IdentifyMe: A Challenging Long-Context Mention Resolution Benchmark for LLMs

Kawshik Manikantan[1], **Makarand Tapaswi**[1], **Vineet Gandhi**[1], **Shubham Toshniwal**[2]
[1]CVIT, IIIT Hyderabad [2]NVIDIA

Abstract

Recent evaluations of LLMs on coreference resolution have revealed that traditional output formats and evaluation metrics do not fully capture the models' referential understanding. To address this, we introduce `IdentifyMe`, a new benchmark for mention resolution presented in a multiple-choice question (MCQ) format, commonly used for evaluating LLMs. `IdentifyMe` features long narratives and employs heuristics to exclude easily identifiable mentions, creating a more challenging task. The benchmark also consists of a curated mixture of different mention types and corresponding entities, allowing for a fine-grained model performance analysis. We evaluate both closed- and open-source LLMs on `IdentifyMe` and observe a significant performance gap (20-30%) between the state-of-the-art sub-10B open models *vs.* closed ones. We observe that pronominal mentions, which have limited surface information, are typically harder for models to resolve than nominal mentions. Additionally, we find that LLMs often confuse entities when their mentions overlap in nested structures. The highest-scoring model, `GPT-4o`, achieves 81.9% accuracy, highlighting the strong referential capabilities of state-of-the-art LLMs while also indicating room for further improvement. [1]

Instruction: Read the text given below. The text has an entity mention marked within “““ {{mention}} (#This is the marked mention) ”””. Extract the mention and find who/what the mention refers to in the text.

Text: The residence of Mr. Peter Pett , the well-known financier , on Riverside Drive is one of the leading eyesores of that breezy and expensive boulevard ......For the thousandth time he felt himself baffled by this calm , goggle-eyed boy who treated him with such supercilious coolness . “ You ought to be out in the open air this lovely morning , ” he said feebly . “ All right . Let 's go for a walk . I will if you will . ” “ I – I have other things to do , ” said Mr. Pett , recoiling from the prospect . “ Well , this fresh-air stuff is overrated anyway . Where 's the sense of having a home if you do n't stop in it ? ” “ When I was your age , I would have been out on a morning like this – er – bowling my hoop . ” “ And look at you now ! ” “ What do you mean ? ” “ Martyr to lumbago . ” “ I am not a martyr to lumbago , ” said Mr. Pett , who was touchy on the subject . “ Have it your own way . All I know is – ” “ Never mind ! ” “ I 'm only saying what mother . . . ” “ Be quiet ! ” Ogden made further researches in the candy box . “ Have some , pop ? ” “ No . ” “ Quite right . Got to be careful at your age . ” “ What do you mean ? ” “ Getting on , you know . Not so young as you used to be . Come in , pop , if you 're coming in . There 's a draft from that door . ” Mr. Pett retired , fermenting . He wondered how another man would have handled this situation . The ridiculous inconsistency of the human character infuriated him . Why should he be a totally different man on Riverside Drive from the person he was in Pine Street ? Why should he be able to hold his own in Pine Street with grown men – whiskered , square-jawed financiers – and yet be unable on Riverside Drive to eject **{{a fourteen-year-old boy}} (#This is the marked mention)** from an easy chair ? ......

Options:

Riverside Drive	Library
The Typewriter Girl	Mr. Pett's Room
Mr. Peter Pett's Residence	Ogden Ford
Elmer Ford	Mr. Peter Pett
Mrs. Pett	$\boxed{\text{None of the Above}}$

Answer: Ogden Ford

Figure 1: Sample instance from the validation set of `IdentifyMe`. The mention of interest is highlighted in the text. The answer options include frequently occurring entities in the text, and None of the Above.

1 Introduction

Coreference Resolution (CR) consists of identifying the entity mentions and clustering them based on the entity identity. It is a fundamental task for text comprehension and can therefore be used to assess a model's textual understanding. While LLMs have made tremendous strides on a wide array of NLP tasks (Brown et al., 2020; OpenAI, 2024a; Gemini Team et al., 2024), their performance on CR has been relatively underwhelming. It remains uncertain to what extent this is due to the LLMs' weak referential abilities, as traditional coreference setups—both datasets and metrics—require LLMs to adhere to varying definitions of mentions, boundaries, and entities across datasets.

