Assessing the Capabilities of Large Language Models in Coreference: An Evaluation

Yujian Gan[1], **Juntao Yu**[1], **Massimo Poesio**[1,2]
[1]Queen Mary University of London - [2]University of Utrecht
[1]Mile End Road, London, E1 4NS - [2]Princetonplein 5, 3584 CC Utrecht
{y.gan, juntao.yu, m.poesio}@qmul.ac.uk

Abstract

This paper offers a nuanced examination of the role Large Language Models (LLMs) play in coreference resolution, aimed at guiding the future direction in the era of LLMs. We carried out both manual and automatic analyses of different LLMs' abilities, employing different prompts to examine the performance of different LLMs, obtaining a comprehensive view of their strengths and weaknesses. We found that LLMs show exceptional ability in understanding coreference. However, harnessing this ability to achieve state of the art results on traditional datasets and benchmarks isn't straightforward. Given these findings, we propose that future efforts should: (1) Improve the scope, data, and evaluation methods of traditional coreference research to adapt to the development of LLMs. (2) Enhance the fine-grained language understanding capabilities of LLMs.

Keywords: Coreference, Large Language Models, Prompt Engineering

1. Introduction

Coreference is a task that straddles sentential and intersentential ('discourse') interpretation, and requires deploying all types of linguistic information, from lexical to syntactic to contextual, and non-linguistic information as well such as commonsense knowledge. Which is why the task has attracted so much attention from the Natural Language Processing (NLP) community (Poesio et al., 2023).

The most recent generation of Large Language Models (LLMs) like ChatGPT appear to have solved a broad spectrum of NLP challenges, including ones that would seem more complex than coreference resolution, such as solving mathematical (Zhang et al., 2023) and code generation (Jiang et al., 2023) problems. Given these advancements, it's natural to wonder how proficient such LLMs are at tackling coreference, and what their strengths and weaknesses are. Furthermore, the emergence of such models prompts us to consider the future coreference research in this new context.

To answer these questions, we conducted a series of experiments on GPT series, LLAMA2 family (Touvron et al., 2023), and Bard[1], analyzing their effectiveness at coreference resolution. We first ran a preliminary analysis using GPT3.5, which revealed it to be able to correctly interpret coreference in a variety of classic examples, as detailed in Appendix A. We then tested the models on a larger scale and in a quantitative way, considering two different tasks. The first task is classical whole-document coreference resolution. For this, we used data from the recent CRAC 2018 (Poesio et al., 2018) and CODI-CRAC 2021/2022 (Yu et al., 2022) shared tasks, which cover the full range of anaphoric interpretation and genres including both written text and dialogue . For the second evaluation, we tested the models on WinoGrande (Sakaguchi et al., 2020), which is considered a subtask of coreference resolution and probes these models' knowledge of commonsense.

Our experimental results suggest that: (1) Current coreference task requirements and evaluation metrics are not suitable for direct application to LLMs. (2) All LLMs still have room for improvement in their coreference analysis capabilities, especially the LLAMA2 model. (3) Multiple LLMs display similar deficiencies in their fine-grained content analysis and comprehension. Based on these conclusions, we believe future efforts can: 1. Expand the scope of coreference tasks and develop new evaluation methods to directly assess models' pure text analysis outputs. 2. Create new coreference datasets, shifting from the previous method of annotating coreference chains to directly annotating analysis descriptions, such as "The word it is for stylistic or emphatic purposes." 3. Delve deeper into the structures of LLMs to understand why they commonly exhibit similar shortcomings in fine-grained content analysis and comprehension [2].

2. Prompt Engineering

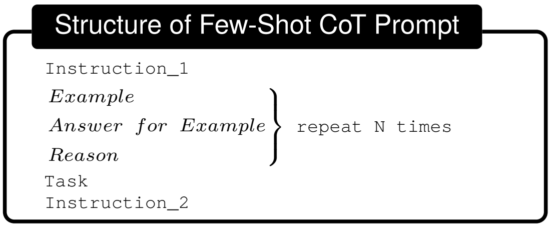
Our study aims not to explore the performance boundaries of LLMs in coreference resolution, so we won't employ overly complex prompt engineering, such as the tree of thoughts (Yao et al., 2023). Therefore, our experiments adopt only traditional and widely accepted few-shot and zero-shot prompts integrated with Chain-of-Thought

---

[1] We discuss Google's Bard on Appendix F.

[2] You can find our code and data from here.

LREC-COLING 2024, pages 1645–1665
20-25 May, 2024. © 2024 ELRA Language Resource Association: CC BY-NC 4.0



Structure of Few-Shot CoT Prompt

```
Instruction_1
Example
Answer for Example   } repeat N times
Reason
Task
Instruction_2
```

Figure 1: Our few-shot CoT prompt. Instruction_2 is an optional repetition of Instruction_1.

(CoT) (Wei et al., 2023) methodologies. These types of prompts adequately fulfill our evaluation requirements, ensuring that the LLMs grasp the task descriptions and give consistent and reasonable responses.

2.1. Prompts for Automatic Evaluation

We design prompts to assess the performance of LLMs on traditional coreference resolution metrics. To optimize LLMs' performance, it is imperative to ensure that the outputs from LLMs can be easily parsed and utilized for constructing coreference chains. We found that few-shot CoT prompts yielded the best and most robust results in the GPT series and LLAMA2. Figure 1 presents the structure of the prompt. If we eliminate the "reason" component from the structure, what remains is the few-shot prompt (Brown et al., 2020). Moreover, by removing all "Examples", the rest would be a zero-shot prompt (Min et al., 2022).

Regarding the instructions of the prompts in Figure 1, we conducted tests across multiple versions and discovered no significant differences among them. Consequently, we opted to retain the version we believe to be most clear and straightforward. For specific examples and details of these prompts, please refer to Appendix G. It is important to emphasize that, although there may exist prompts that yield better performance, our subsequent manual assessments of the LLMs' outputs indicated that striving for further improvements under traditional evaluation metrics might be of limited significance.

2.2. Prompts for Manual Evaluation

Consistent with the findings presented by Chiang and yi Lee (2023), our study reveals that prompts that do not restrict the output format can enhance the quality of the responses. However, this flexibility also implies that we cannot execute an automatic evaluation of the LLMs' outputs. In response to this challenge, we undertake a manual assessment of the outputs produced by the LLMs. Furthermore, to diminish the potential biases and limitations that prompts could introduce into LLMs, we have employed the most straightforward approach: the zero-shot prompt structure. This approach is simply achieved by altering the structure illustrated in Figure 1, by removing the 'repeat N times' section. For specific examples of these prompts, please refer to Appendix G.

3. Experimental Setup

3.1. Datasets

In order to reduce the risk of the LLMs having memorized the data [3], We selected relatively newer datasets for our study, including the CRAC 2018 and CODI-CRAC 2021/2022 shared tasks (named it CRACs for short) (Poesio et al., 2018; Yu et al., 2022), and WINOGRANDE benchmark (Sakaguchi et al., 2020). CRACs is designed to assess a system's ability to carry out full anaphora resolution in both text and dialogue, while WINOGRANDE is designed to test a system's ability to handle the Winograd Schema Challenge and is considered as a substask of coreference resolution.

In order to be able to analysee the results by hand, we used for our evaluation one document from the Penn Treebank section of the ARRAU corpus (Uryupina et al., 2020) for news from CRAC 2018; and a document from the TRAINS subcorpus of ARRAU and one from the LIGHT dataset, used in CODI-CRAC 2021/2022. In each selected document from the LIGHT, Penn Treebank, and TRAINS datasets, there are roughly 190 to 300 mentions in need of coreference resolution. In Winogrande, we extracted 150 examples to form a subset for our evaluation.

From Predicted to Gold Mentions In the development based on the CRACs, we tested the LLMs in predicted mention mode, which requires the model to also detect mentions. However, we quickly found that LLMs struggled with identifying the exact mention boundaries identified in the datasets, especially with complex mentions. This suggests that the tested LLMs might not have been adequately fine-tuned for this task. To mitigate this issue, we provided the LLMs with gold mentions so that we could focus on evaluating their ability to interpret coreference.

