Coreference Without Bells and Whistles

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science
in
Computer Science and Engineering
by Research

by

Kawshik Manikantan Sundar
2019111004
kawshik.manikantan@research.iiit.ac.in

International Institute of Information Technology
Hyderabad - 500 032, INDIA
May, 2025

Copyright © Kawshik Manikantan Sundar, 2025
All Rights Reserved

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled "Coreference Without Bells and Whistles" by Kawshik Manikantan Sundar, has been carried out under my supervision and is not submitted elsewhere for a degree.

Date

Advisors: Prof. Vineet Gandhi and
Prof. Makarand Tapaswi

To All Who Willed This

Acknowledgements

"Which is more important," asked the Big Panda, "the journey or the destination?"
"The company," said the Tiny Dragon

When I began my research, I sought out something challenging, which is why CVIT, propelled by the tales I heard, was my top choice. I never expected it to be the only thing that would stick with my expectations. Take this not lightly as I never imagined myself exploring language models and linguistic challenges amidst discussions of diffusion and convolutions. After a journey filled with significant highs and even greater lows, I am pleasantly surprised to be writing this acknowledgment this early.

I've had the privilege of working with three advisors who complement one another, and together they have fully shaped me technically. I am grateful to Prof. Vineet Gandhi for introducing me to Occam's razor, the principle that the simplest solution is often the best, through his work. His strong belief in work ethic overall (and specifically in mine) is the reason I've been granted so much leeway on the multitude of days when I didn't produce any *"visibly publishable results"*.

No words would suffice to describe the impact Prof. Makarand Tapaswi has had on my technical and professional development. I deeply admire his curiosity, attention to detail, and process-oriented approach to work and life, and I aspire to embody these qualities one day. I genuinely feel that reaching out to Dr. Shubham Toshniwal marked one of the most crucial moments in my research journey. He has always been an approachable senior I have wished for a long, alongside a greatly insightful mentor. His beam of optimism has been a guiding light during my *"dark"* days.

My grateful thanks to my labmates for all the conversations and discussions we have had over the years. I am especially grateful to BVK for his unwavering support and talks during our late visits to JC. He is deeply missed, and I pray that his soul rests in peace. I also sincerely thank Varun for his timely assistance during my rebuttal phase. Most of the enjoyable moments during my college years, including the late-night movies I watched, the restaurants and places I visited, and all my conversations, were shared with my group of friends: Tanvi, Gayatri, Gokul, Akarsh, Ahana, Nikit, and Sriram. To Tanvi, a special thanks for always being there to lift my spirits, whether it was in the middle of the night or at the breakfast table. Shoutout to Gayatri for being a great friend, part-time teacher and student; our sessions together to understand core concepts were very crucial. I am grateful to Sreevatsa for his unwavering support during critical times, often staying up late just to be there for me. My heartfelt thanks to my *"extended"* labmates- Aayushi, Bharat, Harshit, Pragati and Supreet - for making me feel at home on numerous occasions.

Words cannot fully convey my gratitude to my family, who have been equally, if not more, engaged in my research journey. It is no exaggeration when I say that my mom, Meenakshi, deserves to have her name alongside mine on all my papers. She has patiently listened to, understood, and occasionally critiqued every one of my ideas, from the dumbest to the one that eventually got published. Without her love and sacrifices, neither I nor my family would be where we are today. My dad, Sundar, has been the foundation upon which me and my family grew. His love for his children and his steadfast, quiet presence instils in us the confidence to excel. My sister, Niveadhitha is someone I wouldn't hesitate calling at 1 AM at night for help. She is a pillar of support and I know I can rely on her for everything and forever. I would like to thank my grandmothers, my relatives and my teachers whose blessings and prayers have constantly pushed me forward. A hurray to Abhishek and Arjun for being the great sport they are!

I pass on to anyone who stumbles upon this thesis, a lesson my mom taught me and which has become a fundamental part of who I am. A Thirukkural in my mother-tongue Tamil,

"Deiyvaththaan Aagaa Theninum Muyarchithan
Meyvaruththa Kooli Tharum "

"Even if destined by fate not to succeed,
Persistent effort shall endure rewards."

Abstract

Bells and Whistles *(noun, informal)*: *items or features that are useful but not essential.*

Coreference resolution (CR) is the task of identifying text spans that refer to the same entity. It is a fundamental component of natural language understanding with applications in various downstream NLP tasks, such as question answering, knowledge graph construction, and summarization. Despite its significance and the advancements made by neural coreference models, CR models face a major bottleneck: their limited generalization capability.

Prior work attributes this generalization gap to differences in annotations, such as what constitutes a mention (or entity) and varying preferences to span boundaries. For a model to have *strong referential capabilities*, it must adapt to these annotation-specific nuances. However, achieving this level of adaptability remains a significant challenge, even for state-of-the-art (SOTA) models. This challenge is further amplified when evaluating the referential capabilities of large language models (LLMs) in a few-shot setting, where replicating nuanced annotations with just a few examples is highly unrealistic. We observe that these annotation-specific nuances, can be beneficial but are not essential for downstream tasks or for evaluating the core referential capabilities of an LLM. We describe these nuances as *bells and whistles*.

In this work, we redefine the traditional formulation of coreference resolution by shifting focus away from its *bells and whistles*. Instead, we propose task formulations more aligned with practical applications and demonstrate improved generalizability across domains.

Our first contribution introduces an alternative referential task, Major Entity Identification (MEI). MEI simplifies referential tasks by:(a) assuming that target entities are explicitly provided in the input, and (b) focusing exclusively on frequent entities. Assuming entities to be part of the input shifts the responsibility for domain-specific annotation adaptation—determining which entities are annotated—from the training phase to inference. Through extensive experiments, we show that MEI models generalize effectively across domains using both supervised approaches and LLM-based few-shot prompting across multiple datasets. Importantly, MEI aligns with the classification framework, enabling the use of robust, intuitive, and well-understood classification-based evaluation metrics. Beyond its theoretical appeal, MEI also has practical utility as it allows users to efficiently search for all mentions of a specific entity or a group of entities of interest.

Our second major contribution addresses critical shortcomings identified in recent evaluations of large language models (LLMs) on coreference resolution. These studies revealed that traditional output formats and evaluation metrics fail to capture models' referential understanding fully. Traditional evaluation methods require reproducing the entire document along with annotated cluster information or precisely replicating the antecedent span. This introduces additional *bells and whistles*, such as ensuring the accurate reproduction of spans and documents. To tackle this issue, we introduce `IdentifyMe`, a new benchmark for mention resolution that adopts a multiple-choice question (MCQ) format—a widely used evaluation approach for LLMs. With this simplified task design, any failure can now, be attributed exclusively to issues with *mention resolution*. `IdentifyMe` presents long narratives and applies heuristics to eliminate easily identifiable mentions, resulting in a more challenging and rigorous task. The benchmark incorporates a curated mix of various mention types and their corresponding entities, enabling fine-grained analysis of model performance. Notably, LLM performance remains substantially below human-level performance on IdentifyMe, highlighting considerable room for improvement even for advanced models like GPT-4. The evaluation also reveals key weaknesses in current LLMs, particularly with pronominal mentions, nested mentions, and other nuanced cases.

Overall, this work moves beyond traditional coreference resolution formulations, focusing on tasks with practical applicability and providing fresh insights into the referential strengths and weaknesses of current models. We term this approach *Coreference Without Bells and Whistles* — a streamlined perspective that prioritizes utility and understanding of model capabilities over tailored annotation adaptation.

Contents

Chapter Page

1 Introduction . . . 1
1.1 The Components . . . 1
1.1.1 The Entity . . . 1
1.1.2 The Span . . . 1
1.1.3 The Mention . . . 2
1.1.4 The Antecedent . . . 2
1.2 Coreference Resolution . . . 2
1.3 How is Coreference Resolution Achieved? . . . 2
1.4 Generalizability & *Bells and Whistles* . . . 4
1.4.1 Mention Definition . . . 4
1.4.2 Span Boundaries . . . 5
1.4.3 Reproducibility . . . 5
1.5 Contributions . . . 5
1.5.1 Major Entity Identification: A Generalizable Alternative to Coreference Resolution 5
1.5.2 `IdentifyMe`: A Challenging Mention Resolution Benchmark for LLMs . . . 5
1.6 Thesis Outline . . . 6

2 Related Work . . . 7
2.1 Neural models for CR . . . 7
2.2 Datasets . . . 7
2.3 Metrics . . . 8
2.4 Generalization in CR . . . 8
2.5 LLM Evaluation . . . 9
2.6 Entity-centric tasks . . . 9

3 Major Entity Identification: A Generalizable Alternative to Coreference Resolution . . . 10
3.1 Introduction . . . 10
3.2 Task Formulation . . . 12
3.2.1 Notation . . . 12
3.2.2 Formulation . . . 12
3.3 Supervised MEI models . . . 13
3.3.1 Document encoding . . . 13
3.3.2 Mention detection. . . . 13
3.3.3 Identification module. . . . 13
3.3.4 Updates to the working memory. . . . 14

3.4 Few-shot MEI with LLMs . . . 14
3.4.1 Mention detection challenges. . . . 15
3.4.2 MEI with LLMs. . . . 15
3.4.3 Prompt 1. Word-level MEI. . . . 15
3.4.4 Prompt 2. Head2Span retrieval. . . . 16
3.4.5 Preserving structure. . . . 16
3.5 Experimental Setup . . . 16
3.5.1 Datasets. . . . 16
3.5.2 Metrics. . . . 17
3.5.3 Major entity selection. . . . 17
3.5.4 Implementation details. . . . 17
3.6 Experiments: Supervised Models . . . 18
3.6.1 Baselines. . . . 18
3.6.2 Supervised results. . . . 19
3.6.3 Generalization across datasets. . . . 19
3.6.4 Long documents. . . . 19
3.6.5 Computational performance. . . . 20
3.6.6 Performance across number of entities . . . 20
3.7 Experiments: Few-shot prompting . . . 22
3.7.1 Models. . . . 22
3.7.2 Linking Performance w/ Gold Mentions . . . 22
3.7.3 MEI Task Performance with LLMs . . . 23
3.8 Error Analysis . . . 24
3.9 Conclusion . . . 25
3.10 Prompts . . . 26

4 IdentifyMe: A Challenging Mention Resolution Benchmark for LLMs . . . 31
4.1 Introduction . . . 31
4.2 `IdentifyMe` Benchmark . . . 33
4.3 Selecting Mentions for `IdentifyMe` . . . 33
4.3.1 Step 1: Discard easy mentions. . . . 33
4.3.2 Step 2: Ranking mentions by difficulty. . . . 34
4.4 Dataset Statistics . . . 34
4.5 How difficult is `IdentifyMe`? . . . 34
4.6 Experiments . . . 35
4.7 Results . . . 35
4.8 Error Analysis . . . 36
4.8.1 Sample Outputs . . . 38
4.9 Conclusion . . . 38
4.10 Prompts . . . 39

5 Conclusion and Future Scope . . . 43
5.1 Conclusion . . . 43
5.2 Future Scope . . . 44

List of Figures

Figure		Page
1.1	Text excerpt from *Aladdin* with coreference outputs.	1
1.2	Output format of the seq2seq model proposed by *Bohnet et. al* [1]	3
1.3	Few-Shot Doc-Level Coreference Output	4
3.1	CR *vs*. MEI.	11
3.2	Identification module of MEIRa.	14
3.3	Results on the AFT dataset.	20
3.4	Linking speed comparison between MEIRa-S and `longdoc`.	20
3.5	Performance of MEIRa-H across number of target entities.	21
4.1	Sample instance from the validation set of `IdentifyMe`.	32
4.2	Examples of errors in nested mention resolution and dialogue attribution.	38
4.3	Sample instances from `IdentifyMe` that both `GPT-4o` and `Llama-3.1` correctly identified.	39

