIGIR '23, Taipei, Taiwan. Association for Computing Machinery.

Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. The dawn after the dark: An empirical study on factuality hallucination in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10879–10899, Bangkok, Thailand. Association for Computational Linguistics.

Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. *Preprint*, arXiv:2305.20050.

Mary McHugh. 2012. Interrater reliability: The kappa statistic. *Biochemia medica : časopis Hrvatskoga društva medicinskih biokemičara / HDMB*, 22:276–82.

Naoaki Okazaki and Jun'ichi Tsujii. 2010. Simple and efficient algorithm for approximate dictionary matching. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 851–859, Beijing, China. Coling 2010 Organizing Committee.

Mohammad Sadegh Rasooli and Joel R. Tetreault. 2015. Yara parser: A fast and accurate dependency parser. *Computing Research Repository*, arXiv:1503.06733. Version 2.

Aswin Rrv, Nemika Tyagi, Md Nayem Uddin, Neeraj Varshney, and Chitta Baral. 2024. Chaos with keywords: Exposing large language models sycophancy to misleading keywords and evaluating defense strategies. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12717–12733, Bangkok, Thailand. Association for Computational Linguistics.

Mobashir Sadat, Zhengyu Zhou, Lukas Lange, Jun Araki, Arsalan Gundroo, Bingqing Wang, Rakesh Menon, Md Parvez, and Zhe Feng. 2023. DelucionQA: Detecting hallucinations in domain-specific question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 822–835, Singapore. Association for Computational Linguistics.

Pranab Sahoo, Prabhash Meharia, Akash Ghosh, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024. A comprehensive survey of hallucination in large language, image, video and audio foundation models. *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11709–11724.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2023. Towards understanding sycophancy in language models. *Preprint*, arXiv:2310.13548.

Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. 2024. LLM-check: Investigating detection of hallucinations in large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Weihang Su, Yichen Tang, Qingyao Ai, Changyue Wang, Zhijing Wu, and Yiqun Liu. 2024. Mitigating entity-level hallucination in large language models. *Preprint*, arXiv:2407.09417.

Somin Wadhwa, Silvio Amir, and Byron Wallace. 2023a. Revisiting relation extraction in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15566–15589, Toronto, Canada. Association for Computational Linguistics.

Somin Wadhwa, Silvio Amir, and Byron C Wallace. 2023b. Revisiting relation extraction in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15566–15589.

Jerry Wei, Da Huang, Yifeng Lu, Denny Zhou, and Quoc V. Le. 2024. Simple synthetic data reduces sycophancy in large language models. *Preprint*, arXiv:2308.03958.

Yakir Yehuda, Itzik Malkiel, Oren Barkan, Jonathan Weill, Royi Ronen, and Noam Koenigstein. 2024. InterrogateLLM: Zero-resource hallucination detection in LLM-generated answers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9333–9347, Bangkok, Thailand. Association for Computational Linguistics.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. 2024. How language model hallucinations can snowball. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 59670–59684. PMLR.

Yuhao Zhang, Victor Zhong, Danqi Chen, Gabor Angeli, and Christopher D. Manning. 2017. Position-aware attention and supervised data improve slot filling. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, pages 35–45.

A Appendix

A.1 Semantic Similarity Scores

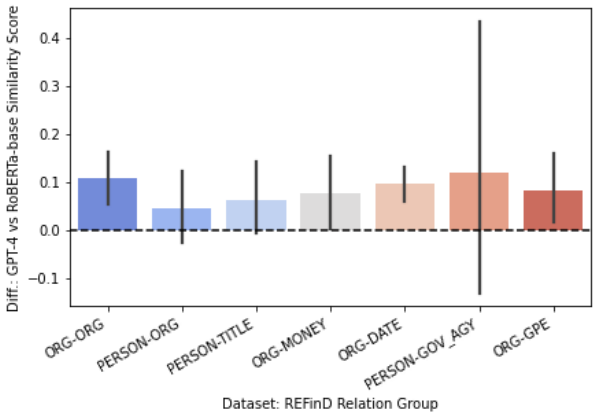

Figure A.3: REFinD: Difference in Semantic Similarity Scores (GPT4 vs SBERT).

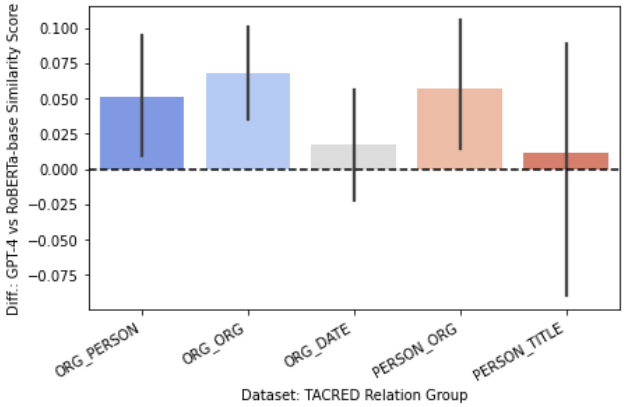

Figure A.4: REFinD: Difference in Semantic Similarity Scores (GPT4 vs SBERT).

Description	**Semantic Similarity**	**REFinD Semi**		**REFinD Open**		**TACRED Semi**		**Tacred Open**	
		>0.7	μ	**>0.7**	μ	**>0.7**	μ	**>0.7**	μ
Constrained Prompt Temp - 0.2	SBERT	34%	$0.54_{\pm 0.30}$	21%	$0.46_{\pm 0.25}$	4%	$0.30_{\pm 0.22}$	5%	$0.26_{\pm 0.22}$
	GPT-4 Prompt	62%	$0.44_{\pm 0.35}$	59%	$0.71_{\pm 0.22}$	11%	$0.35_{\pm 0.26}$	10%	$0.31_{\pm 0.25}$
Constrained Prompt Temp - 0.5	SBERT	41%	$0.55_{\pm 0.33}$	18%	$0.45_{\pm 0.25}$	5%	$0.25_{\pm 0.22}$	3%	$0.24_{\pm 0.20}$
	GPT-4 Prompt	59%	$0.65_{\pm 0.36}$	54%	$0.68_{\pm 0.24}$	8%	$0.30_{\pm 0.24}$	11%	$0.31_{\pm 0.25}$

Table A.4: Various semantic similarity scores from REFinD and TACRED based on prompt type.

A.2 Dataset Statistics

Dataset	Train	Dev	Test	Total
REFinD	9128	1965	1953	**13046**
TACRED	55112	17195	**12184**	84491

Table A.4: Datasets:No_Relation set