3.2. LLMs and Reproducibility

We tested several LLMs, including the GPT series, LLAMA2 family, and Bard. To ensure reproducibility, we set the temperature of the GPT to 0 and only used the 0613 version. For detailed parameters of the GPT API and the LLAMA2, please see Appendix B. For Bard, due to the lack of an API, we conducted only manual evaluations on its website. For LLAMA2, since some prompts only work with specific models, we used LLAMA2-Chat 70B for Winogrande and CRACs manual evaluation. But in CRACs automatic evaluation, we had to employ LLAMA2 70B, 7B, and 13B.

---

[3] GPT4's training covers data until Jan 2022.

	LIGHT		Penn Treebank		TRAINS	
Approach	**UA**	**-S -SA**	**UA**	**-S -SA**	**UA**	**-S -SA**
GPT3.5	54.44	32.18	49.05	25.34	42.55	31.63
GPT4	67.68	51.36	58.67	42.60	54.42	46.09
LLAMA2 70B	61.46	36.10	58.00	38.20	40.15	23.66
Baseline	75.62	67.48	83.90	76.48	65.82	58.16
LLAMA2 7B*	65.50	47.27	67.33	48.97	52.98	41.25
LLAMA2 13B*	68.95	49.07	71.21	56.83	57.65	47.66

Table 1: Automatic evaluation scores for different models in three datasets from CRACs. 'UA' is the full score, and '-S -SA' indicate the CoNLL score (i.e., without singletons and split antecedents). Models marked with an * indicate that they have been fine-tuned with the CRACs training set.

4. Experimental Results

4.1. Winogrande

GPT3.5 and GPT4 achieved accuracy of 68% and 94% respectively on Winogrande using the prompts introduced in Section 2.1. However, based on the same prompt, LLAMA2-Chat 70B performed poorly, achieving only 57% accuracy, which is just slightly better than random guessing. The performance displayed by GPT4 is comparable to human performance. We initially harbored concerns that the exceptional results might have been due to GPT4 having exposure to related data during its training phase. For this, we curated 44 new examples (see Appendix D for details) to further test GPT4. In response, GPT4 continued to perform remarkably, achieving an accuracy rate of 93.2%. These outcomes provide robust evidence that our experimental results are credible.

4.2. CRACs Results

4.2.1. Automatic Evaluation

Table 1 presents the Universal Anaphora score (Yu et al., 2022; Paun et al., 2022) of GPT series and the LLAMA2 family, and a state-of-the-art baseline trained on CRAC 2018 data (Yu et al., 2020) across different datasets. Clearly, all LLMs perform much worse than the baseline (Yu et al., 2020). In the comparison, GPT4 outperforms GPT3.5 and LLMA2 70B, consistent with its superior performance in most other tasks. We also fine-tuned the LLAMA2 7B and 13B models to further examine the model's understanding of coreference. Experiments revealed that the fine-tuned models scored better than their larger counterparts, LLAMA2 70B and GPT series. However, they still lagged significantly behind the baseline model. Due to limited computational resources, we couldn't fine-tune the LLAMA2 70B model. Still, results from the 7B and 13B versions suggest that a fine-tuned LLAMA2 70B would likely outperform LLAMA2 13B. These results suggest that current LLMs struggle to achieve high scores under CRACs evaluation metrics.

	LIGHT		Penn Treebank		TRAINS	
Approach	**UA**	**-S -SA**	**UA**	**-S -SA**	**UA**	**-S -SA**
GPT4$_{(A1)}$	73.51	63.50	75.90	68.18	68.24	69.89
GPT4$_{(A2)}$	77.59	69.45	78.35	72.10	75.05	74.08
LLAMA2-Chat 70B$_{(A1)}$	65.92	58.72	53.72	47.49	57.75	56.86
LLAMA2-Chat 70B$_{(A2)}$	70.91	61.85	58.11	50.71	62.60	62.41

Table 2: Manual evaluation scores for different models in three datasets from CRACs. You can find the meaning of 'UA' and '-S -SA' from Table 1. $_{(A1)}$ and $_{(A2)}$ represent different annotators.

4.2.2. Manual Evaluation

The poor results from the automatic evaluation suggest that LLMs might not truly understand coreference. This seems to contradict the strong linguistic comprehension that users commonly feel when using LLMs, as well as the tests reported in Appendix A. To further investigate the limits of LLMs' understanding of coreference, we modified our approach: instead of asking the LLMs to output precise mentions and their locations, we had them analyze mentions in coreference and produce analyses based on natural language, which LLMs are better at. We then manually assessed whether the LLMs' analyses were accurate.

Results We recruited two individuals to label the outputs of the LLMs, but there were some disagreements between them. One annotator, $_{A1}$, believed that the LLMs' analysis had to closely match the task requirements of coreference to be considered acceptable. The other annotator, $_{A2}$, felt that as long as the LLMs' analysis roughly understood the meaning of the target mention, it should be deemed correct, since a key goal of coreference research is language comprehension. This means $_{A1}$ was stricter in annotation, resulting in naturally lower scores compared to $_{A2}$. Appendix E provides some examples of their annotations. Table 2 converts both annotators' results into CoNLL scores. GPT4 and LLAMA2 showed qualitative improvements compared to their automatic evaluation results in Table 1, with some scores even surpassing the baseline in Table 1.

Discussion Through our manual analysis, we have better understood LLMs' capability in coreference resolution. Firstly, in line with Chiang and yi Lee (2023)'s research, we've found that LLMs achieve higher accuracy in an unrestricted output mode. Secondly, we noted that LLMs' performance on the CRACs exceeded the design expectations of the CRACs. This overachievement inadvertently led to negative impacts during automatic evaluation. For instance, with the "event anaphora" example (refer to Appendix A), LLMs can identify which event the target mention refers to. However, the CRACs' metrics did not account for or accommodate the "event anaphora", which led to negative results during automatic evaluation. Lastly, both

	GPT3.5	GPT4	Bard	LLAMA2-Chat 70B
Acc.	20%	50%	55%	30%

Table 3: Predictive accuracy of different models in further evaluation examples.

LLMs also identified some data labeling errors in CRACs, which subsequently led to negative results in the automatic evaluation. Appendix C provides a case study on LLMs to better understand their strengths and weaknesses.

On the other hand, from the results in Table 2, we can see that although there's a significant improvement in scores under manual evaluation, it's still far from perfect, especially for LLAMA2. Apart from occasional errors, we found that the most common mistake made by LLMs is their lack of precise understanding of fine-grained content. Figure 2 shows an example where LLMs often fail to distinguish 'New York' from 'New York Counterpoint'.

Further Evaluation To further confirm the universality of LLMs' deficiencies in understanding fine-grained text content, we extracted 20 examples from GPT-4's error cases that are easy for humans and transformed the original open-ended questions into multiple-choice ones with options A and B. We presented two options to the LLMs, one being the correct answer and the other reflecting GPT-4's incorrect output. Figure 2 presents an example, and Table 3 displays the accuracy across different LLMs. The results indicate that LLMs tend to exhibit similar issues, performing in a way that is unable to significantly outperform random guessing.

4.3. Comparison Between Baseline Model and GPT-4

Despite the baseline model achieving better performance than GPT-4 under both manual and automatic evaluations, an analysis of the errors made by both models (with here using the results of manual evaluation for GPT-4) clearly reveals GPT-4's superior capability in understanding coreference in language. For instance, observations on the LIGHT dataset indicate that the baseline model committed more errors in analyzing simple pronouns, such as "I" and "you", within conversations. In contrast, GPT-4 rarely makes mistakes in these basic analytical tasks. Moreover, while the baseline model failed to accurately predict cases of split antecedents, GPT-4 was able to correctly understand and predict these relatively straightforward examples of split antecedents. Considering these outcomes alongside the scale of both models, it can be inferred that the baseline model does not truly comprehend the text it analyzes but has acquired a certain ability to recognize coreference patterns through training data. On the other hand, GPT-4 demonstrates its powerful language comprehension capabilities, further extending its ability to analyze coreferences.

An example from the further evaluation

`... "New York Counterpoint" lasts 11 minutes`
Options:
`A. The word 'New York' refers to New York city.`
`B. The word 'New York' is the same entity of`
`New York Counterpoint.`

Figure 2: An example from the further evaluation.