List of Tables

Table Page

3.1 Comparing CR and MEI benchmarks statistically. . . . 12
3.2 Results for models trained jointly on FantasyCoref and LitBank. . . . 18
3.3 Results for models trained on OntoNotes. . . . 19
3.4 Few-shot LLM prompting results assuming the availability of ground-truth mentions. . . 22
3.5 Results on LLMs with different mention detection and linking strategies. . . . 23
3.6 Error Analysis of GPT-4 and MEIRa-H 24
3.7 Qualitative Analysis showcasing different errors made by GPT-4 and MEIRa-H. . . . 25
3.8 Prompt for WL Coreference 27
3.9 Prompt for H2S Retrieval 28
3.10 Prompt for evaluating linking performance 29
3.11 Prompt for Direct version of E2E MEI 30

4.1 Performance of LLMs on `IdentifyMe` val split vs Random split (10 runs) 33
4.2 Validation accuracy of LLMs w/ and w/o CoT. . . . 33
4.3 Comparing the performance of LLMs and Humans on a subset of 50 questions 35
4.4 Performance of various models on the `IdentifyMe` test set. . . . 36
4.5 Performance split by mention type and dataset source. . . . 36
4.6 LLM Performance on Ent. vs NoA 37
4.7 LLM performance on nested mentions versus non-nested mentions 37
4.8 The zero-shot prompt passed to `GPT-4o-mini` to generate representative phrases for each cluster/entity. . . . 40
4.9 QA prompt with CoT used in the test set experiments. . . . 41
4.10 QA prompt without CoT. . . . 42

Chapter *1*

Introduction

There lived [a poor tailor named Mustapha]$_1$, who had [a son called Aladdin]$_2$. [Aladdin]$_2$ was disobedient to [[his]$_2$ father]$_1$ and [mother]$_3$ and spent all [his]$_2$ time idling with [[**his**]$_2$ **friends**]$_4$.When [Aladdin]$_2$ was old enough to learn, [[his]$_2$ father]$_1$ took [him]$_2$ into [[his]$_1$ own shop]$_5$, and taught [him]$_2$ how to use [**the needle**]$_6$ [there]$_5$.

Figure 1.1: Text excerpt from *Aladdin* with coreference outputs.

1.1 The Components

Let us begin the discussion of coreference resolution by discussing the essential components of coreference resolution - *the entity*, *the span*, and *the mention*.

1.1.1 The Entity

An Entity is any individual, location, organization, object or concept mentioned in the text and can be distinctively identified. For the example text, the entity set is { *Mustapha* (people), *Aladdin* (people), *mother* (people), *Aladdin's friends* (people), *Mustapha's shop* (location/object), *the needle* (object/concept) }.

1.1.2 The Span

A span of text refers to a contiguous sequence of words within a larger body of text. In the text, it can be defined as a starting offset and an ending offset. For example, the span [2, 6] indicates the string *a poor tailor named Mustapha* while [13, 13] and [15, 15] represent two separate spans with the same string representation *Aladdin*.

1.1.3 The Mention

A mention is a specific span of text that refers to an entity. For example, the spans {*a poor tailor named Mustapha* ([2,6]), *his father* ([19,20]), *his father* ([30,31]), *his* ([35,35]) } refer to the entity *Mustapha*.

1.1.4 The Antecedent

The antecedent of a mention is any mention that refers to the same entity, and occurs earlier in the document. A mention can have multiple antecedents; for instance, in the case of *his father* in the last line of the excerpt, its antecedents are *his father* and *a poor tailor named Mustapha*, with the former being the nearest antecedent.

1.2 Coreference Resolution

Coreference Resolution is the task of identifying all mentions that correspond to an entity mentioned in the text. The coreference resolution task involves taking a document as input and producing mention clusters, ensuring that only mentions referring to the same entity are grouped together. For example, for the excerpt in Fig. 1.1, the coreference outputs could be represented as: {{*a poor tailor named Mustapha*, *his father*, *his father*, *his*}, {*a son called Aladdin*, *Aladdin*, *his*, *his*, *his*, *Aladdin*, *his*, *him*}, {*mother*}, {*his friends*}, {*his own shop*, *there*}, {*the needle*}}. Here, we have identified all mentions of every entity within the document, listed in Subsection 1.1.1.

Why is coreference important?
Coreference resolution is a critical aspect of natural language understanding, playing a key role in various downstream NLP tasks, including question-answering, summarization, translation, and domain-specific applications such as legal and medical text analysis. Additionally in the age of Large Language Models (LLMs), which focus on natural language generation to tackle a wide range of traditional challenges, it is crucial for these models to have a strong internal capability for referential understanding. For instance, an LLM tasked with question answering must inherently resolve references in the text, providing entity names instead of ambiguous pronouns.

1.3 How is Coreference Resolution Achieved?

Since the seminal work of *Lee et. al* [2], coreference resolution has largely been approached through task-specific models designed for a two-stage process consisting of *mention detection* and *mention resolution*. ***Mention detection*** entails scoring spans (typically up to a maximum length l) to determine whether they represent a mention or a generic span. ***Mention Resolution*** assigns the identified mention

Input: There lived a poor tailor named Mustapha, who had a son called Aladdin.

Prediction: SHIFT: next sentence

Input: There lived a poor tailor named Mustapha, who had a son called Aladdin. Aladdin was disobedient to his father and mother and spent all his time idling with his friends.

Prediction:

A Aladdin → a son called Aladdin

B his → Aladdin

C his father → a poor tailor named Mustapha

D his → his

E his → his

F SHIFT: next sentence

Figure 1.2: Output format of the seq2seq model proposed by *Bohnet et. al* [1]

to an existing entity or creates a new one. This resolution typically involves mapping the current mention to an *Entity Id* or identifying a previous mention as the antecedent, thereby establishing a connection between the mentions resulting in clusters. These models often yield results that, after post-processing of decision labels (*Entity/Antecedent ID* or *None*), directly produce clusters in the form described in Section 1.2.

Recent work [1, 3] have explored leveraging advances in seq2seq models by framing coreference resolution as a seq2seq task. These methodologies typically involve reproducing mentions identified in the text (*Mention Detection*), while *mention resolution* is accomplished by generating an action structure that connects the identified mention to an antecedent phrase (*e.g.* refer to Fig. 1.2) These approaches utilize large pre-trained seq2seq models (e.g., T0 or mT5), which are subsequently fine-tuned on coreference datasets to generate outputs with the desired structure.

In recent times, with the advent of LLMs, efforts have been made to evaluate the zero-shot/few-shot referential capabilities of LLMs. Such evaluations require the generation output to be close to natural language for better performance. Hence, previous efforts towards evaluating coreference resolution [4, 5], have opted to reproduce the entire document with annotated cluster information (refer Fig. 1.3. For evaluating *mention resolution*, works[6, 7] have focused on a QA-style prompting strategy where the model reproduces an antecedent span to form coreference clusters.

In this work, we focus on task-specific models and the few-shot evaluation of LLMs, as seq2seq fine-tuning is computationally intensive and falls outside the scope of this thesis.

Input: There lived a poor tailor named Mustapha, who had a son called Aladdin. Aladdin was disobedient to his father and mother and spent all his time idling with his friends. When Aladdin was old enough to learn, his father took him into his own shop, and taught him how to use the needle there.

Prediction: There lived [a poor tailor named Mustapha] **(#cluster_1)**, who had [a son called Aladdin] **(#cluster_2)**. [Aladdin] **(#cluster_2)** was disobedient to [[his] **(#cluster_2)** father] **(#cluster_1)** and [mother] **(#cluster_3)** and spent all [his] **(#cluster_2)** time idling with [[his] **(#cluster_2)** friends] **(#cluster_4)**. When [Aladdin] **(#cluster_2)** was old enough to learn, [[his] **(#cluster_2)** father] **(#cluster_1)** took [him] **(#cluster_2)** into [[his] **(#cluster_1)** own shop] **(#cluster_5)**, and taught [him] **(#cluster_2)** how to use [the needle] **(#cluster_6)** [there] **(#cluster_5)**.

Figure 1.3: Few-Shot Doc-Level Coreference Output

1.4 Generalizability & *Bells and Whistles*

Over the years, the performance of neural models on CR has improved due to advancements in neural architectures, the increase in model parameters, and the emergence of pretrained encoders. However, performance of neural models trained on one of the datasets significantly decreases when tested across unseen benchmarks across domains raising concerns of poor generalization However, recent research[8] suggests that his generalization gap might be due to annotation differences, with models learning idiosyncracies corresponding to the trained dataset. Since the inability to adapt to different annotations conflate with poor generalizable coreference resolution, it is difficult to measure the generalizability of models currently through multi-dataset evaluation. These annotation differences are termed by us as ***bells and whistles***, features that are helpful to adapt to specifically, but are not essential for downstream tasks or for evaluating the core referential capabilities of an LLM. We highlight some of these differences that we deal with in our work.

1.4.1 Mention Definition

One of the most prevalent differences in annotation is the definition of what constitutes a mention. For example, OntoNotes [9] does not annotate singletons, confounding mention identity with being referential. While OntoNotes covers unrestricted coreference, LitBank limits the markable entities to only specific categories: people (PER), facilities (FAC), locations (LOC), geo-political entities (GPE), organizations (ORG), and vehicles (VEH). Thus models, trained on OntoNotes or Litbank generalize poorly when tested across each other.

1.4.2 Span Boundaries

Span boundaries are crucial in defining a mention (ref Subsection 1.4.1). Prior research [10] has demonstrated that coreference annotations often vary in span boundaries. For instance, golden and system mentions may differ due to the inclusion or exclusion of surrounding punctuation, such as commas or quotation marks. These inconsistencies in boundaries significantly impact coreference evaluation scores.

1.4.3 Reproducibility

Ensuring the reproducibility of exact spans (without minor variations) and the document in fine-tuned seq2seq models and few-shot evaluations of LLMs is crucial, as it can significantly impact performance on referential tasks like coreference and mention resolution. However, poor performance stemming from reproducibility issues does not truly represent the model's referential capabilities.

1.5 Contributions

In our work, we aim to get around these *bells and whistles* i.e. annotation differences through new task formulations more aligned with practical applications and demonstrate improved generalizability across domains. In this section, we describe our two-fold contributions.

1.5.1 Major Entity Identification: A Generalizable Alternative to Coreference Resolution

This study, presented at *EMNLP 2024*, We propose an alternative referential task of Major Entity Identification (MEI), which helps deal with mention annotation differences discussed in Subsection 1.4.1. MEI simplifies referential tasks by assuming that the frequent target entities are explicitly provided in the input. Assuming entities to be explicitly part of the input improves the generalizability of the task by shifting the domain-specific annotation adaptation from the training phase to inference, guided by explicit cues. Through comprehensive experiments, we demonstrate that MEI models generalize effectively across domains, utilizing both supervised methods and LLM-based few-shot prompting on various datasets.