For instance, Le and Ritter (2023) report that on document-level coreference annotation, LLMs perform well at mention linking but struggle with mention detection, particularly due to varying definitions of what constitutes an entity and how mention boundaries are defined. While Manikantan et al. (2024) mitigate the variability of entity definition by assuming major entities as inputs, their

[1]Code for the paper is available at:
https://github.com/KawshikManikantan/IdentifyMe

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 768–777

evaluation remains limited by dataset-specific span boundaries. Recent work by Gan et al. (2024) demonstrates through manual analysis that LLMs perform markedly better when evaluated in an unrestricted output mode. This suggests that traditional evaluations may underestimate LLMs' coreference capabilities, highlighting the need to adapt traditional CR datasets and metrics to better assess LLMs.

Along these lines, we introduce the `IdentifyMe` benchmark for mention resolution in a multiple-choice question (MCQ) format. The MCQ format is commonly used in large language model (LLM) evaluations (Hendrycks et al., 2021) and offers two key advantages. First, its widespread presence in pretraining datasets enables LLMs to answer questions in this format effectively. Second, it eliminates the need for exact antecedent span identification during mention resolution evaluation, thus mitigating errors caused by dataset-specific annotation choices.

To construct the benchmark, we use annotations from two long-text coreference benchmarks, namely LitBank (Bamman et al., 2020) and FantasyCoref (Han et al., 2021). To make the benchmark challenging, we restrict it to pronominal and nominal mentions and apply heuristics for each mention type to filter out easily resolvable cases (Section 2.1). Each MCQ instance consists of text marked with the mention of interest and choices comprising frequently occurring entities in the text and the *None of the Above* (NoA) option. Fig. 1 shows an example in `IdentifyMe`, derived from LitBank.

We evaluate both closed- and open-source LLMs with the following key findings:

- Among the mention types, LLMs perform worse on pronominal mentions (which have limited surface information) than on nominal mentions.
- The instances where *None of the Above* is the correct answer prove particularly challenging for all the models, with open-source models experiencing a performance drop of more than 50%.
- With nested mentions, LLMs frequently confuse entities with overlapping mentions (*e.g.*, $\boxed{\boxed{\text{his}}\text{ mother}}$).
- The highest-scoring model `GPT-4o` scores 81.9% on `IdentifyMe`, highlighting the strong performance of *frontier* LLMs while indicating scope for further improvement in referential capabilities.

2 `IdentifyMe` Benchmark

`IdentifyMe` is an MCQ-based benchmark where, given a text document with a marked mention, the task is to identify the entity the mention refers to. We derive these mentions from two coreference datasets focused on literary texts: LitBank and FantasyCoref. These datasets provide long contexts (1700 words on average for FantasyCoref and 2000 words for LitBank) and feature multiple entities with rich inter-dependencies (*e.g.*, *Mr. and Mrs. Pett*) that make resolving mentions more challenging. While LitBank offers diverse writing styles and linguistic structures, FantasyCoref includes entities that often take on different forms (*e.g.*, disguises and transformations), or undergo title change (*e.g.*, *Prince Rudolph* is called *The Emperor* after his coronation), which further complicates entity mapping.

Coreference annotations cluster mentions that refer to the same entity, but creating an MCQ requires a representative phrase for each entity cluster. We use `GPT-4o-mini` (see Table 9) to generate these phrases based on the mentions and their frequencies. The generated annotations undergo manual review to ensure each entity has a distinct representative phrase.

To prevent confusion, we discard and avoid labeling clusters that: (i) contain annotation errors (*e.g.*, due to cluster merging or splitting (Kummerfeld and Klein, 2013)); (ii) are too small (< 3 mentions) or difficult or ambiguous to label (*e.g.*, entities like *some money*); (iii) are plural, as they often lack explicit surface forms that can be derived from mentions.

An MCQ is created from a document using mentions from labeled clusters, with all labeled entities provided as options. To ensure benchmark quality, we exclude short context documents (< 1000 words) or those where the discarded entities represent more than 50% of the mentions.