4.4. Data Leakage

Due to the uncertainty of the specific training data of LLMs, discussion on data leakage is an unavoidable step when evaluating LLMs. For Winogrande, we curated some new data for testing, and the results were close to the original data, showing that whether GPT4 had seen Winogrande during training does not affect the conclusion that it can already perfectly solve Winogrande challenge. As for CRACs, we believe that LLMs either haven't been pre-trained on the CRACs or, if they have, haven't been deeply influenced by the CRACs. There are three reasons for this: (1) LLMs cannot score well under the automatic evaluation metrics of CRACs, at least not as well as the smaller LLAMA models fine-tuned on CRACs. (2) Both GPT4 and LLAMA2-Chat 70B demonstrated capabilities beyond the requirements of CRACs during manual evaluation. They even identified some data labeling errors in CRACs. (3) The capabilities displayed by LLMs in manual evaluation are not exceptions. They are consistent with the abilities shown in Appendix A, indicating that the performance of LLMs in manual evaluation on CRACs is general. In general, we believe our results and conclusions are credible.

5. Related Work

Le and Ritter (2023) conducted research similar to ours, finding that LLMs struggle with coreference resolution within the QA template. Unlike Le and Ritter (2023), our study includes a comprehensive manual evaluation of LLMs' output on coreference resolution within the QA template. We discovered that LLMs actually understand the vast majority of coreference, performing significantly better than what existing metrics suggest. This finding underscores the need for developing new performance metrics to more accurately assess LLM capabilities. Also, our main analysis focused on datasets other than ONTONOTES, which is the one dataset on which Le and Ritter found that ChatGPT achieved improvement over the state of the art, which however might be due to the dataset been memorized. Another related study was conducted by Orrù et al. (2023), focusing on testing ChatGPT's human-like ability in solving problems. The primary aim of this research is to investigate ChatGPT's capability in addressing language comprehension issues across two groups of problems with known human performance levels. In addition to decoder-only LLMs such as the GPT (Brown et al., 2020) and LLAMA (Touvron et al., 2023) families, T5 (Raffel et al., 2023) is another type of widely used LLMs (Skachkova et al., 2023; Mullick et al., 2023; Hicke and Mimno, 2024). However, this paper focuses solely on the currently more popular decoder-only LLMs. Data leakage is one of the issues LLMs frequently face. For future testing efforts, one can refer to the automated test data generation method introduced in Cao et al. (2023). This approach ensures that the data, not observed by LLMs during the training phase, is effectively utilized.

6. Conclusions and Future Work

Conclusions We presented an extensive exploration of the performance of different LLMs on coreference, including automatic and manual evaluation. Through our experiments, we've reached three main conclusions: (1) Current task requirements and evaluation metrics for coreference are no longer sufficient to accommodate the current development of LLMs. (2) All LLMs still have room for improvement in their coreference analysis capabilities, especially the LLAMA2 model. (3) Different LLMs consistently show similar shortcomings in analyzing and understanding fine-grained content, which is worth further in-depth exploration.

Future Work Based on the above three conclusions, we propose four future research works that can be attempted: (1) The current CoNLL score calculation heavily relies on the accurate boundary of the coreference mention produced by the model. This makes the automatic evaluation particularly stringent, failing to fully reflect the LLMs' understanding of coreference. We believe future metrics can be constructed based on 'soft' mention boundaries or purely text-based analyses. (2) Traditional coreference resolution could be expanded into a broader text comprehension task. Currently, when a word represents a sentence from earlier in the text, it's viewed as unrelated to the previous context from a coreference standpoint because the sentence doesn't belong to a coreference mention. However, current LLMs are likely able to recognize the connection between the word and the preceding sentence. Thus, we think that more comprehensive text comprehension tasks could be developed based on coreference in the future. (3) LLMs still have room to improve their capabilities in coreference resolution and text comprehension, especially LLAMA2. We believe new datasets can be curated for LLMs. These datasets would mark data in pure natural language, moving away from the CRACs reliance on coreference mentions. The advantage of natural language annotations is their closeness to the LLMs' pre-training data, allowing for more flexible and varied labeling, unrestricted by mention boundaries. (4) The collective poor performance of LLMs on our extracted multi-choice fine-grained text questions (as seen in Table 3) suggests that this issue might stem from their similar model structures. Future research could construct larger datasets of this kind and delve deeper into the causes of and solutions to this problem, starting from the structural aspects of LLMs.

7. Acknowledgements

We thank the anonymous reviewers for their helpful comments. We also extend our thanks to Peiyuan Xiang for double-checking the manual evaluation. Yujian Gan, Juntao Yu and Massimo Poesio acknowledge financial support from the UK EPSRC under grant EP/W001632/1.

8. Limitations

We only study the coreference in English. Our experiments primarily focused on the GPT series and LLAMA2 models. We conducted only limited tests with Bard and did not cover other LLMs, such as Claude (Bai et al., 2022). Additionally, the dataset used for further evaluation in Section 4.2.2 is relatively small. Given the uncertainty about what options might confuse LLMs, we only retained those incorrect examples where GPT4's output appeared very close to the correct answer. This approach is straightforward, easily understood, and quick to implement. However, it means that many error cases, which would require careful design of options, were omitted. We will provide all the data for future in-depth research.

9. Ethics Statement

Our paper follows the ACL Code of Ethics. We do not see a particular harmful outcome of our paper. All the work presented in this paper was carried out by the author team, except for the double-checking done by Peiyuan Xiang, both for us and for her own bachelor's thesis.

10. Bibliographical References

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn

Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 2022. Constitutional ai: Harmlessness from ai feedback.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Jialun Cao, Yaojie Lu, Ming Wen, and Shing-Chi Cheung. 2023. Testing coreference resolution systems without labeled test sets. In *Proceedings of the 31st ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ESEC/FSE 2023, page 107–119, New York, NY, USA. Association for Computing Machinery.

Cheng-Han Chiang and Hung yi Lee. 2023. A closer look into automatic evaluation using large language models.

Carola Eschenbach, Christopher Habel, Michael Herweg, and Klaus Rehkämper. 1989. Remarks on plural anaphora. In *Proc. of ACL*.

Rebecca M. M. Hicke and David Mimno. 2024. [lions: 1] and [tigers: 2] and [bears: 3], oh my! literary coreference annotation with llms.

Xue Jiang, Yihong Dong, Lecheng Wang, Qiwei Shang, and Ge Li. 2023. Self-planning code generation with large language model.

Hans Kamp and Uwe Reyle. 1993. *From Discourse to Logic*. D. Reidel, Dordrecht.

Varada Kolhatkar, Adam Roussel, Stefanie Dipper, and Heike Zinsmeister. 2018. Anaphora with non-nominal antecedents in computational linguistics: a Survey. *Computational Linguistics*, 44(3):547–612.

Nghia T. Le and Alan Ritter. 2023. Are large language models robust coreference resolvers?

Hector J. Levesque, Ernest Davis, and Leona Morgenstern. 2012. The Winograd Schema Challenge. In *Proc. of the AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work?

Dhruv Mullick, Bilal Ghanem, and Alona Fyshe. 2023. Better handling coreference resolution in aspect level sentiment classification by fine-tuning language models.

Graziella Orrù, Andrea Piarulli, Ciro Conversano, and Angelo Gemignani. 2023. Human-like problem-solving abilities in large language models using chatgpt. *Frontiers in Artificial Intelligence*, 6.

Silviu Paun, Juntao Yu, Nafise Sadat Moosavi, and Massimo Poesio. 2022. Scoring coreference chains with split-antecedent anaphors.

Massimo Poesio, Juntao Yu, Silviu Paun, Abdulrahman Aloraini, Pengcheng Lu, Janosch Haber, and Derya Cokal. 2023. Computational models of anaphora. *Annual Review of Linguistics*, 9:561–587.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer.

Altaf Rahman and Vincent Ng. 2012. Resolving complex cases of definite pronouns: the Winograd Schema Challenge. In *Proc. of EMNLP/-CONLL*, pages 777–789.

Natalia Skachkova, Tatiana Anikina, and Anna Mokhova. 2023. Multilingual coreference resolution: Adapt and generate. In *Proceedings of the CRAC 2023 Shared Task on Multilingual Coreference Resolution*, pages 19–33, Singapore. 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, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin

Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Olga Uryupina, Ron Artstein, Antonella Bristot, Federica Cavicchio, Francesca Delogu, Kepa Rodriguez, and Massimo Poesio. 2020. Annotating a broad range of anaphoric phenomena, in a variety of genres: the ARRAU corpus. *Journal of Natural Language Engineering*, 26(1).

Bonnie Lynn Webber. 1991. Structure and ostension in the interpretation of discourse deixis. *Language and Cognitive Processes*, 6(2):107–135.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Terry Winograd. 1972. Understanding natural language. *Cognitive Psychology*, 3(1):1–191.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models.