1.5.2 `IdentifyMe`: A Challenging Mention Resolution Benchmark for LLMs

As noted in Subsection 1.4.2 and Subsection 1.4.3, traditional LLM evaluation relies on reproducing entire documents or specific spans like mention or antecedent spans. Any failure to accurately replicate the test reflects negatively on the model's performance in referential tasks. However, such poor results blur the line between the model's inability to reproduce spans and its true referential capabilities. To tackle this issue, we introduce `IdentifyMe`, a new benchmark for *mention resolution* that adopts a multiple-choice question (MCQ) format—a widely used evaluation approach for LLMs. By simplifying

the task design, failures can be solely attributed to *mention resolution* issues. `IdentifyMe` features long narratives and uses heuristics to remove easily identifiable mentions, creating a more challenging task. The benchmark combines diverse mention types and entities, enabling fine-grained performance analysis. This work was accepted as a short paper (oral presentation) at *NAACL 2025*.

1.6 Thesis Outline

Overall, this thesis is divided into five chapters, for each of which a quick description is given below:

- **Chapter 2** provides a concise overview of prior research on neural coreference resolution models and their generalizability. Additionally, it explores the evaluation of large language models (LLMs) for their referential capabilities and examines other task formulations that share similarities with our proposed new tasks.
- **Chapter 3** provides an in-depth exploration of Major Entity Identification, including the task's formulation, an analysis of model architectures, and their generalizability. Additionally, it assesses the end-to-end referential capabilities of large language models (LLMs).
- **Chapter 4** presents further details on `IdentifyMe`, including its statistics and the insights gained into the mention resolution capabilities of LLMs.
- **Chapter 5** presents a brief overview of our contributions and outlines the potential directions for future work.

Chapter *2*

Related Work

2.1 Neural models for CR

Neural models for coreference resolution (CR) have become the de facto standard in supervised settings since the seminal works of [2, 11]. This marked a significant shift from traditional approaches that relied heavily on manually engineered features and rule-based systems [12, 13, 14]. These neural models treat all possible text spans as potential mentions and learn the probability of each span being an antecedent of another. By leveraging deep contextual embeddings to derive span representations, they significantly outperformed traditional models. Major advances came with the use of pre-trained encoders further fine-tuned for the coreference resolution task [15, 16]. Notably,[16] introduced a pre-training objective specifically designed to better represent and predict text spans, leading to improved performance on span-based tasks such as coreference resolution. Recent works have explored tailored architectures for handling diverse linguistic relations in coreference[17], while others have leveraged advances in pre-trained sequence-to-sequence models [1, 3]. As neural models continue to become more sophisticated, attention has turned toward improving their efficiency. For example,[18] proposed using word-level spans as candidate mentions, significantly reducing computational complexity. Another line of research focuses on using single dense representations for entity clusters. Incremental approaches to coreference resolution using such representations have shown to improve efficiency with minimal accuracy loss[19, 20] have been introduced.

2.2 Datasets

Several influential datasets have been pivotal in the advancement and evaluation of coreference resolution models. OntoNotes[21] is a large-scale, multi-genre corpus containing approximately 2.9 million words drawn from news, broadcast conversations, web data, and more. It offers rich linguistic annotations—including part-of-speech tags, named entities, and coreference relations—and has become a widely used benchmark for coreference tasks. LitBank[22], on the other hand, centers on literary texts, providing a challenging dataset with longer documents and more complex coreference patterns that

often involve characters and abstract entities. In contrast, FantasyCoref [23] targets the fantasy genre, introducing unique challenges due to fictional worlds and entities.

A comparison of these datasets highlights key differences in size, genre, and annotation practices. For instance, OntoNotes features shorter documents on average (467 words) compared to LitBank (2105 words) and FantasyCoref (1500 words). They also differ in their annotation guidelines: OntoNotes and FantasyCoref exclude singleton mentions, whereas LitBank includes them. Another notable distinction lies in what constitutes a mention. While OntoNotes and FantasyCoref annotate all noun phrases, LitBank restricts annotation to specific entity types such as people, locations, organizations, facilities, geopolitical entities, and vehicles.

Prior studies [24, 25, 8] have demonstrated that lexical and syntactic variation across datasets can significantly reduce model performance when transferring from one dataset to another. This suggests that models often learn dataset-specific biases instead of robust, generalizable coreference representations. To tackle these generalization challenges, researchers have explored domain adaptation techniques and harmonized annotation frameworks to enable joint training and evaluation across datasets. Our work contributes to this ongoing line of research.

2.3 Metrics

Coreference resolution systems are typically evaluated using a set of established metrics. The MUC score [26], one of the earliest, measures performance based on entity links. The B3 score [27] evaluates precision and recall by examining the overlap of coreference chains at the mention level. CEAFe [28], in contrast, emphasizes entity-level alignment. To provide a more holistic assessment, the CoNLL score averages the F1 scores of MUC, B3, and CEAFe. Alternative metrics such as BLANC and LEA, introduced in [29] and [30] respectively, were designed to address limitations of traditional metrics. BLANC balances coreference and non-coreference (singleton) decisions, while LEA incorporates entity-aware scoring to fairly evaluate clusters of varying sizes.

Ongoing research continues to explore and refine evaluation metrics to better capture the complexities of coreference resolution in different settings. In our approach, Major Entity Identification reframes coreference resolution as a classification problem, allowing the use of widely-adopted and intuitive classification metrics such as Macro-F1 and Micro-F1.

2.4 Generalization in CR

Generalization in coreference resolution (CR) continues to pose a significant challenge, as models often struggle to perform well across diverse domains and datasets [24, 25, 8]. To address the limitations in generalization, researchers have explored several strategies. Feature addition involves enriching the input representations with linguistic cues or external knowledge. For example,[31] demonstrated the benefits of incorporating semantic information from knowledge graphs to improve coreference

resolution. [17] utilized tailored architectures for specific linguistic relations and coreference types. Joint training methods aim to improve generalization by training CR models on multiple datasets, promoting domain-invariant learning [32, 33]. Additionally, active learning frameworks have also been explored in [34, 35, 36]. Rather than relying on additional training data, we argue for an alternative formulation where the burden of domain adaptation is offloaded from training to inference.

2.5 LLM Evaluation

LLM Evaluation on referential tasks has largely been conducted in limited settings, such as the sentence-level Winograd Schema Challenges (WSC) [37], clinical pronoun resolution [7] and instance-level Q&A [38]. *Le and Ritter* [4] conducted the first document-level evaluation of LLMs for CR but assumed an oracle-mention detector. In contrast, in MEI we conduct end-to-end evaluations.

In our work on `IdentifyMe`, we observe from prior studies that some, such as [4, 5], report that the performance of LLMs on CR has been relatively underwhelming, with models facing challenges even in mention detection. Recent work by *Gan et al.* [6] has conducted an in-depth analysis, revealing that the span-based output format commonly used in typical CR setups underestimates LLMs' remarkable referential understanding. They suggest adapting CR datasets and task metrics to support LLM evaluations, and our work takes a step in this direction.

2.6 Entity-centric tasks

Entity-centric tasks similar to MEI include character identification, where either annotations are restricted to a subset of entities [39] or custom models are developed to extract mentions of specific characters from TV show transcripts [40, 41]. MEI differs from these works by adopting a generalized task formulation independent of annotation strategies and entity selection. Another task, Entity Linking [42] extracts distinct entities from a document and links them to external Knowledge Bases. In contrast, MEI focuses on retrieving all mentions (including nominals and pronominals) of a specific set of key entities, extracted solely from the context of the document.

Chapter *3*

Major Entity Identification: A Generalizable Alternative to Coreference Resolution

3.1 Introduction

Coreference resolution (CR) is the task of finding text spans that refer to the same entity. CR is a fundamental language understanding task relevant to various downstream NLP applications, such as question-answering [43], building knowledge graphs [44], and summarization [45]. Despite the importance of CR and the progress made by neural coreference models [18, 1, 3], domain generalization remains an issue even with the best-performing supervised models [32, 33].

The lack of domain generalization in CR models can largely be attributed to differences in annotation guidelines of popular CR benchmarks, specifically annotation guidelines about what constitutes a mention [8]. For example, OntoNotes [9] does not annotate singletons, confounding mention identity with being referential. Thus, models trained on OntoNotes generalize poorly [33]. The importance of mention detection for CR generalization is further highlighted by [36], showing that solely annotating mentions is sufficient and more efficient for adapting pre-trained CR models to new domains (in comparison to annotating coreference chains). Similarly, GPT-4 struggles with zero-/few-shot mention prediction, but with ground-truth mentions, its CR performance [4] is competitive with that of supervised models [33].

Given these observations, we hypothesize that current CR models, including large language models, generalize well at *mention clustering* but struggle to generalize on *mention detection* due to idiosyncrasies of different domains/benchmarks. We put forth an alternative referential task where the entities of interest are known and provided as additional input. Assuming entities to be part of the input offloads the required domain adaptation from training to inference. Specifically, we propose the task of Major Entity Identification (MEI), where we assume the major entities of the narrative, to be provided as input along with the text (see Fig. 3.1a). We focus on major entities for the following reasons: (a) Specifying major entities of a narrative is intuitively easier. (b) A handful of major entities often dominate any discourse. Table 3.1a shows that in FantasyCoref roughly 16% of entities (942 of 5829) contribute to 63% of the mentions (35938 of 56968).

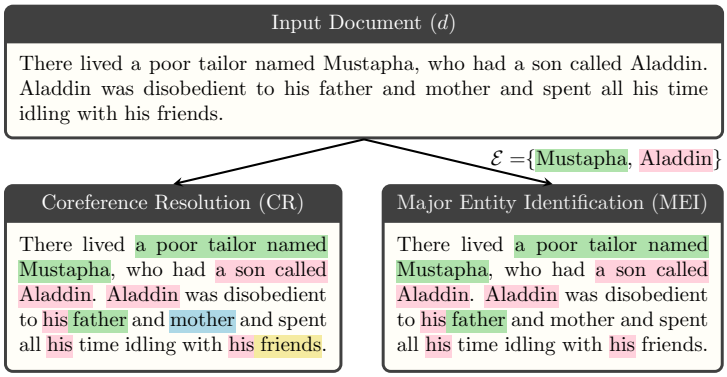


Figure 3.1: CR *vs*. MEI.

(a) The CR task aims to detect and cluster all mentions into different entities, shown in various colors. MEI takes major entities as additional input and aims to detect and classify the mentions that refer only to these entities.

In this work, we adapt two literary CR benchmarks, namely LitBank [46] and FantasyCoref [23] by identifying frequently occurring entities as major entities and customizing a state-of-the-art coreference model [33] to MEI. Our tests for generalizability reveal that while there is a big gap in CR performance between in- and out-of-domain models [33], this performance gap is much smaller for MEI (Section 3.6).

To test this hypothesis further, we evaluate large language models (LLMs) for MEI in a few-shot learning setup. On CR, LLMs are shown to struggle with mention detection and perform worse than supervised models [4]. Contrary to this, on MEI, top LLMs (*e.g*. GPT-4) are only slightly behind supervised models (Section 3.7). These experiments in the supervised setting and the few-shot setting demonstrate that the MEI task is more generalizable than CR.

Additionally, we argue that MEI is easier to evaluate than CR. The MEI task can be viewed as a classification task in which any text span either refers to one of the input entities or the null class (*minor* entities and other non-mention spans). The classification metrics maintain consistent granularity, proportionally penalize perturbations, and exhibit high discriminatory power while intuitively meeting multiple desired specifications [47, 48].