2.1 Selecting Mentions for `IdentifyMe`

Based on previous works which utilize rule-based linguistic patterns to perform (Zhou and Su, 2004; Lee et al., 2013) or analyze (Haghighi and Klein, 2009; Otmazgin et al., 2023) coreference resolution, we propose a two-step heuristic to identify challenging mentions.

Step 1: Discard easy mentions. We apply two criteria to filter out mentions that can be easily resolved due to syntactic similarity:

Nominal fuzzy score: We calculate the fuzzy similarity[2] between a nominal mention and its entity's representative phrase, allowing for variations in word order and subsets. We discard mentions with similarity scores above 75%, as these cases typically provide obvious surface-form clues for identification.

Net distractor score: We categorize pronominal mentions based on attributes like gender, number, and animacy (LingMess (Otmazgin et al., 2023)). For a candidate marked pronominal mention, nearby pronouns of the same category that refer to the same entity can provide disambiguating context. However, pronouns that either share the category but refer to different entities, or refer to the same entity but have different categories, can increase ambiguity. We define the `Net-Distractor-Score` as the difference between the count of ambiguity-increasing and disambiguating neighboring pronouns. We discard mentions with non-positive scores (≤ 0).

Step 2: Ranking mentions by difficulty. Filtered mentions are ranked from most to least difficult: for nominals, a low `Nominal-Fuzzy-Score` is preferred; and for pronouns, a high `Net-Distractor-Score` is preferred. Additionally, the distance between the marked mention and other mentions of the same entity also indicate difficulty. We consider distances to the nearest mention, the nearest nominal mention, and the nearest mention resembling the representative phrase as further criteria for ranking. All the individual criteria are combined using Copeland's method (Copeland, 1951), evaluating pairwise wins and losses to determine the final ranking.

2.2 Dataset Statistics

`IdentifyMe` comprises the 1800 most challenging questions based on our ranking method, drawn from 159 documents (64 from LitBank, 95 from FantasyCoref). We randomly select 600 of these questions as a validation set for prompt tuning and ablation experiments. Each question includes a *None of the Above (NoA)* option to encourage more confident entity selection. To test NoA detection, we remove the correct entity from 10% of the questions, making NoA the correct choice. Both validation and test splits maintain balance across source datasets and mention types (pronominals and nominals).

[2] https://github.com/seatgeek/thefuzz

Model	Random (10 runs)	`IdentifyMe` (Val.)
`Mistral-7B`	64.8 ± 2.1	55.3
`GPT-4o-mini`	70.5 ± 1.9	63.3
`GPT-4o` *	83.8	80.7

Table 1: Performance of models on the `IdentifyMe` validation set vs. comparable-sized evaluation set consisting of randomly chosen mentions (repeated 10 times).

Model/Approach	Accuracy
`Mistral-7B`	46.0
`Llama-3.1-8B`	50.0
`GPT-4o-mini`	62.0
`Gemini-1.5-Flash`	66.0
`GPT-4o`	70.0
Human-1	92.0
Human-2	94.0

Table 2: Performance of various models and human annotators on a subset of 50 questions from `IdentifyMe`.

2.3 Does `IdentifyMe` have Hard Mentions?

We conduct an ablation experiment to assess the effectiveness of our mention selection process. As a baseline, we randomly sample mentions and evaluate model performance on their identification. The performance drops of 9.5% for `Mistral-7B` and 7.2% for the more robust `GPT-4o-mini` demonstrate that `IdentifyMe` captures more challenging mentions compared to random sampling (see Table 1).

2.4 Human Evaluation on `IdentifyMe` Subset

We perform human evaluation on a randomly selected subset of 10 FantasyCoref documents from the test split of `IdentifyMe`. A set of 50 mention resolution questions are extracted from these documents, comprising 25 nominals and 25 pronominal mentions. As seen in Table 2, there is a significant performance gap of $\sim$23% between humans and the best performing LLM, `GPT-4o`. This confirms that there is substantial scope for improvement and `IdentifyMe` poses a challenge to current LLMs.