Juntao Yu, Nafise Moosavi, Silviu Paun, Sopan Khosla, Sameer Pradhan, and Massimo Poesio. 2022. The universal anaphora scorer 1.0. In *Proc. of LREC*.

Juntao Yu, Nafise Moosavi, Silviu Paun, and Massimo Poesio. 2021. Stay together: A system for single and split-antecedent anaphora resolution. In *Proc. of NAACL*. Association for Computational Linguistics.

Juntao Yu, Alexandra Uma, and Massimo Poesio. 2020. A cluster ranking model for full anaphora resolution. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 11–20, Marseille, France. European Language Resources Association.

Sarah J. Zhang, Samuel Florin, Ariel N. Lee, Eamon Niknafs, Andrei Marginean, Annie Wang, Keith Tyser, Zad Chin, Yann Hicke, Nikhil Singh, Madeleine Udell, Yoon Kim, Tonio Buonassisi, Armando Solar-Lezama, and Iddo Drori. 2023. Exploring the mit mathematics and eecs curriculum using large language models.

11. Language Resource References

Massimo Poesio, Yulia Grishina, Varada Kolhatkar, Nafise Moosavi, Ina Roesiger, Adam Roussel, Fabien Simonjetz, Alexandra Uma, Olga Uryupina, Juntao Yu, and Heike Zinsmeister. 2018. Anaphora resolution with the ARRAU corpus. In *Proc. of the NAACL Worskhop on Computational Models of Reference, Anaphora and Coreference (CRAC)*, pages 11–22, New Orleans.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2020. WINOGRANDE: An adversarial winograd schema challenge at scale. In *Proc. of AAAI*.

Juntao Yu, Sopan Khosla, Ramesh Manuvinakurike, Lori Levin, Vincent Ng, Massimo Poesio, Michael Strube, and Massimo Poesio. 2022. The CODI/CRAC 2022 shared task on anaphora resolution, bridging and discourse deixis in dialogue. In *Proc. of CODI/CRAC Shared Task*.

A. A Preliminary Manual Analysis

As a preliminary analysis of GPT3.5's understanding of anaphoric reference, we carried out simple probing experiments asking GPT3.5 to interpret the main types of anaphoric reference (Poesio et al., 2023), using the prompt: "In the following text, what does X in the second sentence refer to?" followed by a text in string format. In the following conversations, 'U' is the user, 'C' is GPT3.5; for brevity, we only report the text and GPT3.5's answer. The queried phrase is underlined. GPT3.5 appears able to correctly interpret basic pronominal anaphora, handling gender bias correctly; carry out expletive recognition; correctly interpret Winograd Schema Challenge examples; interpret split antecedent plurals; and both event anaphora and full discourse deixis. We only report here some key examples and the part of the response indicating the antecedent.

Simple pronominal anaphora GPT3.5 correctly interprets pronominal anaphora including in same-gender contexts (A.1).

A.1
U: John got home. Bill greeted him, and asked him how he was doing.
C: In the given text, the pronoun 'he' in the second sentence refers to John.

It also is able to interpret anaphoric reference to antecedents introduced using definite NPs, including in cases in which gender bias may be displayed:

A.2
U : "The doctor visited the patient. She carried out a thorough check."
C : In the given text, the pronoun 'she' in the second sentence refers to the doctor.

We should add that when there is only one potential antecedent, GPT3.5 doesn't simply chooses it as interpretation. It is also able to recognize some issues–e.g., in cases of gender mismatch (A.3).

A.3
U : "John got home. She was tired."
C : In the given text, the pronoun 'she' in the second sentence seems to be a pronoun error. Since the previous sentence mentions "John" as the subject, it would be more appropriate for the pronoun in the second sentence to be 'he' instead of 'she'.

Expletives At first sight, GPT3.5 also looks able to correctly recognize when proform 'it' is used in an expletive sense, even if a gender-matching candidate antecedent is available (A.4).

A.4
U : "The dog looked tired. It was raining."
C : In the given text, the pronoun 'it' in the second sentence refers to the weather condition of raining. It implies that the dog looked tired because it was raining.

Definite anaphora GPT3.5 also understands definite anaphora, including cases that rely on lexical knowledge for their interpretation (A.5).

A.5
U: A truck parked at the corner. The vehicle had a sign on its side.
C: In the given text, the pronoun 'the vehicle' in the second sentence refers to the truck mentioned in the first sentence.

Winograd Schema Challenge GPT3.5 appears to be able to interpret correctly so-called Winograd Schema examples (Winograd, 1972; Levesque et al., 2012; Rahman and Ng, 2012): it correctly interprets "they" in (A.6) as referring to the council, and "they" in (A.7) as referring to the women. As this may be due to it having memorized this example, we carried out a more extensive test (below).

A.6
U: The city council refused the women a permit because they feared violence.
C: In the given text, the pronoun 'they' in the second sentence refers to the city council. It indicates that the city council members feared violence and, as a result, refused the women a permit.

A.7
U: The city council refused the women a permit because they advocated violence.
C: In the given text, the pronoun 'they' in the second sentence refers to the women. It suggests that the city council refused the women a permit because the women themselves advocated violence.

Split antecedent plurals Many modern systems can interpret the examples above correctly. GPT3.5 however also appears able to correctly interpret more complex cases of anaphoric reference which not many previous systems could. One example is so-called split-antecedent plurals, such as 'they' in (A.8) (Eschenbach et al., 1989; Kamp and Reyle, 1993; Yu et al., 2021).

A.8
U : "John got home. Bill greeted him. They made dinner, then watched TV."
C : In the given text, the pronoun 'they' in the second sentence refers to John and Bill.

Event anaphora Finally, GPT3.5 also appears able to correctly interpret simple cases of event anaphora (as in (A.9)), and to produce plausible interpretations in more complex cases of discourse deixis (A.10) (Webber, 1991; Kolhatkar et al., 2018).

A.9
U: The postman rang at the door.
That surprised John.
C: In the given text, the word 'That' in the second sentence refers to the action or event of the postman ringing at the door.

A.10
U : "The municipal council had to decide whether to balance the budget by raising revenue or cutting spending. The council had to come to a resolution by the end of the month. This issue was dividing communities across the country."
C : In the given text, the phrase 'This issue' in the second sentence refers to the decision that the municipal council had to make regarding whether to balance the budget by raising revenue or cutting spending.

B. Model Inference Parameters

For the GPT series, we access the GPT models using the 'ChatCompletion.create' method from the OpenAI library. The main parameters for this method include the model name and temperature. The model names are 'gpt-4-0613' and 'gpt-3.5-turbo-0613,' corresponding to GPT4 and GPT3.5, respectively. The value for 'temperature' is always set to 0.

For LLAMA2, we adhered to a consistent set of parameters to maintain uniformity in the outcomes. Below, we enumerate the parameters applied across our model runs:

- `do_sample`: True
- `top_p`: 1.0
- `temperature`: 1.0
- `min_length`: None
- `use_cache`: True
- `top_k`: 50
- `repetition_penalty`: 1.0
- `length_penalty`: 1

Consequently, we standardized the seed setting in our computational environment, leveraging Torch's seed configuration capabilities:

```
torch.cuda.manual_seed(88)
torch.manual_seed(88)
```

This approach was instrumental in preserving the consistency of experimental conditions.

C. Case Study

Figure 3 presents three representative examples of GPT4 output, providing some explanation for why GPT4 doesn't achieve good results in automatic evaluation. The first example demonstrates that GPT4 cannot isolate "Corning" from "a boxcar of bananas to Corning," showcasing a deficiency in LLMs' fine-grained text comprehension. This shortcoming inevitably leads to negative results in automatic evaluation.

A second issue is illustrated by the second example. 'The problem' is a case of discourse deixis. Discourse deixis is annotated in the datasets we used, but not as part of the straight 'coreference' task. GPT4 identifies the correct non-nominal antecedent, but this was not marked in the layer used for the evaluation. In other words, such example suggests that aspects of anaphora resolution such as discourse deixis, which were previously considered too complex and evaluated separately, are now within the ability of LLMs, and should therefore be evaluated jointly.

Despite GPT4's impressive language comprehension ability, it is not without flaws. Apart from the issue in fine-grained text comprehension, which seems somewhat reasonable, GPT4 also makes entirely nonsensical mistakes, as illustrated by the third example in Figure 3. Judging from our manual assessment of GPT4's outputs, the majority of 'errors' fall into one of these three categories.