Furthermore, MEI, by its definition, disregards insignificant and smaller clusters known to inflate the CR metrics [47, 49, 50]. As an aside, formulating MEI as a classification task allows for a trivial parallelization across candidate spans (See Subsection 3.6.5).

Finally, MEI's explicit mapping of mentions to predefined entities improves its usability over CR in downstream applications that focus on mentions of specific entities. MEI effectively replaces tailored heuristics employed to extract CR cluster(s) referring to entities of choice in such applications (entity understanding [51], sentiment and social dynamics analysis [41, 52]).

	LitBank		FantasyCoref	
Statistics	CR	MEI	CR	MEI
# of Mentions	29103	16985	56968	35938
# of Non singletons	23340	16985	56968	35938
Mean ant. dist.	55.31	36.95	57.58	30.24
# of Clusters	7927	490	5829	942
Avg. cluster size	3.67	34.66	9.77	38.15

Table 3.1: Comparing CR and MEI benchmarks statistically.

(a) MEI has fewer but larger clusters, and a smaller mean antecedent distance (Mean ant. dist.). Our formulation's frequency-based criterion for deciding major entities means that singleton mentions are typically not a part of MEI.

3.2 Task Formulation

3.2.1 Notation

For a document d, let $\mathcal{E} = \{e_j\}_{j=1}^{L}$ be the set of L major entities that we wish to identify. We define $\mathcal{M}_{\text{all}}$ as the set of all mentions that could refer to any entity and subsequently $\mathcal{M}_j \subseteq \mathcal{M}_{\text{all}}$ as the set of mentions that refer to a major entity e_j. Furthermore, we denote $\mathcal{M} = \bigcup_j \mathcal{M}_j$ as the set of mentions that refer to one of the major entities while mentions that do not correspond to any major entity are designated as $\mathcal{M}_{\text{other}} = \mathcal{M}_{\text{all}} \setminus \mathcal{M}$.

3.2.2 Formulation

In MEI, the input consists of the document d and designative phrases $\mathcal{P} = \{p(e_j)\}_{j=1}^{L}$ where $p(e_j)$ succinctly represents the entity e_j. For example, in Fig. 3.1a, the phrases *"Aladdin"* and *"Mustapha"* uniquely represent Aladdin and his father who appear in *"Aladdin And The Wonderful Lamp"*. Note that in CR, the designative phrases $\mathcal{P}$ are not part of the input.

In contrast to CR's clustering foundations, MEI starts with a prior for each entity (the designative phrase) and can be formulated as an open set classification, where every mention is either classified as one of the major entities or ignored. Formally, MEI aims to assign each mention $m \in \mathcal{M}_j$ to e_j and mentions $m \in \mathcal{M}_{\text{other}}$ to $\varnothing$, a null entity.

3.3 Supervised MEI models

We propose MEIRa, **M**ajor **E**ntity **I**dentification via **Ra**nking, which draws inspiration from the entity ranking formulation [19, 20] and maintains an explicit representation for entities. The MEIRa models consist of 3 steps: encoding the document, proposing candidate mentions, and an identification (id) module that tags mentions with major entities or the null entity.

3.3.1 Document encoding

is performed using a Longformer-Large [53], ϕ, that we finetune for the task. Mentions (or spans) are encoded as $\mathbf{m}_i = \phi(m_i, d)$ by concatenating the first, last, and an attention-weighted average of the token representations within the mention span. In MEI, an additional input is the set of designative phrases $\mathcal{P}$ for the major entities. Since each phrase is derived from the document itself, we also obtain its encoding using the backbone: $\mathbf{e}_j = \phi(p(e_j), d)$.

3.3.2 Mention detection.

Similar to prior efforts [33], we use a mention proposal network that predicts high-scoring candidate mentions. This step finds all mentions $\mathcal{M}_{\text{all}}$ and not just the ones corresponding to the major entities $\mathcal{M}$.Training a model to only detect mentions of major entities would confuse it leading to poor performance.

3.3.3 Identification module.

As illustrated in Fig. 3.2a, we initialize a working memory $\mathcal{E}^W = [\mathbf{e}_j]_{j=1}^L$ as a list of L major entities based on their designative phrase representations. Given a mention m_i, the id module computes the most likely entity as:

$$[s_i^*, e_i^*] = \max_{j=1\ldots L} f([\mathbf{m}_i, \mathbf{e}_j, \chi(m_i, e_j)]) \,, \tag{3.1}$$

where $f()$ is an MLP that predicts the score of tagging mention m_i with the entity e_j, and $\chi(m_i, e_j)$ encodes metadata. The output s_i^* corresponds to the highest score and e_i^* is the top-scoring entity. Based on the score, m_i is assigned to:

$$y(m_i) = \begin{cases} e_i^* & \text{if } s_i^* > \tau \,, \\ \varnothing & \text{otherwise} \ , \end{cases} \tag{3.2}$$

where τ is a threshold (set to 0 in practice).

The metadata $\chi(m_i, e_j)$ contains a distance (position) embedding representing the log distance between the mention m_i and the last tagged instance of the entity e_j. If no mention is yet associated with the entity, we use a special learnable embedding.

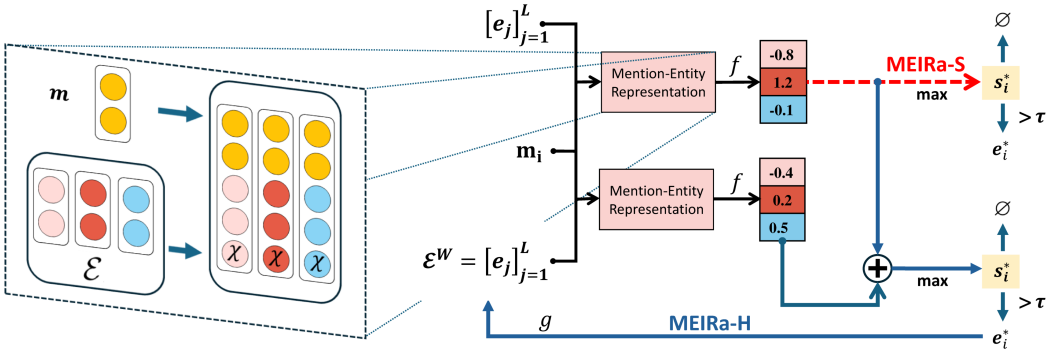


Figure 3.2: Identification module of MEIRa.

(a) A mention encoding $\mathbf{m}_i$ is concatenated with each entity's embedding in $\mathcal{E}^W$ and the metadata $\chi(m_i, e_j)$. Network f scores the likelihood of assigning m_i to each major entity. If the highest score s_i^* is above the threshold τ, m_i is associated with the highest scoring major entity e_i^* or discarded. In MEIRa-S, the entity memory $\mathcal{E}^W$ remains static. For MEIRa-H (blue path), the assigned entity's working memory is updated, and both the static (top half) and updated working memory (bottom half) are utilized to compute a final score.

3.3.4 Updates to the working memory.

We investigate two approaches:

1. **MEIRa-S**tatic:As the name suggests, the working memory $\mathcal{E}^W$ of the entity representations remains constant ($\mathcal{E}^{W(0)}$) and is not updated with new mention associations. This makes the approach highly parallelizable.

2. **MEIRa-H**ybrid: Similar to traditional CR, this variation maintains a dynamic working memory $\mathcal{E}^W$, which is updated with every new mention-id association. Specifically, assuming m_i is assigned to e_j^*, the working memory would be updated using a weighted mean operator g as $\mathbf{e}_j \leftarrow g(\mathbf{e}_j, \mathbf{m}_i)$, similar to [20]. To prevent error accumulation, we evaluate the mentions against $\mathcal{E}^W$ and the initial entity representations ($\mathcal{E}^{W(0)}$), then compute the average score. This hybrid approach reaps benefits from both, the initial clean designative phrases and the dynamic updates.

Following [20], the mention detection and identification modules are trained end-to-end using separate cross-entropy loss functions.

3.4 Few-shot MEI with LLMs

We propose a prompting strategy to leverage LLMs for MEI, addressing their challenges in CR.

3.4.1 Mention detection challenges.

CR or MEI can be addressed using separate few-shot prompting strategies for mention detection and mention clustering/identification. However, [4] found that this strategy faced significant challenges with mention detection, performing worse than a deterministic mention detector. Thus, they assume access to an oracle mention detector and focus on evaluating LLMs' linking capabilities.

An alternative is to use an external supervised mention detector instead of the oracle. However, this requires annotated training data and may not align with a true few-shot LLM prompt paradigm. Additionally, supervised mention detectors often fail to generalize across CR datasets due to annotation variability [49].

3.4.2 MEI with LLMs.

We demonstrate that transitioning from CR to MEI addresses this gap in mention detection and proposes an end-to-end, few-shot prompting approach for MEI. Inspired by [18], we develop a prompting strategy that first performs MEI at word-level (rather than span), followed by a prompt to retrieve the span corresponding to the word. In addition to the document d and the set of phrases $\mathcal{P}$, we also provide entity identifiers (*e.g.* #1, #2) to the LLM.
We will use the following example:

Document:
That lady in the BMW is Alice's mom.

Major Entities:
1. *Alice*;
2. *Alice's mother*.

3.4.3 Prompt 1. Word-level MEI.

Mention detection with LLMs is challenging due to the frequent occurrence of nested mentions. We overcome this by prompting the LLM to tag each word. Specifically, through few-shot examples, we ask the LLM to detect and tag the **syntactic heads**[1] (e.g., *lady*, *Alice*, *mom*) of mentions that refer to the major entities. Other words are left untagged (implicitly assigned to $\varnothing$, the null entity). To create the few-shot examples, a contiguous set of words annotated with the same entity is considered as a span and its syntactic head is extracted using `spaCy` [54].

The ideal output for the example above is:

"That lady#2 in the BMW is Alice#1's mom#2..".

[1] A syntactic head of a phrase is a word (*lady*) that is central to the characteristics of the phrase (*The lady in the BMW*).

Note that, even though the span *"BMW"* might be a valid mention, it is not annotated as it does not refer to one of the major entities. The exact prompt used for this is provided in Table 3.8.

3.4.4 Prompt 2. Head2Span retrieval.

The entity tagged heads are passed to the Head2Span (H2S) module, along with the document to retrieve the span. The prompt consists of the document pre-annotated with the positions of the head, where each candidate head-word is followed by a "#" and is instructed to be replaced by the complete span (including any existent determiners and adjectives).
For the input:

> *That lady# in the BMW is Alice#'s mom#.*

The expected ideal output is

> *That lady (That lady in the BMW) in the BMW is Alice(Alice's)'s mom (Alice's mom).*

Table 3.9 shows the H2S prompt.

3.4.5 Preserving structure.

We pose MEI as a structured generation task, prompting LLMs to reproduce documents and generate MEI tags at specific locations. Proprietary models like GPT-4 generally reproduce documents faithfully but for rare failures, we use the Needleman-Wunsch algorithm [55] to align documents and extract tags In the case of open-source models, we employ regular expression-based constrained decoding with the `outlines` library [56]

3.5 Experimental Setup

3.5.1 Datasets.

We evaluate three literary datasets chosen for their longer length and identifiable major entities, particularly the key narrative elements such as characters or plot devices. Table 3.1a compares statistical aspects of MEI and CR, revealing that MEI features fewer clusters (entities) but larger cluster sizes (more mentions per cluster).