3 Experiments

Models. Among closed-source models, we evaluate `GPT-4o` (OpenAI, 2024a),

Model	w/o CoT	w/ CoT
Mistral-7B	**55.3**	53.8
Llama-3.1-8B	50.2	**59.7**
GPT-4o-mini	63.3	**67.0**

Table 3: Validation accuracy of LLMs w/ and w/o CoT.

Model	Total (1200)	Nominal (600)	Pronominal (600)
Random	8.0	7.6	8.5
Mistral-7B	51.5	52.5	50.5
Llama-3.1-8B	53.3	53.2	53.5
GPT-4o-mini	63.3	67.7	59.0
Gemini-1.5-Flash	73.9	77.7	70.0
GPT-4o	**81.9**	**85.2**	**78.7**

Table 4: Performance of various models on the IdentifyMe test set.

GPT-4o-mini (OpenAI, 2024b), and Gemini-1.5-Flash[3] (Gemini Team et al., 2024). Due to computational constraints, we limit the evaluation of open-source models to sub-10B variants: Llama-3.1-8B (Meta-AI, 2024) and Mistral-7B (Jiang et al., 2023).

MCQ setup. The selected mention is highlighted in the original text by enclosing it with special tokens (*e.g.* "... eject *a fourteen-year old boy* from ..." → "... eject {{*a fourteen-year old boy*}} (#This is the marked span) from ...". A zero-shot prompt instructs the model to retrieve and resolve the mention and identify who or what it refers to from a given set of entities and NoA (detailed prompt in Appendix A.3).

Inference details. For open-source models, we use regex-based constrained decoding with the outlines library (Willard and Louf, 2023) to limit answers to specific entity representative phrases. We also experiment with a chain-of-thought (CoT) approach (Wei et al., 2023), instructing the model to explain its reasoning before answering the question. As seen in Table 3, using CoT improves the model performance (*e.g.*, +9.5% for Llama-3.1-8B, +3.7% for GPT-4o-mini). Based on these results, we use the CoT decoding for evaluation over the test set. For details on prompts used and decoding regular expressions, see Appendix A.3.

[3]Due to safety filters, evaluated on 1197 questions

Model	Nominal FC (300)	Nominal LB (300)	Pronominal FC (300)	Pronominal LB (300)
Mistral-7B	39.0	66.0	51.7	49.3
Llama-3.1-8B	42.3	64.0	55.0	52.0
GPT-4o-mini	60.7	74.7	63.3	54.7
Gemini-1.5-Flash	72.1	83.3	73.7	66.3
GPT-4o	**79.3**	**91.0**	**81.3**	**76.0**

Table 5: Performance split by mention type and dataset source. FC: FantasyCoref, LB: LitBank.

3.1 Results

Table 4 presents the overall LLM performance on the IdentifyMe test set, along with a breakdown by nominal and pronominal mention types. The Random baseline, where answers are uniformly randomly chosen, achieves 8% on our benchmark. Although all LLMs outperform the Random baseline, open-source models show considerable room for improvement, with Llama-3.1-8B reaching only 53.3% accuracy. GPT-4o is the top-performing model with an accuracy of 81.9%. Meanwhile, GPT-4o-mini, an affordable closed-source option, surpasses smaller open-source models but lags behind top performers like GPT-4o and Gemini-1.5-Flash. Across mention types, all closed-source models perform significantly better at resolving nominal mentions than pronominal ones.

Table 5 presents the performance split across mention types and source datasets. For nominal mentions, the FantasyCoref (FC) instances are, on average, considerably more challenging than those from LitBank (LB). This could be because of the higher surface similarity across FantasyCoref entities (*e.g. The eldest princess*, *The youngest princess*). In contrast, LitBank's pronominal mentions are harder to resolve than FantasyCoref's, possibly due to its complex linguistic structure.