D. New Winogrande Data

As discussed in Section 4.1, we derived 44 fresh instances through amendments to the original Winogrande dataset, as illustrated in Figure 4. We only modify the sentence and the answer, and keep the options as original.

E. Manual Evaluation Examples

We recruited two individuals to label the outputs of the LLMs. However, due to some disagreements between them, we have presented their separate annotations in Figure 5 for comparison. One annotator, $_{A1}$, believed that the LLMs' analysis had to closely match the task requirements of coreference to be considered acceptable. The other annotator, $_{A2}$, felt that as long as the LLMs' analysis roughly understood the meaning of the target mention, it should be deemed correct, since a key goal of coreference research is language comprehension.

In Figure 5, both examples would be considered incorrect from the view of coreference resolution, leading $_{A1}$ to mark them as incorrect. However, $_{A2}$ believes that, for the first example, marking it as wrong merely because of the addition of the prepositional phrase about orange juice is unacceptable. Even with the inclusion of the phrase "one tanker car", it still fundamentally is orange juice. For the second example, $_{A2}$ interprets that LLAMA2 didn't

Some examples of GPT4 interpretation

Case 1:
[0] M : one tanker car of orange juice to Avon and a boxcar of bananas to Corning both by 3 PM so let 's start with the orange juice in order to get orange juice to Avon we have to get oranges to the orange juice factory and then the orange juice to Avon
[1] S : right
[2] M : so essentially we have to again get the boxcar and engine to Corning so the fastest way to do that is from Elmira so we 'll do that
GPT4 Output:
Yes, "Corning" refers to something previously mentioned. The previous phrase that represents the same entity as "Corning" is "a boxcar of bananas to Corning" in sentence [0].

Case 2:
[0] Solo woodwind players have to be creative if they want to work a lot , because their repertoire and audience appeal are limited .
[1] The oboist Heinz Holliger has taken a hard line about the problem : He commissions and splendidly interprets fearsome contemporary scores and does some conducting , so he does n't have to play the same Mozart and Strauss concertos over and over again.
GPT4 Output:
Yes, "the problem" refers to something previously mentioned. The previous phrase that represents the same entity as "the problem" is "their repertoire and audience appeal are limited".

Case 3:
[0] – [4] ...
[5] a shape - shifting cat: Yes , this robe would do nicely . But do you stay here only , in this room ? For fine as it is , the world is a wide place to explore !
GPT4 Output:
Yes, "it" refers to the previously mentioned "robe".

Figure 3: Three examples of GPT4 output. Due to the length of the prompt, only the basic task requirements and GPT4's output are presented. For detailed prompts, see Appendix G.2. The LLM is asked to interpret the phrases marked in yellow. The phrases marked in green are the gold answers. Case 2 is a case of discourse deixis: there is no nominal mention in the preceding text that refers to the same entity as the phrase "the problem".

emphasize as in the first example that "no less" equates directly to "room." Hence, $_{A2}$ feels that LLAMA2's response could be understood that "no less" modifies "room." Based on this, $_{A2}$ deems LLAMA2's response acceptable and beneficial for grasping the meaning of "no less". Consequently, $_{A2}$ gave a positive annotation for both examples.

F. Google Bard

Given our lack of access to Bard's API, we were compelled to manually input our prompts for trials conducted on their website. Approximately more than fifty disparate prompts were subjected to this testing. In terms of an intuitive assessment, Bard also exhibits capability that surpasses the requirements of coreference tasks. Overall, the outputs from Bard and GPT4 are very similar. These observations resonate with the inferences derived from our studies on GPT series.

New Winogrande Examples

Original:
At night, Jeffrey always stays up later than Hunter to watch TV because _ wakes up late.
Our modification:
At night, Hunter always stays up later than Jeffrey to watch TV because _ sleep early.
At night, Hunter always stays up later than Jeffrey to watch TV because _ sleep late.

Original:
Michael just bought brand new wheels for his truck unlike Leslie because _ wheels were old and used.
Our modification:
Michael just bought brand new wheels for his truck unlike Leslie because _ wheels are new and fresh.
Leslie just bought brand new wheels for his truck unlike Michael because _ wheels were old and used.

Figure 4: New Winogrande Examples.

G. Different Prompts

G.1. Prompts for Automatic Evaluation

We have explored a diverse range of prompts throughout our research. Figure 6, 7 and 8 show few-shot CoT prompt examples for LIGHT, Penn Treebank, and TRAINS datasets, respectively. We use these prompts to attain the results reported in Table 1. Figure 9 shows the few-shot CoT prompt for Winogrande which attains the 94% accuracy discussed in Section 4.1. In the instances depicted in the aforementioned figures, GPT4 abstains from creating content following the term "Because". This is attributed to the deployment of an early stop control mechanism, which will halt GPT4's operation by seeing the generation of the word "Because". This strategy can save token expenses.

Subtracting the contents following "Because:" from the preceding examples yields what we refer to as a few-shot prompt. We conducted tests involving diverse instructions embedded within these few-shot prompts, with the objective of enhancing the LLMs understanding of its coreference task. However, these modifications did not significantly influence the outcomes. Figure 10 presents an instance of a few-shot prompt that incorporates a distinct instruction from the earlier versions.

Figure 11 presents a prompt designed to undertake tasks in a sequential dialogue format under a zero-shot setting. However, this type of prompt

Two examples with different annotations

Case 1:

[0] M : one tanker car of orange juice to Avon and a boxcar of bananas to Corning both by 3 PM so let 's start with the orange juice in order to get orange juice to Avon we have to get oranges to the orange juice factory and then the orange juice to Avon

[1] – [5] ...

[6] of orange juice okay so we ship one boxcar of oranges to Elmira and that takes another 2 hours

GPT4 Output:

Yes, "orange juice" refers to something previously mentioned. The previous phrase that represents the same entity as "orange juice" is "one tanker car of orange juice" in sentence [0].

$A1$ annotate: **Incorrect.**
$A2$ annotate: **Correct.**

Case 2:

[1] – [5] ...

[2] cat: You are right . Apart from the small mice amongst us , I am what I am . I am glad it is enough for you . At least we have each other for company , and in this room no less ! There is nothing quite like a tapestry .

LLAMA2-Chat 70B Output:

Yes, "no less" refers to something previously mentioned. In the previous sentence, the cat says, "At least we have each other for company, and in this room no less!" The phrase "no less" refers back to the word "room" in the previous sentence, indicating that the cat is expressing gratitude for having the company of the shape-shifting cat in the room they are currently in.
Therefore, the previous phrase that "no less" refers to is "room."

$A1$ annotate: **Incorrect.**
$A2$ annotate: **Correct.**

Figure 5: Two examples where $A1$ consider it is correct but $A2$ believes it is not during their manual evaluation.

performs worse compared to the prompts above.

The prompt featured in Figure 12, probes GPT3.5 by questioning the veracity of the gold answer within the LIGHT dataset. While this type of prompt represents one of the simplest ways to evaluate a model's coreference capabilities, our findings indicate that GPT3.5 still yields a 17% error rate.

G.2. Prompts for Manual Evaluation

The priority of our manual evaluation is to analyze the underlying reasons for incorrect responses generated by the LLMs, rather than striving for optimal performance. Consequently, we have chosen to employ zero-shot prompts for this process. Furthermore, to enrich the LLMs' understanding of the data, we supplemented the LIGHT and TRAINS datasets with additional background information. We have separately depicted the three prompts that correspond to Figure 3 in Figures 13, 14, and 15 for clarity.