1. *LitBank* [46] annotates coreference in 100 literary texts, each averaging around 2000 words. Following prior work [33], we utilize the initial cross-validation split, dividing the documents into training, validation, and test sets with an 80:10:10 ratio

2. *FantasyCoref* [23] provides OntoNotes [9]-style[2] coreference annotations for 211 documents from Grimm's Fairy Tales, with an average length of approximately 1700 words. The dataset includes 171 training, 20 validation, and 20 test documents.

3. *Additional Fantasy Text (AFT)* [23] provides annotations for long narratives: (a) Aladdin (6976 words), (b) Ali Baba and the Forty Thieves (6911 words), and (c) Alice in Wonderland (13471 words).

3.5.2 Metrics.

In contrast to CR, MEI facilitates the use of simple classification metrics. We define standard precision and recall for each major entity considered as an individual class of its own.

For a dataset $\mathcal{D} = \{d_1, \ldots, d_{|\mathcal{D}|}\}$, the evaluation metrics are defined as follows:

$$\text{Macro-F1} = \frac{\sum_{d \in \mathcal{D}} \sum_{e_j \in \mathcal{E}_d} F1(e_j)}{\sum_{d \in \mathcal{D}} |\mathcal{E}_d|}, \text{ and} \tag{3.3}$$

$$\text{Micro-F1} = \frac{1}{|\mathcal{D}|} \sum_{d \in \mathcal{D}} \frac{\sum_{e_j \in \mathcal{E}_d} F1(e_j) \cdot |\mathcal{M}_j|}{\sum_{e_j \in \mathcal{E}_d} |\mathcal{M}_j|}. \tag{3.4}$$

Macro-F1 is the average F1-score of entities across the dataset, while Micro-F1 is the frequency-weighted F1-score of entities within a document, averaged across the dataset.

3.5.3 Major entity selection.

We select as major entities, the top-k entities ranked as per the frequency of occurrences. We use k=5 for LitBank and FantasyCoref after visualizing the frequency plots of their training sets. For longer documents in AFT, we select up to 9 entities to ensure coverage of all key entities from the story. We also enforce that every entity $e_j \in \mathcal{E}$ has a mention count $|\mathcal{M}_j| \geq 5$. We derive the representative span for each selected e_j from the set of mentions $\mathcal{M}_j$ by selecting the most commonly occurring name or nominal mention.

3.5.4 Implementation details.

Supervised models: Model hyperparameters are derived from [33]. To ensure consistent performance across different numbers of target entities, we randomly select a subset of major entities at each training iteration. Supervised models were trained five times with random seeds, and we present aggregated results as the mean and standard deviation.

LLMs: We follow a few-shot prompting mechanism across the setups and experiments. Prompts that perform referential tasks consist of 3 examples of 6 sentences each. These 3 examples contain a mixture

[2]The exact guidelines are documented here

	FantasyCoref		LitBank	
Model	Macro-F1	Micro-F1	Macro-F1	Micro-F1
Coref-ID	72.5±2.2	78.8±2.7	79.7±2.7	80.6±3.7
Coref-CM	77.7±1.8	82.4±2.2	74.1±2.5	76.0±3.0
Coref-FM	77.9±1.7	83.2±2.2	77.4±2.3	80.6±4.7
MEIRa-S	**80.7±0.6**	**84.9±0.5**	80.8±0.8	81.8±1.0
MEIRa-H	80.3±1.4	84.3±2.0	**82.3±1.2**	**83.2±2.5**

Table 3.2: Results for models trained jointly on FantasyCoref and LitBank.

of narrative styles (narratives, dialogues), types of entities (major, non-major entities), categories of mentions (names, nominals, pronouns), and plurality. Additionally, before producing the MEI output, we ask the LLM to describe each major entity briefly. We find that this additional step improves performance. For the H2S prompt, we provide 9 sentences as examples, balancing the number of pre- and post-modifiers to the head. All examples were selected from LitBank's train set and kept constant throughout the experiments. We set the temperature to 0 for all the models to ensure consistent and reproducible outputs.

3.6 Experiments: Supervised Models

3.6.1 Baselines.

We train the `longdoc` model [33] for CR and perform the following three inference-time adaptations for MEI:

Coref-ID: `longdoc` uses active lists of entity representations, resolving coreference by associating mentions with existing clusters or generating new ones. During inference, we disable the cluster creation step and pre-fill the entity list with the encoded vector representations of the major entities. Hence, all the detected mentions either get mapped to one of the major entities or are discarded.

Coref-Cosine Map (Coref-CM): Since CR clusters obtained from `longdoc` lack explicit entity association, we employ the Kuhn-Munkres (KM) algorithm [57] to find the optimal matching cluster for each major entity. The cost matrix uses the cosine similarity between the encoded representation of the major entities and the predicted cluster embeddings, both derived from `longdoc`.

Coref-Fuzzy Map (Coref-FM): This method uses the KM algorithm to derive optimal mappings by constructing a cost matrix from accumulated fuzzy-string matching scores between designative phrases and the predicted cluster's mention strings.

	FantasyCoref		LitBank	
Model	Macro-F1	Micro-F1	Macro-F1	Micro-F1
Coref-ID	63.4±1.8	69.5±3.6	58.0±2.4	57.7±1.0
Coref-CM	72.8±0.3	76.5±0.5	61.0±5.9	61.2±5.2
Coref-FM	71.2±1.5	75.2±1.3	66.1±2.1	67.1±3.9
MEIRa-S	**75.7±1.5**	78.5±1.2	74.6±1.1	74.7±1.6
MEIRa-H	74.7±1.0	**78.5±0.8**	**77.2±1.9**	**78.6±2.7**

Table 3.3: Results for models trained on OntoNotes.

3.6.2 Supervised results.

In this experiment, we train MEIRa and the baseline models on the joint training set of LitBank and FantasyCoref. Subsequently, we assess their performance on the individual test sets, with results summarized in Table 3.2. Overall, MEIRa models consistently outperform the baselines on both metrics while also exhibiting better stability with a lower variance. The considerable variance observed in the performance of baseline methods across all experiments underscores the non-trivial nature of identifying clusters corresponding to major entities within the output clusters provided by the CR algorithms. MEIRa-H and MEIRa-S exhibit competitive parity on FantasyCoref (children stories), while MEIRa-H edges out on LitBank dataset, showcasing its adaptability in elaborate sentence constructions.

3.6.3 Generalization across datasets.

To evaluate the generalization capabilities of MEIRa and baseline models, we train them on the OntoNotes dataset and then test their performance on LitBank and FantasyCoref. The results are presented in Table 3.3. When compared with Table 3.2, we observe a significant performance drop across the baseline models (*e.g.* for Coref-ID, the average Micro-F1 scores drop from 80.6 to 57.7 on LitBank). The performance gap for the baseline models is more pronounced on LitBank than on FantasyCoref because LitBank's annotation strategies differ more significantly from those of OntoNotes. The observations aligns with previous work [33], that showcase poor generalization of models trained for CR. In contrast, MEIRa models recover most of the underlying performance on both the datasets (MEIRa-H drops a little from 83.2 to 78.6 on LitBank Micro-F1), demonstrating MEI as a more adaptable task, bringing robustness over varying annotation strategies.

3.6.4 Long documents.

Fig. 3.3 presents results on the AFT dataset of the models trained using a combined training set of LitBank and FantasyCoref. MEIRa models significantly outperform the baseline models, with

	AFT	
Model	Macro-F1	Micro-F1
Coref-ID	68.1±5.9	78.7±6.1
Coref-CM	71.1±2.8	82.4±4.2
Coref-FM	71.1±4.7	83.2±4.7
MEIRa-S	81.6±1.4	88.8±1.3
MEIRa-H	**82.8±1.1**	**89.5±1.0**

Figure 3.3: Results on the AFT dataset.

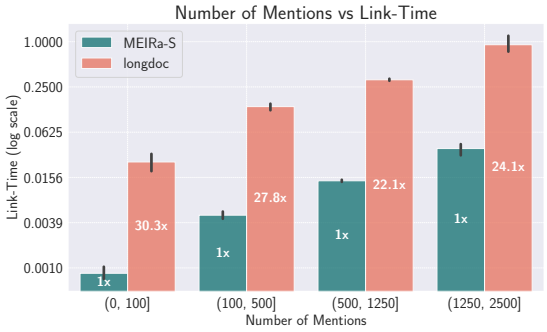


Figure 3.4: Linking speed comparison between MEIRa-S and longdoc.

MEIRa-H gaining 11.7% in Macro-F1 over the best baseline. The results demonstrate the efficacy of MEIRa models on resolving key entities in longer narratives.

3.6.5 Computational performance.

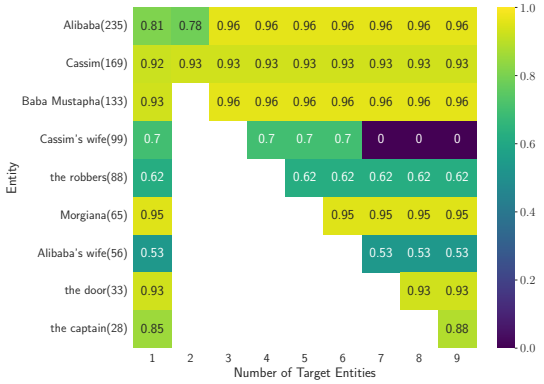
MEIRa-S supports parallel batched processing since it does not update the working memory after associating mentions, *i.e*. the mentions need not be processed sequentially from left to right. Hence, post-mention detection (common to all models), MEIRa-S is about 25× faster than longdoc when assessed across LitBank, FantasyCoref and AFT datasets on an NVIDIA RTX 4090 (see Fig. 3.4 in the appendix). Additionally, with the model's small memory footprint during inference, the entire process can also be parallelized across chunks of documents making it extremely efficient. Hence, we pose MEIRa-S as a faster while competitive alternative to MEIRa-H (that requires dynamic updates and has similar computational performance as longdoc).

The classification formulation and the lack of an update step in MEIRa-S makes it a more efficient alternative to MEIRa-H and CR models. Fig. 3.4 displays the speed-up obtained in the identification module when assessed across documents with varying numbers of mentions. MEIRa-S consistently clocks a 20x efficiency across all ranges.

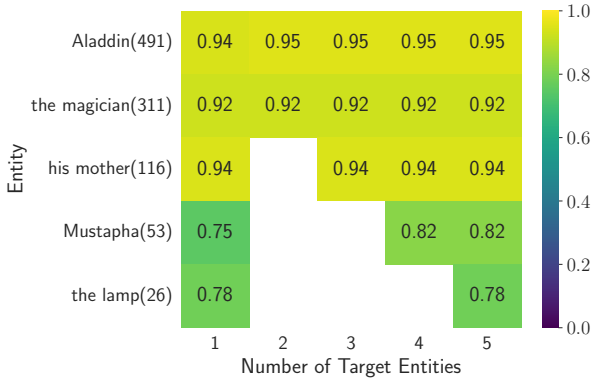
3.6.6 Performance across number of entities

For consistency, the experiments of the main paper are evaluated across all the selected major entities (chosen using the thresholds defined in Section 3.5). A natural extension is to assess the model's performance with varying numbers of entities of choice. For instance, if one is interested in only two key characters, can these models maintain consistency when provided with their designative phrases?