3.2 Error Analysis

Comparing entities *vs.* NoA. Table 6 provides the accuracy distribution when the correct option is an entity (Ent) *vs.* NoA. Furthermore, we classify errors into three categories: (a) ground truth is an entity and the model chooses another entity (Ent-Ent), (b) ground truth is an entity, but the model predicts NoA (Ent-NoA), and (c) ground truth is NoA, but the model chooses an entity (NoA-Ent). Open-source models perform extremely poorly on the NoA subset (120 MCQs), leading to high NoA-Ent errors. Among closed-source models,

Sample Error by GPT-4o

. . . “M’ama ... non m ’ ama ... ” the prima donna sang , and “ M’ama ! ” , with a final burst of love triumphant , as she pressed the dishevelled daisy to her lips and lifted her large eyes to the sophisticated countenance of the little brown Faust-Capoul , who was vainly trying , in a tight purple velvet doublet and plumed cap , to look as pure and true as {{ his artless victim}} (#This is the marked mention) . Newland Archer , leaning against the wall at the back of the club box , turned his eyes from the stage and scanned the opposite side of the house . . .

Ground Truth: Madame Nilsson
Predicted Answer: M. Capoul

Figure 2: An error by GPT-4o in resolving a nested mention where the model incorrectly resolves *his artless victim* to the entity referred to by *his* i.e. *M. Capoul*.

Model	Accuracy Ent	NoA	#Misclassifications Ent-Ent	Ent-NoA	NoA-Ent
Mistral-7B	57.0	1.7	453	11	118
Llama-3.1-8B	59.2	0.8	438	3	119
GPT-4o-mini	63.4	62.5	221	174	45
Gemini-1.5-Flash	78.6	30.3	192	38	83
GPT-4o	**82.9**	**73.3**	**135**	**50**	**32**

Table 6: Left: Model accuracy for MCQs with correct answer as an entity (Ent, 1080 samples) *vs*. NoA (120 samples). Right: Number of misclassifications within entities (Ent-Ent) or with NoA (Ent-NoA, NoA-Ent).

Gemini-1.5-Flash achieves sub-par performance on NoA MCQs ($\downarrow$ 48.3%) and prefers to select an entity when the answer is NoA (83/120). Interestingly, GPT-4o and GPT-4o-mini are much more resilient on NoA questions, with drops of only $\downarrow$ 9.6% and $\downarrow$ 0.9%, respectively.

Nested mentions. The dataset contains 352 instances of nested mentions, where the span of one mention overlaps with another. Table 7 shows that the accuracy of nested mentions is comparable to the overall accuracy. However, when models err in resolving these mentions, about 40% of these errors are because the predicted entity corresponds to an overlapping mention. Figure 2 illustrates a sample nested mention error made by GPT-4o.

Model	Accuracy Non-nested	Nested	Span Error
Mistral-7B	50.1	54.8	40.3
Llama-3.1-8B	53.2	53.7	42.9
GPT-4o-mini	60.8	69.3	34.3
Gemini-1.5-Flash	73.3	75.1	36.8
GPT-4o	82.1	81.5	47.7

Table 7: LLM performance on nested mentions (352 of 1200) versus non-nested mentions. The Span Error column indicates the error for nested mentions where the predicted entity corresponds to an overlapping mention.

4 Conclusion

We present IdentifyMe, a challenging MCQ benchmark designed for the evaluation of the referential capabilities of LLMs. Our analysis reveals several key challenges for LLMs, including: (i) pronominal resolution which has limited surface form information, (ii) questions where “None of the Above" is the correct answer, and (iii) nested mentions that require distinguishing between overlapping spans. GPT-4o scores 81.9% on IdentifyMe, highlighting the strong referential capabilities of *frontier* LLMs while still leaving ample room for improvement. We believe the IdentifyMe benchmark, with its curated mix of diverse and challenging mentions, will serve as an effective tool for fine-grained assessment of state-of-the-art LLMs’ referential capabilities.

5 Limitations

The IdentifyMe has several limitations: it covers only the literary domain, includes only nominal and pronominal mentions, and excludes plural entities. The source datasets we used are publicly available, and our preliminary investigations suggest limited contamination risk, as none of our evaluated LLMs could accurately reproduce the original CoNLL annotations for complete stories. While we significantly transformed the original coreference annotations to construct our benchmark, we acknowledge the potential possibility of data contamination.

References

David Bamman, Olivia Lewke, and Anya Mansoor. 2020. An Annotated Dataset of Coreference in English Literature. In *LREC*.