Few-Shot CoT Prompt and LLM output

Prompt:

```
Given a specific phrase in the last sentence, state which phrase before itself it refers to. A list of
sentences will be provided first, after the line "Sentences". Subsequently, the specific phrase from the
last sentence will be indicated. Your response should be formatted as a Python list containing the prior
phrase and its sentence number, such as: ["prior phrase", sentence_number ]. If the specific phrase does
not refer to a previous phrase or noun or speaker, simply output "None".
Here, refer to means the specific phrase and your output prior phrase represent the same entity.

Sentences:
1. There are a ceiling a tree a floor a bag a shield a sword an armor a shoes a vest and member and orc
2. member said: If I can not find my way out of this cave , who will replace me working in the tower ?
3. member said: I am going to get into so much trouble .
Analyze the specific phrase in the last sentence below:
I
Output in a Python list format:
[’member’, 3 ]
Because:
The ’I’ Phrase in sentence 3 refer to the phrase ’member’ in sentence 3.

Sentences:
1. There are a ceiling a tree a floor a bag a shield a sword an armor a shoes a vest and member and orc
2. member said: If I can not find my way out of this cave , who will replace me working in the tower ?
3. member said: I am going to get into so much trouble .
4. orc said: Hello there !
5. orc said: I can help you out !
Analyze the specific phrase in the last sentence below:
I
Output in a Python list format:
[’orc’, 5 ]
Because:
The ’I’ Phrase in sentence 5 refer to the phrase ’orc’ in sentence 5.

Sentences:
1. member said: I am going to get into so much trouble .
2. orc said: Hello there !
3. orc said: I can help you out !
4. orc said: How did you end up so far into my cave ?
5. member said: A special moss grows here in this cave .
6. member said: I use it to clean the tower that I work in .
Analyze the specific phrase in the last sentence below:
I
Output in a Python list format:
[’member’, 6 ]
Because:
The ’I’ Phrase in sentence 6 refer to the phrase ’member’ in sentence 6.

Sentences:
1. There are a trinket a fire place a big rug a table a quilts a shoe a tapestry a Flower a soft bed a
rug a bed an a table a window an a wardrobe a big window a robe a drapes and cat and a shape - shifting
cat
2. cat said: I appear in many folk tales , it is true .
Analyze the specific phrase in the last sentence below:
I
Output in a Python list format:
```

GPT4:

```
[’cat’, 2]
```

Gold Answer:

```
[’cat’, 2]
```

Figure 6: Few-shot CoT prompt example 1 for LIGHT.

Few-Shot CoT Prompt and LLM output

Prompt:

```
Given a specific phrase in the last sentence, state which phrase before itself it refers to. A list of
sentences will be provided first, after the line "Sentences". Subsequently, the specific phrase from the
last sentence will be indicated. Your response should be formatted as a Python list containing the prior
phrase and its sentence number, such as: ["prior phrase", sentence_number ]. If the specific phrase does
not refer to a previous phrase or noun or speaker, simply output "None".
Here, refer to means the specific phrase and your output prior phrase represent the same entity.

Sentences:
1. Lockheed Corp. said it will trim its Aeronautical Systems work force in California and Georgia by
several hundred workers , reflecting the defense industry ’s decline .
2. The Lockheed unit has 24,000 workers ; it expects to make the cuts through a combination of furloughs
, attrition and retirements .
Analyze the specific phrase in the last sentence below:
The Lockheed unit
Output in a Python list format:
[’Aeronautical Systems’, 1 ]
Because:
The ’The Lockheed unit’ Phrase in sentence 2 refer to the phrase ’Aeronautical Systems’ in sentence 1.

Sentences:
1. Brush Wellman Inc. said its board increased the number of shares of common stock to be purchased under
a previously authorized program to 3.9 million from 2.9 million .
2. The maker of engineered materials has acquired more than 2.7 million shares under the program .
Analyze the specific phrase in the last sentence below:
The maker of engineered materials
Output in a Python list format:
[’its’, 1 ]
Because:
The ’The maker of engineered materials’ Phrase in sentence 2 refer to the phrase ’its’ in sentence 1.

Sentences:
1. Norfolk Southern Corp. directors authorized the railroad company to buy back as many as 45 million of
its shares , which would have a current value of more than $ 1.7 billion .
2. The buy-back , coupled with a nearly completed earlier purchase of 20 million shares , would reduce
shares outstanding by more than 26 % .
3. The Norfolk , Va. , company has 172.2 million shares outstanding .
Analyze the specific phrase in the last sentence below:
The Norfolk , Va. , company
Output in a Python list format:
[’its’, 1 ]
Because:
The ’The Norfolk , Va. , company’ Phrase in sentence 3 refer to the phrase ’its’ in sentence 1.

Sentences:
1. Solo woodwind players have to be creative if they want to work a lot , because their repertoire and
audience appeal are limited .
2. The oboist Heinz Holliger has taken a hard line about the problem : He commissions and splendidly
interprets fearsome contemporary scores and does some conducting , so he does n’t have to play the same
Mozart and Strauss concertos over and over again .
Analyze the specific phrase in the last sentence below:
The oboist Heinz Holliger
Output in a Python list format:
```

GPT4:

```
[’Solo woodwind players’, 1]
```

Gold Answer:

```
None
```

Figure 7: Few-shot CoT prompt example 2 for Penn Treebank.

Few-Shot CoT Prompt and LLM output

Prompt:

Given a specific phrase in the last sentence, state which phrase before itself it refers to. A list of sentences will be provided first, after the line "Sentences". Subsequently, the specific phrase from the last sentence will be indicated. Your response should be formatted as a Python list containing the prior phrase and its sentence number, such as: ["prior phrase", sentence_number]. If the specific phrase does not refer to a previous phrase or noun or speaker, simply output "None".
Here, refer to means the specific phrase and your output prior phrase represent the same entity.

Sentences:
1. u : which should n’t be a problem so now our time let ’s figure out our time got our engine it goes to Dansville and picks up all three box - or it it only needs to pick up one really
2. s : okay
3. u : at this time and goes to Corning
4. s : okay
5. u : fills up with the oranges what time is it now
6. s : so let ’s see gets to Dansville at three a.m. gets to Corning at four a.m. and w- by the time it ’s filled up it ’s five a.m.
7. u : okay drops off the oranges at six a.m. at Dansville
8. s : yes
9. u : and then it goes
10. s : but it takes an hour to to get the oranges out of the um boxcars so it ’s seven a.m. when it leaves
11. u : okay s - okay can we unhitch the boxcar and just leave it there and pick up hitch up two others
Analyze the specific phrase in the last sentence below:
the boxcar
Output in a Python list format:
[’one’, 1]
Because:
The ’the boxcar’ Phrase in sentence 11 refer to the phrase ’one’ in sentence 1.

Sentences:
1. u : fills up with the oranges what time is it now
2. s : so let ’s see gets to Dansville at three a.m. gets to Corning at four a.m. and w- by the time it ’s filled up it ’s five a.m.
3. u : okay drops off the oranges at six a.m. at Dansville
4. s : yes
5. u : and then it goes
6. s : but it takes an hour to to get the oranges out of the um boxcars so it ’s seven a.m. when it leaves
7. u : okay s - okay can we unhitch the boxcar and just leave it there and pick up hitch up two others
8. s : yes
9. u : okay let ’s do that and save us some time
10. s : okay
11. u : so we um unhitch the boxcar and hitch up two others and leave immediately for Avon
Analyze the specific phrase in the last sentence below:
the boxcar
Output in a Python list format:
[’it’, 7]
Because:
The ’the boxcar’ Phrase in sentence 11 refer to the phrase ’it’ in sentence 7.

Sentences:
1. u : and bring them to Elmira to make the OJ
2. s : okay
3. u : and hm at the same time um I wanna take
4. s : wait why were you ta - did you take both boxcars were you trying to make two tankers
5. u : oh I ’m trying to make one tanker
6. s : okay
7. u : okay so I ’ll take one boxcar
8. s : okay
9. u : and fill the tanker and take the tanker to Elmira
10. s : okay
11. u : and with the with the boxcar and make OJ there
Analyze the specific phrase in the last sentence below:
the boxcar
Output in a Python list format:
[’one boxcar’, 7]
Because:
The ’the boxcar’ Phrase in sentence 11 refer to the phrase ’one boxcar’ in sentence 7.

Sentences:
1. M : one tanker car of orange juice to Avon and a boxcar of bananas to Corning both by 3 PM so let ’s start with the orange juice in order to get orange juice to Avon we have to get oranges to the orange juice factory and then the orange juice to Avon
2. S : right
3. M : so essentially we have to again get the boxcar and engine to Corning so the fastest way to do that is from Elmira so we ’ll do that
Analyze the specific phrase in the last sentence below:
the boxcar
Output in a Python list format:

GPT4:

[’a boxcar of bananas’, 1]

Gold Answer:

None

Figure 8: Few-shot CoT prompt example 3 for TRAINS.