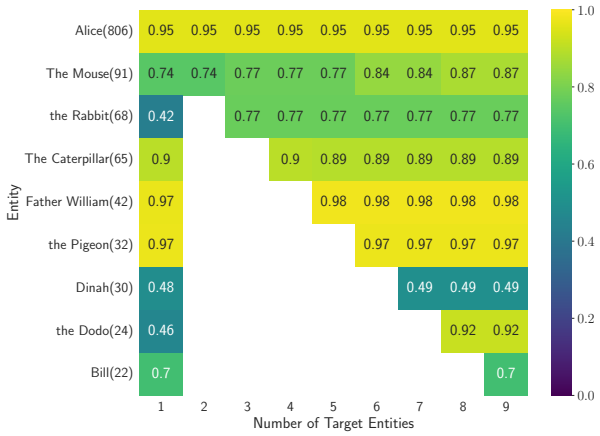
In this section, we address this concern and evaluate the MEI models with varying numbers of input entities. We present the per-entity F1-score of all entities across the AFT dataset. The results for MEIRa-H are showcased in Fig. 3.5a, Fig. 3.5b and Fig. 3.5c. The first column of the heatmap shows



(a) Ali Baba and the Forty Thieves



(b) Aladdin



(c) Alice in Wonderland

Figure 3.5: Performance of MEIRa-H across number of target entities.

the per-entity F1-score when it is the sole target entity in the document. For e.g., the value in the first column in Fig. 3.5a corresponding to the entity *Baba Mustapha* (0.93) indicates the performance of the model when *Baba Mustapha* is the only target entity.

As we move across the columns of a particular row (ignoring the first column), the column number indicates the number of target entities used at inference. For instance, if the column number is k, the target entities are the top-k frequent entities. Again, the 4^{th} column in the row corresponding to *Baba Mustapha* indicates its individual F1-score in the experiment where the four input entities are *Alibaba*, *Cassim*, *Baba Mustapha* and *Cassim's wife*.

There are a few individual cases where the performance significantly varies with modifying the number of input entities. For example, *Cassim's wife* is confused with *Alibaba's wife* after the latter's introduction. However, overall, the per-entity F1-score remains consistent across varying numbers of input entities across all three documents. These results demonstrate the effectiveness of MEIRa-H for applications requiring variable numbers of target entities. This consistency is mainly due to the variable

Model	FantasyCoref		LitBank	
	Macro-F1	Micro-F1	Macro-F1	Micro-F1
MEIRa-H	88.5	91.0	86.1	85.4
GPT-4	**90.7**	**92.0**	**88.8**	**91.6**
GPT-3.5	69.2	74.2	74.3	75.8
Code Llama-34B	67.0	72.4	68.9	73.1
Llama3-8B	53.8	60.6	50.2	53.4
Mistral-7B	67.3	75.8	61.6	73.9

Table 3.4: Few-shot LLM prompting results assuming the availability of ground-truth mentions.

entity training, where a randomly chosen subset of major entities is selected in each iteration. Excluding this procedure leads to significant fluctuation in performance while modifying the number of target entities.

3.7 Experiments: Few-shot prompting

3.7.1 Models.

We experiment with GPT-4[3] [58], GPT-3.5[4], Code Llama-34B [59], Mistral-7B [60], and Llama3-8B.[5] Following [4], we use the instruction-tuned versions for open-source models. These models were chosen for their ability to handle the extended context required for our benchmarks.

3.7.2 Linking Performance w/ Gold Mentions

We first evaluate all the models assuming the availability of an oracle mention detector. The experimental configuration is aligned with that of [4], albeit with the distinction that we assess them for the MEI task rather than for CR. The prompt used in our setup is provided in Table 3.10. For comparison, we also perform inference on golden mentions with MEIRa-H.

The results in Table 3.4 show that GPT-4 surpasses the supervised MEIRa-H model in this setup. Among LLMs, GPT-4 is easily the best-performing model. Code Llama-34B performs the best among open-source models, closely followed by Mistral-7B. While Code Llama-34B is tailored for the code domain, surprisingly, it outperforms strong LLMs suited for natural language. This result corroborates a similar finding by [4] for CR and related evidence regarding code pretraining aiding entity tracking [61].

[3]Specifically, `gpt-4-1106-preview`

[4]Specifically, `gpt-3.5-turbo-0125`

[5]https://ai.meta.com/blog/meta-llama-3/

Model	FantasyCoref		LitBank	
	Macro-F1	Micro-F1	Macro-F1	Micro-F1
MEIRa-H	**80.3**	**84.3**	82.3	83.2
GPT-4 w/ Ext det	80.1	82.2	78.7	83.9
GPT-4 with varying prompting strategies				
Single prompt	51.8	57.5	61.1	70.7
Two-stage prompt	70.5	74.9	76.5	81.3
Word-level MEI + spaCy H2S				
GPT-4	77.1	79.4	**82.5**	**85.5**
GPT-3.5	50.1	54.4	60.1	63.1
Code Llama-34B	30.0	31.4	22.7	23.2
Llama3-8B	29.2	32.1	20.5	26.0
Mistral-7B	19.4	21.9	12.9	14.0

Table 3.5: Results on LLMs with different mention detection and linking strategies.

We find that Code Llama-34B performs close to GPT-3.5 for FantasyCoref, though a sizable gap persists in the Macro-F1 metric for LitBank, potentially due to its linguistic complexity.

3.7.3 MEI Task Performance with LLMs

In this section, we present the results for the end-to-end MEI task using LLMs. We compare all the approaches from Section 3.4 and relevant baselines with the results summarized in Table 3.5. To limit the combinations of LLMs and approaches for our experiments, we first compare all the approaches in tandem with GPT-4 and then present results for the best-performing approach with other LLMs.

The first straightforward approach of using a *Single Prompt* to retrieve all the mentions of major entities in a single pass results in a significant performance drop compared to MEIRa-H (prompt in Table 3.11). The reason is that while GPT-4 outperforms MEIRa-H at mention linking, its mention detection performance, especially with nested mentions, is much worse compared to MEIRa-H.[6]

To further underscore the importance of mention detection, we compare against *GPT-4 w/ Ext det*, which utilizes an external pre-trained mention detector followed by prompt-based linking (prompt in Table 3.10 of Appendix). We train the mention detector on the PreCo dataset [62], which achieves a 93.8% recall and 53.1% precision on the combined FantasyCoref and LitBank validation sets. *GPT-4 w/ Ext det* performs at par with the fully supervised MEIRa-H, again highlighting the strong mention linking capabilities of GPT-4.

[6]The failure to detect nested mentions is despite best efforts to provide illustrative examples in the few-shot prompt. [4] report similar findings with earlier GPT versions.

Error Type	MEIRa-H	GPT-4
Missing Major	162	793
Major-Major	210	154
Major-Other	243	0
Other-Major	200	516
Extra-Major	461	896
Total	1276	2359

Table 3.6: Error Analysis of GPT-4 and MEIRa-H

(a) Breakdown of errors by MEIRa-H and GPT-4 on the combined LitBank and FantasyCoref test set.

Next, we present the results of our proposed *Two-stage prompt*, motivated by the *Single prompt* method's failure with nested mentions. The first prompt asks GPT-4 to perform word-level MEI, by limiting the task to syntactic heads only. The second prompt then performs the task of mapping the identified syntactic heads to full mention spans. The results strongly validate our proposed approach with a relative improvement of more than 10% over the *Single prompt* method across all metrics and datasets. We also explore replacing the second step, i.e., head-to-span (H2S) retrieval, with an external tool. Specifically, we invert `spaCy`'s span-to-head mapping to obtain a head-to-span retriever.[7]

GPT-4 significantly improves in this setup, outperforming even the supervised model on LitBank. Given the strong performance of *GPT-4 + spaCy H2S*, we evaluate the open-source LLMs in only this setting. We observe a wide gap between GPT-4 and the open-source models. Llama3-8B outperforms other open-source models on both datasets in Micro-F1 and stays competitive with the larger Code Llama-34B in Macro-F1. However, this contrasts with Llama3-8B's significant lag in the idealized golden mention setting, which solely evaluates the model's linking capabilities.

3.8 Error Analysis

We classify MEI errors into five categories: (1) *Missing Major:* Not detecting a mention $m \in \mathcal{M}$. (2) *Major-Major:* Assigning a mention $m \in \mathcal{M}_j$ to any other major entity $\mathcal{E} \setminus e_j$. (3) *Major-Other:* Assigning a mention $m \in \mathcal{M}$ to $\varnothing$. (4) *Other-Major:* Assigning a mention $m \in \mathcal{M}_{\text{other}}$ to any major entity in $\mathcal{E}$. (5) *Extra-Major:* Detecting extra mentions $m \notin \mathcal{M}_{\text{all}}$ and assigning to any major entity in $\mathcal{E}$.

Results combined over the LitBank and FantasyCoref test sets are presented in Table 3.6a. Missing Major and Extra-Major contribute most of the errors for GPT-4, highlighting the scope for improvement in mention detection and span retrieval. Mention detection also remains a challenge in MEIRa-H, the

[7]For the test set gold mentions of the two datasets, there were only two cases where spans had the same head. We handled these two cases manually.

Golden Mentions	Presently [a small boy]$_0$ came walking along the path – [an urchin of nine or ten]$_0$ ..... [Winterbourne]$_1$ had immediately perceived that [he]$_1$ might have the honor of claiming [him]$_2$ as a fellow countryman. "Take care [you]$_2$ don't hurt [your]$_2$ teeth," [he]$_1$ said, paternally ..... [My]$_2$ mother counted them last night, and one came out right afterwards. She said she'd slap [me]$_2$ if any more came out. [I]$_2$ can't help it. It's this old Europe ..... If [you]$_2$ eat three lumps of sugar, [your]$_2$ mother will certainly slap [you]$_2$," [he]$_1$ said. "She's got to give [me]$_2$ some candy, then," rejoined [[his]$_1$ young interlocutor]$_2$.
GPT-4 Output	Presently [a small boy]$_0$ came walking along the path – [an urchin of nine or ten]$_0$ ..... [Winterbourne]$_1$ had immediately perceived that [he]$_1$ might have the honor of claiming [him]$_2$ as a fellow countryman. "Take care you don't hurt your teeth," [he]$_1$ said, paternally ..... [My]$_2$ mother counted them last night, and one came out right afterwards. [She]$_2$ said [she]$_2$'d slap [me]$_2$ if any more came out. [I]$_2$ can't help it. [It]$_2$'s this old Europe ..... If you eat three lumps of sugar, [your]$_2$ mother will certainly slap [you]$_2$," [he]$_1$ said. "[She]$_2$'s got to give [me]$_2$ some candy, then," rejoined [his]$_2$ young interlocutor.
MEIRa-H Output	Presently a small boy came walking along the path – [an urchin of nine or ten] ..... [Winterbourne]$_1$ had immediately perceived that [he]$_1$ might have the honor of claiming [him]$_2$ as a fellow countryman. "Take care [you]$_2$ don't hurt [your]$_2$ teeth," [he]$_1$ said, paternally ..... [My]$_2$ mother counted them last night, and one came out right afterwards. She said she'd slap [me]$_2$ if any more came out. [I]$_2$ can't help it. It's this old Europe ..... If [you]$_2$ eat three lumps of sugar, [your]$_2$ mother will certainly slap [you]$_2$," [he]$_1$ said. "She's got to give [me]$_2$ some candy, then," rejoined [[his]$_1$ young interlocutor]$_2$.

Table 3.7: Qualitative Analysis showcasing different errors made by GPT-4 and MEIRa-H.