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. In *NeurIPS*.

A. Copeland. 1951. A Reasonable Social Welfare Function. In *Seminar on Applications of Mathematics to Social Sciences*.

Yujian Gan, Massimo Poesio, and Juntao Yu. 2024. Assessing the Capabilities of Large Language Models in Coreference: An Evaluation. In *LREC-COLING*.

Gemini Team, Petko Georgiev, and Ving Ian Lei. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context.

Aria Haghighi and Dan Klein. 2009. Simple Coreference Resolution with Rich Syntactic and Semantic Features. In *EMNLP*.

Sooyoun Han, Sumin Seo, Minji Kang, Jongin Kim, Nayoung Choi, Min Song, and Jinho D. Choi. 2021. FantasyCoref: Coreference Resolution on Fantasy Literature Through Omniscient Writer's Point of View. In *Fourth Workshop on Computational Models of Reference, Anaphora and Coreference*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring Massive Multitask Language Understanding. In *ICLR*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Jonathan K. Kummerfeld and Dan Klein. 2013. Error-Driven Analysis of Challenges in Coreference Resolution. In *EMNLP*.

Nghia T. Le and Alan Ritter. 2023. Are Large Language Models Robust Coreference Resolvers? *arXiv preprint arXiv:22305.14489*.

Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2013. Deterministic Coreference Resolution Based on Entity-Centric, Precision-Ranked Rules. *Computational Linguistics*.

Kawshik Manikantan, Shubham Toshniwal, Makarand Tapaswi, and Vineet Gandhi. 2024. Major Entity Identification: A Generalizable Alternative to Coreference Resolution. In *EMNLP*.

Meta-AI. 2024. The Llama 3 Herd of Models.

OpenAI. 2024a. GPT-4 Technical Report.

OpenAI. 2024b. GPT-4o-mini: Advancing Cost-Efficient Intelligence.

Shon Otmazgin, Arie Cattan, and Yoav Goldberg. 2023. LingMess: Linguistically Informed Multi Expert Scorers for Coreference Resolution. In *EACL*.

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. In *NeurIPS*.

Brandon T Willard and Rémi Louf. 2023. Efficient Guided Generation for LLMs. *arXiv preprint arXiv:2307.09702*.

GuoDong Zhou and Jian Su. 2004. A High-Performance Coreference Resolution System using a Constraint-based Multi-Agent Strategy. In *COLING*.

Error in a Dialog by GPT-4o

. . . " Well , Watson , what do you make of it ? " Holmes was sitting with his back to me , and I had given him no sign of my occupation . " How did you know what I was doing ? I believe you have eyes in the back of **{{your}} (#This is the marked mention)** head . " " I have , at least , a well-polished , silver-plated coffee-pot in front of me , " said he . . .

Ground Truth: Sherlock Holmes
Predicted Answer: Dr. John Watson

Figure 3: A sample error made by GPT-4o where Sherlock Holmes and Dr. John Watson are engaged in a dialog. The instance is particularly hard because the dialog speakers are not marked and need to be inferred.

A Appendix

A.1 Model Details

Table 8 presents the precise model identifies used in this work.

Model	Model Identifier
Mistral-7B	mistralai/Mistral-7B-v0.3
Llama-3.1-8B	meta-llama/Llama-3.1-8B
GPT-4o-mini	gpt-4o-mini-2024-07-18
Gemini-1.5-Flash	gemini-1.5-flash-001
GPT-4o	gpt-4o-2024-08-06

Table 8: Details of all the models evaluated in the paper.

A.2 Sample Model Outputs

Figure 3 presents another example error where the model is confused due to complicated first and second person references within dialog.

Figure 4 and 5 present instances which both GPT-4o and Llama3.1 get right. Their explanation makes sense as well.

A.3 Prompts

Table 9 presents the prompt passed to GPT-4o-mini to generate representative phrases for each cluster/entity. Table 10 and Table 11 present the prompts and regular expressions used for inference with and without CoT respectively.