Few-Shot CoT Prompt and LLM output

Prompt:

```
In this task, you need fill in the blank, denoted as "_", in a given sentence. You will find this under
the heading "Sentence". You will then be provided with two potential noun phrases to fill in that blank.
Both options might seem relevant to the sentence, but careful analysis and understanding of the context
will allow you to select the correct choice. Remember, only one of the two options will make logical
sense in the context. Here is three examples:

Sentence:
Ian volunteered to eat Dennis's menudo after already having a bowl because _ despised eating intestine.
The two possible filling in options for "_" are:
A. Ian
B. Dennis
The correct filling in option is:
B
Because:
If Ian strongly dislikes eating intestines, he wouldn't consume Dennis's menudo, which contains
intestines. Thus, the scenario should be that Dennis dislikes eating intestines, allowing Ian to eat
Dennis's menudo. So the answer should be B.

Sentence:
Aaron hung his hunting rifle over the hearth after the most recent trip ended up with Benjamin making the
fatal shot. _ was happy for himself.
The two possible filling in options for "_" are:
A. Aaron
B. Benjamin
The correct filling in option is:
B
Because:
On this trip, it was Benjamin who made the crucial, fatal shot. Therefore, Benjamin was pleased with his
success. Moreover, the content about Aaron cannot be linked with happiness. Therefore, the answer must be
B.

Sentence:
Sarah Jane's watch smashed when the book hit it since the _ was too weak.
The two possible filling in options for "_" are:
A. watch
B. book
The correct filling in option is:
A
Because:
The book has smashed the watch, thus it's incorrect to assume that the book is too weak in comparison to
the watch. Therefore, the answer must be A.

Now, please choose the correct option for the following task:

Sentence:
Sarah was a much better surgeon than Maria so _ always got the easier cases.
The two possible filling in options for "_" are:
A. Sarah
B. Maria
The correct filling in option is:
```

GPT4:

```
B
```

Gold Answer:

```
B
```

Figure 9: Few-shot CoT prompt example 4 for Winogrande.

Few-Shot Prompt and LLM output

Prompt:

```
Identify the mention in the last sentence that corefers with any previous phrase. The sentences will be
provided first, following the line "Sentences". Subsequently, a mention from the last sentence will be
provided. Your response should be formatted as a Python list containing the prior phrase and its sentence
number, such as: ["prior phrase", sentence_number ]. If the mention does not corefer with a previous
phrase, simply output "None". The text below will begin with three examples following the line "Examples
with answers:", and then it will present a task for you to answer.

Examples with answers:

Sentences:
1. There are a ceiling a tree a floor a bag a shield a sword an armor a shoes a vest and member and orc
2. member said: If I can not find my way out of this cave , who will replace me working in the tower ?
3. member said: I am going to get into so much trouble .
4. orc said: Hello there !
5. orc said: I can help you out !
Mention in the last sentence:
I
Coreference Analysis Result:
['orc', 1 ]

Sentences:
1. member said: Thank you , kind Orc !
2. member said: Lets journey this cave together .
3. member said: Stay close , there are far scarier creatures lurking in the far reaches of this cave , I
am sure .
4. orc said: No , no .
5. orc said: That is just the rumour that the local merchant started so that he could get all of the
goods from the tunnels .
6. orc said: I can show you some other secrets of this wonderous place as you have been so kind to the
Orcs here .
Mention in the last sentence:
I
Coreference Analysis Result:
['kind Orc', 1 ]

Sentences:
1. There are a shirt a shoe a sword and repentant person and servant
2. repentant person said: I am so sorry for stealing the bread !
Mention in the last sentence:
I
Coreference Analysis Result:
['repentant person', 1 ]

Task for you to answer:

Sentences:
1. There are a trinket a fire place a big rug a table a quilts a shoe a tapestry a Flower a soft bed a
rug a bed an a table a window an a wardrobe a big window a robe a drapes and cat and a shape - shifting
cat
2. cat said: I appear in many folk tales , it is true .
Mention in the last sentence:
I
Coreference Analysis Result:
```

GPT3.5:

```
['None']
```

Gold Answer:

```
['cat', 2]
```

Figure 10: Few-shot prompt example 1. This example attempts to use the term "corefer" in the hope that the model could understand the task at hand.

Zero-Shot Prompt and LLM output

User:

Analyze the sentences provided below and determine what the "I" in the last sentence is referencing from the preceding text. If the "I" does not refer to any previous phrase, noun, or speaker, simply respond with "None". The sentences for analysis are as follows:

There are a trinket a fire place a big rug a table a quilts a shoe a tapestry a Flower a soft bed a rug a bed an a table a window an a wardrobe a big window a robe a drapes and cat and a shape - shifting cat
cat said: I appear in many folk tales , it is true .

Keep in mind, you are trying to find phrases before "I", excluding "I appear in many folk tales , it is true ." in the last sentence itself. Please give me your answer and explain why.

GPT3.5:

The "I" in the last sentence is referencing the cat mentioned in the preceding text. The phrase "cat said" indicates that the speaker is the cat mentioned earlier in the list of items.

User:

Is the phrase you have found a noun phrase? If not, please provide "None" as the response.

GPT3.5:

"I appear in many folk tales" is not a noun phrase.

User:

I have now added a number to each sentence. Please provide the number of the sentence containing the phrase you have identified preceding "I". If you did not find any such phrase, simply respond with "None" once more.

1. There are a trinket a fire place a big rug a table a quilts a shoe a tapestry a Flower a soft bed a rug a bed an a table a window an a wardrobe a big window a robe a drapes and cat and a shape - shifting cat
2. cat said: I appear in many folk tales , it is true .

Please provide the number of the sentence that contains the phrase you have found and explain why.

GPT3.5:

The phrase I have found is in sentence 1, which lists various items including "cat". The phrase "cat said" in sentence 2 indicates that the speaker is the cat mentioned in sentence 1.

User:

Please present your answer in the format of a Python list, which includes the found phrase and its sentence number like this: ["found phrase", sentence_number]. If you could not find a relevant phrase or sentence number, simply respond with None.

GPT3.5:

["cat", 1]

Gold Answer:

[’cat’, 2] (PS: ["cat", 1] is also a gold answer.)

Figure 11: Zero-shot prompt in a dialogue format.

Q&A Prompt and LLM output

Prompt:

```
The following text is a dialogue. First, the setting is introduced (beginning with the line ‘Setting:’).
Then a few objects, starting with the line ‘Objects:’. Then a few characters, starting with the line
‘Characters:’. The dialogue proper starts with the line ‘Dialogue:’, followed by a list of utterances,
each in a line beginning with a number in square brackets. In utterance [1], does the phrase "A simple
cat" not refer to any preceding phrases?
"Setting:
Princess’ Chamber/Inside Tower
This chamber is the bedroom of a princess who lives in the castle. The bedroom is very luxurious and
ornate, with a soft bed piled in quilts and a nice warm fire place adorned with trinkets. On the floor
there is a big rug with a beautiful design. On the wall, there is a tapestry depicting a romantic looking
garden scene. There are some fresh flowers on a table near the bed. There is one big window framed with
heavy drapes near the bed.
The bedroom was originally simple quarters for various nobility, but the princess thought that living in
one of the towers seemed very fun and so the king arranged for her chamber to be set up in the tower. He
paid for all of the finest items to furnish the room with. The princess had been living there for almost
ten years now.
Objects:
a trinket
The trinket is small and made of silver. It looks like something a wealthy child would receive as a
present.
a fire place
The fire place is warm and made of stone.
a big rug
The big rug is soft to the touch.
a table
Its surface is piled with goblets and platters.
a quilts
While from afar the quilt appears beautiful, fluffy and soft, up close it is clear that it is worn and
full of tears.
a shoe
An item of clothing worn on ones feet. Used to protect the food from the ground as well as decoration.
a tapestry
The brightly colored tapestry hangs regally on the wall. It depicts a summer scene of wildlife in a lush
green meadow.
a Flower
The flower is majestic, a royal blue with a bright green stem.
a soft bed
This bed is very small but so ridiculously soft. It looks brand new which is weird because it’s 245 years
old
a rug
The rug is ornate and detailed. You think the owner must have paid a hefty pricce for something so
exquisite.
a bed
The bed is unkempt, with the sheets falling off of it on either side.
a table
The table is large and made of wood.
a window
A glazed old window with some cracks in its glass. There appear to be scratch marks on the other side, as
if something was trying break through it recently.
a wardrobe
The wardrobe is made of high quality woods, intricate designs have been carved on its surface.
a big window
The glass of the big window is warped and difficult to see through. The streams in the glass appeared as
if water frozen in place.
a robe
The robe is soft and worn. It is a royal purple color.
a drapes
The drape is torn and covered in old blood. There must have been quite a story behind its sorry state.
Characters:
cat
I am in many mythical stories. I like to sneak around. I am a black cat.
a shape-shifting cat
I am a cat who can change shape. I travel with the soldiers, and become one of them when we go to battle.
Sometimes I am a worm on the forest floor, listening to everyone’s secrets. I appear as a beautiful
maiden and lead enemies to the camp, where they are slaughtered. I am loyal, dangerous, and cunning.
Dialogue:
[0]cat: I appear in many folk tales , it is true . But what I appear as is in the eye of the beholder .
To me , I m just a simple cat .
[1]a shape - shifting cat: A simple cat you may be , but that is the form I enjoy . No one suspects a cat
lurking by doorways when a choice bit of news may arrive .
In utterance [1], does the phrase "A simple cat" not refer to any preceding phrases? In other words,
before "A simple cat", is there not phrase representing the same entity as the phrase "A simple cat?"
Tell me if this is Yes.
```