(a) Errors are color-coded as follows: Missing Major, Others-Major, Extra-Major, Major-Major, and Major-Other

model making most of the mistakes in the Extra-Major category. GPT-4 distinguishes major entities more clearly than MEIRa-H but tends to over-associate other mentions with major entities, resulting in higher Other-Major and Extra-Major errors. Note that GPT-4 has zero errors in the Major-Other category due to the prompt design, which only allows annotating major entities. Examples of these errors are visualized in Table 3.7a.

3.9 Conclusion

CR models are limited in their generalization capabilities owing to annotation differences and general challenges of domain adaptation. We propose MEI as an alternative to CR, where the entities relevant to the input text are provided as input along with the text. Our experiments demonstrate that MEI is more suited for generalization than CR. Additionally, MEI can be viewed as a classification task that enables the use of intuitive metrics.

A trivially parallelized variation (MEIRa-S), gives a 25x speedup over a comparable CR model, making it more suitable for longer narratives. Unlike CR, the formulation of MEI allows few-shot prompted LLMs to effectively compete with trained models. Our novel two-stage prompting and robust baseline methods empower top-performing LLMs like GPT-4 to achieve this. Our analysis indicates that this task holds promise for effectively evaluating the long-context referential capabilities of LLMs in an end-to-end manner.

Potential applications of MEI include domains such as film and literature, where metadata about salient entities can be sourced from external databases like IMDb or SparkNotes. Additionally, MEI can be applied to the analysis of documents like of financial and legal reports, when the user is familiar with the relevant entities. Lastly, recent research [63] indicates that LLMs can assist or automate the extraction of salient entities, a direction we intend to explore in future work.

3.10 Prompts

We provide exact prompts for all the few-shot prompting experiments. Please note that not all the major entities listed in the few shot examples are necessary to be present in the text.

Instruction

You will receive a Text along with a list of Key Entities and their corresponding Cluster IDs as input. Your task is to perform Coreference Resolution on the provided text to categorize "each word belonging to a cluster" with its respective cluster id. Also briefly describe the key entities in 1-2 sentences before starting the coreference task.
Follow the format below to label a word with its cluster ID:
word#cluster_id
Please keep in mind:
- Ensure the output adheres to the specified format for easy parsing.
- Classify the words in the given text without altering any of the other content.

Example Input:

Key Entities:
1. Katharine Hilbery - #1
2. Mr. Denham - #2
3. Mrs. Hilbery - #3
4. Mr. Hilbery - #4
5. Mr. Fortescue - #5

Text:
CHAPTER I It was a Sunday evening in October , and in common with many other young ladies of her class , Katharine Hilbery was pouring out tea . Perhaps a fifth part of her mind was thus occupied , and the remaining parts leapt over the little barrier of day which interposed between Monday morning and this rather subdued moment , and played with the things one does voluntarily and normally in the daylight . But although she was silent , she was evidently mistress of a situation which was familiar enough to her , and inclined to let it take its way for the six hundredth time , perhaps , without bringing into play any of her unoccupied faculties . A single glance was enough to show that Mrs. Hilbery was so rich in the gifts which make tea-parties of elderly distinguished people successful , that she scarcely needed any help from her daughter , provided that the tiresome business of teacups and bread and butter was discharged for her .

Example Output:

Description of Key Entities present in the text:
#1 - Katharine Hilbery: A young and apparently rich lady and the daughter of Mrs. Hilbery. She and Mrs. Hilbery were organising a party for some distinguished elders.
#3 - Mrs. Hilbery: She is the mother of Katharine Hilbery and is a well-to-do member of the society and a very efficient and able hostess

Coreference:
CHAPTER I It was a Sunday evening in October , and in common with many other young ladies of her#1 class , Katharine#1 Hilbery#1 was pouring out tea . Perhaps a fifth part of her#1 mind was thus occupied , and the remaining parts leapt over the little barrier of day which interposed between Monday morning and this rather subdued moment , and played with the things one does voluntarily and normally in the daylight . But although she#1 was silent , she#1 was evidently mistress of a situation which was familiar enough to her#1 , and inclined to let it take its way for the six hundredth time , perhaps , without bringing into play any of her#1 unoccupied faculties . A single glance was enough to show that Mrs.#3 Hilbery#3 was so rich in the gifts which make tea-parties of elderly distinguished people successful , that she#3 scarcely needed any help from her#3 daughter#1 , provided that the tiresome business of teacups and bread and butter was discharged for her#1 .

Table 3.8: Prompt for WL Coreference

Instruction
Any word marked with # is supposed to be the head of a noun phrase. Expand this head to contain determiner and adjective phrases. Do not remove or add new words while expanding. Stick to the format.
Example Input:
Montraville# was a **Lieutenant#** in the **army#** : **Belcour#** was his brother **officer#** : they had been to take leave of their **friends#** previous to their departure for **America#** , and were now returning to **Portsmouth#** , where the **troops#** waited orders for embarkation
Example Output:
Montraville **(Montraville)** was a Lieutenant **(a Lieutenant in the army)** in the army **(the army)** : Belcour **(Belcour)** was his brother officer **(his brother officer)** : they had been to take leave of their friends **(their friends)** previous to their departure for America **(America)** , and were now returning to Portsmouth **(Portsmouth)** , where the troops **(the troops)** waited orders for embarkation
Example Input:
Arriving at the verge of the **town#** , he dismounted , and sending the **servant#** forward with the horses , proceeded toward the **place#** , where , in the midst of an extensive pleasure **ground#** , stood the **mansion#** which contained the lovely Charlotte **Temple#** .
Example Output:
Arriving at the verge of the town **(the town)** , he dismounted , and sending the servant **(the servant)** forward with the horses , proceeded toward the place **(the place)** , where , in the midst of an extensive pleasure ground **(an extensive pleasure ground)** , stood the mansion **(the mansion which contained the lovely Charlotte Temple)** which contained the lovely Charlotte Temple **(the lovely Charlotte Temple)** .
Example Input:
"You are a benevolent **fellow#** ," said a young **officer#** to him one day and I have a great mind to give you a fine subject to exercise the goodness of your heart upon.
Example Output:
"You are a benevolent fellow **(a benevolent fellow)** ," said a young officer **(a young officer)** to him one day and I have a great mind to give you a fine subject to exercise the goodness of your heart upon.

Table 3.9: Prompt for H2S Retrieval

Instruction

Annotate all the entity mentions in the following text with coreference clusters. Use Markdown tags to indicate clusters in the output, with the following format [mention] (#cluster_name). Do not modify any text outside (), only add text inside parenthesis. The cluster names of the key entities are already provided, mark the mentions of the entity with the corresponding cluster name. Mark the mentions of the other entities with (#others). Also briefly describe the key entities in 1-2 sentences before starting the coreference task.

Example Input:

Key Entities:
1. Katharine Hilbery (#katharine_hilbery)
2. Mr. Denham (#mr._denham)
3. Mrs. Hilbery (#mrs._hilbery)
4. Mr. Hilbery (#mr._hilbery)
5. Mr. Fortescue (#mr._fortescue)

Text:
CHAPTER I It was a Sunday evening in October, and in common with [many other young ladies of [her] (#) class] (#) , [Katharine Hilbery] (#) was pouring out tea . Perhaps a fifth part of [her] (#) mind was thus occupied , and the remaining parts leapt over the little barrier of day which interposed between Monday morning and this rather subdued moment , and played with the things one does voluntarily and normally in the daylight . But although [she] (#) was silent , [she] (#) was evidently [mistress] (#) of a situation which was familiar enough to [her] (#) , and inclined to let it take its way for the six hundredth time , perhaps , without bringing into play any of [her] (#) unoccupied faculties . A single glance was enough to show that [Mrs. Hilbery] (#) was so rich in the gifts which make tea-parties of [elderly distinguished people] (#) successful , that [she] (#) scarcely needed any help from [[her] (#) daughter] (#) , provided that the tiresome business of teacups and bread and butter was discharged for [her] (#) .

Example Output:

Description of Key Entities present in the text:

#katharine_hilbery - Katharine Hilbery: A young and apparently rich lady and the daughter of Mrs. Hilbery. She and Mrs. Hilbery were organising a party for some distinguished elders.

#mrs._hilbery - Mrs. Hilbery: She is the mother of Katharine Hilbery and is a well-to-do member of the society and a very efficient and able hostess

Coreference:
CHAPTER I It was a Sunday evening in October , and in common with [many other young ladies of [her] (#katharine_hilbery) class] (#others) , [Katharine Hilbery] (#katharine_hilbery) was pouring out tea . Perhaps a fifth part of [her] (#katharine_hilbery) mind was thus occupied , and the remaining parts leapt over the little barrier of day which interposed between Monday morning and this rather subdued moment , and played with the things one does voluntarily and normally in the daylight . But although [she] (#katharine_hilbery) was silent , [she] (#katharine_hilbery) was evidently [mistress] (#others) of a situation which was familiar enough to [her] (#katharine_hilbery) , and inclined to let it take its way for the six hundredth time , perhaps , without bringing into play any of [her] (#katharine_hilbery) unoccupied faculties . A single glance was enough to show that [Mrs. Hilbery] (#mrs._hilbery) was so rich in the gifts which make tea-parties of [elderly distinguished people] (#others) successful , that [she] (#mrs._hilbery) scarcely needed any help from [[her] (#mrs._hilbery) daughter] (#katharine_hilbery) , provided that the tiresome business of teacups and bread and butter was discharged for [her] (#katharine_hilbery) .

Table 3.10: Prompt for evaluating linking performance

Instruction

Annotate all the entity mentions that refer to the key entities provided. The mention needs to include determiners and adjectives, if present. Use Markdown tags to indicate clusters in the output, with the following format [mention] (#cluster_name). The cluster names of the key entitites are already provided. Mark the mentions of the entity with the corresponding cluster name. Also briefly describe the key entities in 1-2 sentences before starting the coreference task.

Example Input:

Key Entities:
1. Katharine Hilbery (#katharine_hilbery)
2. Mr. Denham (#mr._denham)
3. Mrs. Hilbery (#mrs._hilbery)
4. Mr. Hilbery (#mr._hilbery)
5. Mr. Fortescue (#mr._fortescue)

Text:
CHAPTER I It was a Sunday evening in October , and in common with many other young ladies of her class , Katharine Hilbery was pouring out tea . Perhaps a fifth part of her mind was thus occupied , and the remaining parts leapt over the little barrier of day which interposed between Monday morning and this rather subdued moment , and played with the things one does voluntarily and normally in the daylight . But although she was silent , she was evidently mistress of a situation which was familiar enough to her , and inclined to let it take its way for the six hundredth time , perhaps , without bringing into play any of her unoccupied faculties . A single glance was enough to show that Mrs. Hilbery was so rich in the gifts which make tea-parties of elderly distinguished people successful , that she scarcely needed any help from her daughter , provided that the tiresome business of teacups and bread and butter was discharged for her .

Example Output:

Description of Key Entities present in the text:

#katharine_hilbery - Katharine Hilbery: A young and apparently rich lady and the daughter of Mrs. Hilbery. She and Mrs. Hilbery were organising a party for some distinguished elders.