Correct Resolution by GPT-4o and Llama-3.1

. . . John had not much affection for his mother and sisters , and an antipathy to me . He bullied and punished me ; not two or three times in the week , nor once or twice in the day , but continually : every nerve I had feared him , and every morsel of flesh in my bones shrank when he came near . There were moments when I was bewildered by the terror he inspired , because I had no appeal whatever against either his menaces or his inflictions ; the servants did not like to offend {{their young master}} (#This is the marked mention) by taking my part against him , and Mrs. Reed was blind and deaf on the subject : she never saw him strike or heard him abuse me , though he did both now and then in her very presence , more frequently , however , behind her back . . .

Ground Truth: John Reed

GPT-4o
Explanation: The mention "their young master" refers to John Reed, as he is the one who is described as being feared by the servants and is the son of Mrs. Reed, making him the young master of the household.
Predicted Answer: John Reed

Llama-3.1
Explanation: The mention refers to the person who is being obeyed by the servants, and is the one who bullies and punishes the narrator, Jane Eyre. The mention is used to describe the relationship between the servants and John Reed.
Predicted Answer: John Reed

Figure 4: Sample instance from IdentifyMe that both GPT-4o and Llama-3.1 get right.

Correct Resolution by GPT-4o and Llama-3.1

. . . And when once Grethel was inside , she intended to shut the oven and let her bake in it , and then she would eat her , too . But Grethel saw what she had in her mind , and said , " I do not know how I am to do it ; how do {{you}} (#This is the marked mention) get in ? " " Silly goose , " said the old woman , " The door is big enough ; just look , I can get in myself ! " and she crept up and thrust her head into the oven . . . .

Ground Truth: **The Old Woman**

GPT-4o
Explanation: In the context, "you" is used by Grethel to ask the old woman how she would get into the oven, referring directly to the old woman.
Predicted Answer: The Old Woman

Llama-3.1
Explanation: The mention refers to the old woman, who is the wicked witch, and is instructing Grethel on how to get into the oven to cook her brother Hansel. She is trying to trick Grethel into getting into the oven herself.
Predicted Answer: The Old Woman

Figure 5: Sample instance from IdentifyMe that both GPT-4o and Llama-3.1 get right.

Instruction
You are provided with information about entities in a document. For each entity, you are given a list of different mentions, along with the number of occurrences of each mention in the format mention (count). Derive an appropriate representative label for each entity from the given mentions. Use the following guidelines: - Prefer names over other noun phrases (nominals). - If the entity appears to be a narrator but lacks a specific name, label the entity as "Narrator". - Ensure the label is as precise and descriptive as possible. - Avoid removing possessive pronouns from the representative label if they are included. - Do not produce any other extra text. Follow the below format: Entity 0: Label 0 Entity i: Label i
Example Input:
Information: Entity 0: i(34), me(17), my(9), myself(3), ishmael(1), my soul(1) Entity 1: the most absent-minded of men(1), that man(1) Entity 2: an artist(1) Entity 3: the commodore on the quarter-deck(1), their leaders(1) Entity 4: your insular city of the manhattoes(1), the city of a dreamy sabbath afternoon(1) Entity 5: the poor poet of tennessee(1) Entity 6: the world(2), this world(1) Entity 7: cato(1) Entity 8: this shepherd(1), the shepherd(1) Entity 9: narcissus(1)
Example Output:
Entity 0: Ishmael Entity 1: The Most Absent-Minded Man Entity 2: An Artist Entity 3: The Commodore Entity 4: City of the Manhattoes Entity 5: The Poor Poet of Tennessee Entity 6: The World Entity 7: Cato Entity 8: The Shepherd Entity 9: Narcissus

Table 9: The zero-shot prompt passed to GPT-4o-mini to generate representative phrases for each cluster/entity.