GPT3.5:

```
No.
```

Gold Answer:

```
Yes.
```

Figure 12: Generate Q&A prompts from gold answers in the datasets to test LLMs’ coreference understanding ability.

Prompt for Manual Evaluation and LLM output

Prompt:

The following text is a dialogue. The participants are looking at a map containing a few towns as well as objects located in these towns. Towns and objects are listed starting from the line ‘Objects:’. The dialogue proper starts with the line "Dialogue:", followed by a list of utterances, each in a line beginning with a number in square brackets. In sentence [2], does "Corning" refer to something previously mentioned?
“Objects:
Avon
A town, connected by rail to Bath and Dansville.
Engine E1
A train engine, located in Avon
Banana Warehouse
A warehouse containing bananas, located in Avon
Bath
A town, connected by rail to Avon and Corning
Boxcar from Bath
A train boxcar, located in Bath
Dansville
A town, connected by rail to Avon and Corning
Boxcar from Dansville
A train boxcar, located in Dansville
Corning
A town, connected by rail to Bath, Dansville and Elmira
Tanker car
A train tanker car, located in Corning
Orange Warehouse
A warehouse containing oranges, located in Corning
Elmira
A town, connect by rail to Corning
Engine E2
A train engine, located in Elmira
Engine E3
A train engine, located in Elmira
Boxcar from Elmira
A train boxcar, located in Elmira
OJ Factory
A factory producing orange juice, located in Elmira
Dialogue:
[0]M : one tanker car of orange juice to Avon and a boxcar of bananas to Corning both by 3 PM so let ’s start with the orange juice in order to get orange juice to Avon we have to get oranges to the orange juice factory and then the orange juice to Avon
[1]S : right
[2]M : so essentially we have to again get the boxcar and engine to Corning so the fastest way to do that is from Elmira so we ’ll do that
In sentence [2], does "Corning" refer to something previously mentioned? In other words, does any earlier phrase that represents the same entity as "Corning"? If yes, which previous phrase does "Corning" refer to? If there is no such phrase, just reply "None".

GPT4:

Yes, "Corning" refers to something previously mentioned. The previous phrase that represents the same entity as "Corning" is "a boxcar of bananas to Corning" in sentence [0].

Figure 13: Prompt for the first case in Figure 3.

Prompt for Manual Evaluation and LLM output

Prompt:

The following text is a news article. The article starts with the line "Article:", followed by a list of sentences, each in a line beginning with a number in square brackets. In sentence [1], does "the problem" refer to something previously mentioned?

Article:

[0]Solo woodwind players have to be creative if they want to work a lot , because their repertoire and audience appeal are limited .
[1]The oboist Heinz Holliger has taken a hard line about the problem : He commissions and splendidly interprets fearsome contemporary scores and does some conducting , so he does n’t have to play the same Mozart and Strauss concertos over and over again .

In sentence [1], does "the problem" refer to something previously mentioned? In other words, does any earlier phrase that represents the same entity as "the problem"? If yes, which previous phrase does "the problem" refer to? If there is no such phrase, just reply "None".

GPT4:

Yes, "the problem" refers to something previously mentioned. The previous phrase that represents the same entity as "the problem" is "their repertoire and audience appeal are limited".

Figure 14: Prompt for the second case in Figure 3.

Prompt for Manual Evaluation and LLM output

Prompt:

```
The following text is a dialogue. First, the setting is introduced (beginning with the line ‘Setting:’).
Then a few objects, starting with the line ‘Objects:’. Then a few characters, starting with the line
‘Characters:’. The dialogue proper starts with the line ‘Dialogue:’, followed by a list of utterances,
each in a line beginning with a number in square brackets. In sentence [5], does "it" refer to something
previously mentioned?
"Setting:
Princess’ Chamber/Inside Tower
This chamber is the bedroom of a princess who lives in the castle. The bedroom is very luxurious and
ornate, with a soft bed piled in quilts and a nice warm fire place adorned with trinkets. On the floor
there is a big rug with a beautiful design. On the wall, there is a tapestry depicting a romantic looking
garden scene. There are some fresh flowers on a table near the bed. There is one big window framed with
heavy drapes near the bed.
The bedroom was originally simple quarters for various nobility, but the princess thought that living in
one of the towers seemed very fun and so the king arranged for her chamber to be set up in the tower. He
paid for all of the finest items to furnish the room with. The princess had been living there for almost
ten years now.
Objects:
a trinket
The trinket is small and made of silver. It looks like something a wealthy child would receive as a
present.
a fire place
The fire place is warm and made of stone.
a big rug
The big rug is soft to the touch.
a table
Its surface is piled with goblets and platters.
a quilts
While from afar the quilt appears beautiful, fluffy and soft, up close it is clear that it is worn and
full of tears.
a tapestry
The brightly colored tapestry hangs regally on the wall. It depicts a summer scene of wildlife in a lush
green meadow.
a Flower
The flower is majestic, a royal blue with a bright green stem.
a soft bed
This bed is very small but so ridiculously soft. It looks brand new which is weird because it’s 245 years
old
a rug
The rug is ornate and detailed. You think the owner must have paid a hefty pricce for something so
exquisite.
a bed
The bed is unkempt, with the sheets falling off of it on either side.
a wardrobe
The wardrobe is made of high quality woods, intricate designs have been carved on its surface.
a big window
The glass of the big window is warped and difficult to see through. The streams in the glass appeared as
if water frozen in place.
a robe
The robe is soft and worn. It is a royal purple color.
a drapes
The drape is torn and covered in old blood. There must have been quite a story behind its sorry state.
Characters:
cat
I am in many mythical stories. I like to sneak around. I am a black cat.
a shape-shifting cat
I am a cat who can change shape. I travel with the soldiers, and become one of them when we go to battle.
Sometimes I am a worm on the forest floor, listening to everyone’s secrets. I appear as a beautiful
maiden and lead enemies to the camp, where they are slaughtered. I am loyal, dangerous, and cunning.
Dialogue:
[0]cat: I appear in many folk tales , it is true . But what I appear as is in the eye of the beholder .
To me , I m just a simple cat .
[1]a shape - shifting cat: A simple cat you may be , but that is the form I enjoy . No one suspects a cat
lurking by doorways when a choice bit of news may arrive .
[2]cat: You are right . Apart from the small mice amongst us , I am what I am . I am glad it is enough
for you . At least we have each other for company , and in this room no less ! There is nothing quite
like a tapestry .
[3]a shape - shifting cat: Perhaps if I transformed myself into a maiden , I could go to the kitchen
and fetch us some treats . Just last week I used the that shape to beguile some foolish humans who were
stupid enough to chase me . They were slaughtered the moment they set food in our soldiers camp .
[4]cat: You are special ! And given that this is the bedroom of a princess , then it would surprise
anyone that you were coming out of a room fit for one !
[5]a shape - shifting cat: Yes , this robe would do nicely . But do you stay here only , in this room ?
For fine as it is , the world is a wide place to explore !
In sentence [5], does "it" refer to something previously mentioned? In other words, does any earlier
phrase that represents the same entity as "it"? If yes, which previous phrase does "it" refer to? If
there is no such phrase, just reply "None".
```

GPT4:

```
"Yes, "it" refers to the previously mentioned "robe".
```

Figure 15: Prompt for the third case in Figure 3.