Instruction
Read the text given below. The text has an entity mention marked within """ {{mention}} (#This is the marked mention) """. Extract the mention and find who/what the mention refers to in the text.
Example Input:
Text: Chapter 1 It is a truth universally acknowledged, that a single man in possession of a good fortune, must be in want of a wife. However little known the feelings or views of such a man may be on his first entering a neighbourhood, this truth is so well fixed in the minds of the surrounding families, that he is considered the rightful property of some one or other of their daughters. "My dear Mr. Bennet," said his lady to him one day, . . . Chapter 2 Mr. Bennet was among the earliest of those who waited on Mr. Bingley. He had always intended to visit him, though to the last always assuring {{his wife}} (#This is the marked mention) that he should not go; and till the evening after the visit was paid she had no knowledge of it. It was then disclosed in the following manner. Observing his second daughter employed in trimming a hat, he suddenly addressed her with: "I hope Mr. Bingley will like it, Lizzy." "We are not in a way to know what Mr. Bingley likes," said her mother resentfully, "since we are not to visit" . . . I do not know how you will ever make him amends for his kindness; or me, either, for that matter. At our time of life it is not so pleasant, I can tell you, to be making new acquaintances every day; but for your sakes, we would do anything. Lydia, my love, though you are the youngest, I dare say Mr. Bingley will dance with you at the next ball. **Options for the answer:** Mary Kitty Mrs. Bennet Mrs. Long Elizabeth Mr. Bingley Mr. Bennet Lydia Netherfield Park None of the Above Note that you can select the None of the Above option (The mention refers to: None of the Above), if the mention does not refer to any other entity/option. Also provide explanations in 1-2 sentences for the same. Do not produce any other extra text. Follow the below format: - Mention: - Explanation: - The mention refers to:
Decoding Regex (Constrained Decoding):
- Mention: \{{[A-Za-z ,\'\.]{1,125}\}} - Explanation: \{{[A-Za-z ,\'\.]{150,350}\}} - The mention refers to: (Mary|Kitty|Mrs. Bennet|Mrs. Long|Elizabeth|Mr. Bingley| Mr. Bennet|Lydia|Netherfield Park|None of the Above)
Example Output:
- Mention: his wife - Explanation: The mention refers to Mrs. Bennet. The pronoun 'his' refers to Mr. Bennet, and 'wife' refers to the person who is married to Mr. Bennet. So, the mention refers to Mrs. Bennet. - The mention refers to: Mrs. Bennet

Table 10: QA prompt with CoT used in the test set experiments.

Instruction
Read the text given below. The text has an entity mention marked within """ {{mention}} (#This is the marked mention) """. Extract the mention and find who/what the mention refers to in the text.
Example Input:
Text: Chapter 1 It is a truth universally acknowledged, that a single man in possession of a good fortune, must be in want of a wife. However little known the feelings or views of such a man may be on his first entering a neighbourhood, this truth is so well fixed in the minds of the surrounding families, that he is considered the rightful property of some one or other of their daughters. "My dear Mr. Bennet," said his lady to him one day, . . . Chapter 2 Mr. Bennet was among the earliest of those who waited on Mr. Bingley. He had always intended to visit him, though to the last always assuring {{his wife}} (#This is the marked mention) that he should not go; and till the evening after the visit was paid she had no knowledge of it. It was then disclosed in the following manner. Observing his second daughter employed in trimming a hat, he suddenly addressed her with: "I hope Mr. Bingley will like it, Lizzy." "We are not in a way to know what Mr. Bingley likes," said her mother resentfully, "since we are not to visit" . . . I do not know how you will ever make him amends for his kindness; or me, either, for that matter. At our time of life it is not so pleasant, I can tell you, to be making new acquaintances every day; but for your sakes, we would do anything. Lydia, my love, though you are the youngest, I dare say Mr. Bingley will dance with you at the next ball. **Options for the answer:** Mary Kitty Mrs. Bennet Mrs. Long Elizabeth Mr. Bingley Mr. Bennet Lydia Netherfield Park None of the Above Note that you can select the None of the Above option (The mention refers to: None of the Above), if the mention does not refer to any other entity/option. Do not produce any other extra text. Follow the below format: - Mention: - The mention refers to:
Decoding Regex (Constrained Decoding):
- Mention: \{{[A-Za-z ,\'\.]{1,125}\}} - The mention refers to: (Mary|Kitty|Mrs. Bennet|Mrs. Long|Elizabeth|Mr. Bingley|Mr. Bennet|Lydia|Netherfield Park|None of the Above)
Example Output:
- Mention: his wife - The mention refers to: Mrs. Bennet

Table 11: QA prompt without CoT.