#mrs._hilbery - Mrs. Hilbery: She is the mother of Katharine Hilbery and is a well-to-do member of the society and a very efficient and able hostess

Coreference:
CHAPTER I It was a Sunday evening in October , and in common with many other young ladies of [her] (#katharine_hilbery) class , [Katharine Hilbery] (#katharine_hilbery) was pouring out tea . Perhaps a fifth part of [her] (#katharine_hilbery) mind was thus occupied , and the remaining parts leapt over the little barrier of day which interposed between Monday morning and this rather subdued moment , and played with the things one does voluntarily and normally in the daylight . But although [she] (#katharine_hilbery) was silent , [she] (#katharine_hilbery) was evidently mistress of a situation which was familiar enough to [her] (#katharine_hilbery) , and inclined to let it take its way for the six hundredth time , perhaps , without bringing into play any of [her] (#katharine_hilbery) unoccupied faculties . A single glance was enough to show that [Mrs. Hilbery] (#mrs._hilbery) was so rich in the gifts which make tea-parties of elderly distinguished people successful , that [she] (#mrs._hilbery) scarcely needed any help from [[her] (#mrs._hilbery) daughter] (#katharine_hilbery) , provided that the tiresome business of teacups and bread and butter was discharged for [her] (#katharine_hilbery) .

Table 3.11: Prompt for Direct version of E2E MEI

Chapter *4*

IdentifyMe: A Challenging Mention Resolution Benchmark for LLMs

4.1 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 [37, 58, 64], 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, [4] 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 [5] mitigate the variability of entity definition by assuming major entities as inputs, their evaluation remains limited by dataset-specific span boundaries. Recent work by [6] 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 [65] 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 [46] and FantasyCoref [23]. 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 4.3). Each MCQ instance consists of text marked with the mention of interest and choices

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	None of the Above

Answer: Ogden Ford

Figure 4.1: Sample instance from the validation set of `IdentifyMe`.

(a) The mention of interest is highlighted in the text. The answer options include frequently occurring entities in the text, and None of the Above.

comprising frequently occurring entities in the text and the *None of the Above* (NoA) option. Fig. 4.1a 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.*, his 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.

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

Table 4.1: Performance of LLMs on IdentifyMe val split vs Random split (10 runs)

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 4.2: Validation accuracy of LLMs w/ and w/o CoT.

4.2 IdentifyMe Benchmark

IdentifyMe is an MCQ-based benchmark where given a document with a marked mention, the task is to identify the entity to which it refers. 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 changes in title (*e.g.*, *Prince Rudolph* is called *The Emperor* after his coronation), which further complicate entity mapping.

Coreference annotations group mentions that refer to the same entity into unlabeled clusters. However, to create an MCQ with entities as options, we need to assign each cluster a representative phrase. We use GPT-4o-mini (see Table 4.8) to generate phrases for each entity based on their mentions and frequencies. These annotations are manually reviewed to ensure a distinct phrase for each entity.

To prevent confusion, we discard and avoid labeling clusters that: (i) contain annotation errors (*e.g.*, due to cluster merging or splitting [50]); (ii) are too small (< 3 mentions) or difficult/ambiguous to label (*e.g.*, entitites 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 create a high quality benchmark, we exclude short context documents (< 1000 words) or those where the discarded entities represent more than 50% of the mentions.

4.3 Selecting Mentions for IdentifyMe

Based on previous works which utilize rule-based linguistic patterns to perform [66, 67] or analyze [68, 17] coreference resolution, we propose a two-step heuristic to identify challenging mentions.

4.3.1 Step 1: Discard easy mentions.

We apply two criteria to filter out mentions that can be easily resolved due to syntactic similarity:

A. Nominal fuzzy score calculates the fuzzy similarity[1] (flexible to order and subset perturbations) between a nominal mention and the corresponding entity's representative phrase. Mentions with a score of 75% or higher are discarded, as we expect that they are easier to correctly identify.

B. Net distractor score. We categorize pronominal mentions based on attributes like gender, number, and animacy (LingMess [17]). Nearby pronominal mentions with the same category as the marked mention (pronominal), and referring to the same entity, may help identify the mention easily. On the other hand, those from a different category but same entity *or* same category but different entity may make it harder to identify. We define the `Net-Distractor-Score` of the marked mention as the number of neighboring pronominals that hinder identification minus the number that help it. We discard mentions that score ≤ 0.

4.3.2 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 [69], evaluating pairwise wins and losses to determine the final ranking.

4.4 Dataset Statistics

`IdentifyMe` includes the 1800 most difficult questions based on the ranking method described above. These questions come from 159 documents (64 from LitBank, 95 from FantasyCoref). Of these, 600 randomly selected questions are used as the validation set for prompt tuning and ablation experiments. A *None of the Above (NoA)* option is included with all the questions to encourage models to respond with greater certainty. To evaluate whether the models can pick NoA as the correct choice, we remove the (original) correct entity for 10% of the questions. The dataset is balanced across the source datasets and mention-types (pronominals and nominals) in both the validation and test splits.

4.5 How difficult is `IdentifyMe`?

We conduct an experiment to assess the effectiveness of our mention selection process. Instead of the approach outlined above, we pick mentions at random and evaluate the ability of models to identify them. The performance gap of 9.5% for `Mistral-7B` and 7.2% for the more robust `GPT-4o-mini` indicates that `IdentifyMe` contains more challenging mentions (see Table 4.1).

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

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

Table 4.3: Comparing the performance of LLMs and Humans on a subset of 50 questions

We perform human evaluation on a randomly selected subset of 50 questions from FantasyCoref (FC), spanning 10 documents—comprising 25 nominal and 25 pronominal instances. As seen in Table 4.3, a significant performance gap (23%) is observed between humans and the best LLM (GPT-4o), highlighting substantial room for improvement.

4.6 Experiments

Models. Among closed-source models, we evaluate GPT-4o [58], GPT-4o-mini [70], and Gemini-1.5-Flash[2] [64]. Due to computational constraints, we limit the evaluation of open-source models to sub-10B variants: Llama-3.1-8B [71] and Mistral-7B [60].

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 4.10).

Inference details. For open-source models, we use regex-based constrained decoding with the outlines library [56] to limit answers to specific entity representative phrases. We also experiment with a chain-of-thought (CoT) approach [72], instructing the model to explain its reasoning before answering the question. As seen in Table 4.2, 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 and decoding regular expressions, see Appendix 4.10.

4.7 Results

Table 4.4 presents the overall LLM performance on the IdentifyMe test set, along with a breakdown by nominal and pronominal mention types. The Random baseline, selecting answers uniformly at random, achieves 8% on our benchmark. Although all LLMs outperform Random, 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

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

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 4.4: Performance of various models on the `IdentifyMe` test set.

Model	Nominal		Pronominal	
	FantasyCoref (300)	LitBank (300)	FantasyCoref (300)	LitBank (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-mini`	**79.3**	**91.0**	**81.3**	**76.0**

Table 4.5: Performance split by mention type and dataset source.

`Gemini-1.5-Flash`. Across mention types, all closed-source models perform significantly better at resolving nominal mentions than pronominal ones.

Table 4.5 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.

4.8 Error Analysis

Comparing entities *vs*. NoA. Table 4.6a 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

Model	Accuracy		#Misclassifications		
	Ent	NoA	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 4.6: LLM Performance on Ent. vs NoA

(a) 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).

Model	Accuracy		Span Error
	Non-nested	Nested	
`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 4.7: LLM performance on nested mentions versus non-nested mentions

(a) The Span Error column indicates the error for nested mentions (352 of 1200) where the predicted entity corresponds to an overlapping mention.

(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, `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 4.7a 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 4.2a illustrates a sample nested mention error made by `GPT-4o`.

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

(a) An example nested mention error: 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*.

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

(b) An example dialogue error: 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.

Figure 4.2: Examples of errors in nested mention resolution and dialogue attribution.

4.8.1 Sample Outputs

Figure 4.2b presents another example error where the model is confused due to complicated first and second person references within dialog.

Figure 4.3a and 4.3b present instances which both GPT-4o and Llama3.1 get right. Their explanation makes sense as well.

4.9 Conclusion

We present IdentifyMe, a challenging MCQ benchmark designed for the evaluation of the referential capabilities of LLMs. Analysis shows that LLMs’ have difficulty with: (i) pronominal resolution without clear surface form clues, (i) questions that require rejecting all wrong options by selecting “None of the Above”, 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 assessing the referential capabilities of state-of-the-art LLMs.

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

(a) 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

(b) Sample instance from IdentifyMe that both GPT-4o and Llama-3.1 get right.

Figure 4.3: Sample instances from IdentifyMe that both GPT-4o and Llama-3.1 correctly identified.

4.10 Prompts

Table 4.8 presents the prompt passed to GPT-4o-mini to generate representative phrases for each cluster/entity. Table 4.9 and Table 4.10 present the prompts and regular expressions used for inference with and without CoT respectively.

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 4.8: 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 4.9: 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 4.10: QA prompt without CoT.

Chapter *5*

Conclusion and Future Scope

In this thesis, we introduced a novel referential task, Major Entity Identification (MEI), and a new MCQ benchmark called `IdentifyMe`. These contributions advance our understanding of coreference resolvers and large language models (LLMs). The task formulations address common *bells and whistles* in coreference resolution, such as inconsistencies in what constitutes a mention, variations in span boundaries, and challenges in reproducing spans and documents for LLM evaluation. MEI is a new referential task that assumes that the frequent target entities are explicitly provided in the input. This resolves issues pertaining to varied mention and entity definitions across benchmarks, ensuring a fairer evaluation of referential capabilities. Conversely, `IdentifyMe` comprises MCQ questions designed to assess the *mention resolution* abilities of LLMs, devoid of any *bells and whistles*. Therefore, any shortcomings on this benchmark can be attributed solely to *mention resolution* failures.

5.1 Conclusion

We conclude this thesis by highlighting the major contributions. From the work on Major Entity Identification, we highlight the following findings:

- Our experiments demonstrate that MEI is more suited for generalization than CR.
- MEI can be viewed as a classification task that enables the use of intuitive metrics.
- MEIRa-S gives a 25x speedup over a comparable CR model, making it more suitable for longer narratives.
- Unlike CR, the formulation of MEI allows few-shot prompted LLMs to effectively compete with trained models.
- Our novel two-stage prompting and robust baseline methods empower top-performing LLMs like GPT-4 to achieve this.

- Our analysis indicates MEI holds promise for effectively evaluating the long-context referential capabilities of LLMs in an end-to-end manner.
- Applications of MEI include domains such as film and literature, where metadata about salient entities can be sourced from external databases like IMDb or SparkNotes.

We then continue our study on the evaluation of LLMs for their referential capability in a more faithful setup. For this, we propose `IdentifyMe` and make the following observations:

- `IdentifyMe` is a challenging MCQ benchmark designed for the evaluation of the referential capabilities of LLMs.
- We present analysis that shows that LLMs' have difficulty with pronominal resolution without clear surface form clues,
- We also note poor performance in questions that require rejecting all wrong options by selecting "None of the Above".
- LLMs also struggle with nested mentions that require distinguishing between overlapping spans.
- `GPT-4o` scores 81.9% on `IdentifyMe`, highlighting the strong referential capabilities of *frontier* LLMs.
- However, on a sample of a benchmark where human evaluations were conducted, `GPT-4o` falls short of human performance by more than 23%, leaving ample room for improvement.

Our work highlights the need to adapt traditional coreference resolution formulations to evolving applications and requirements. Major Entity Identification offers a step toward standardizing annotation efforts in this domain. Together, our contributions facilitate clearer evaluation of referential capabilities, highlighting both strengths and limitations of models. We hope this work motivates future research to develop improved coreference systems for applications such as story understanding, legal and financial analysis, and to assess LLMs on core linguistic tasks, ultimately enhancing their performance across downstream applications.

5.2 Future Scope

Some directions for future work:

- Greater emphasis should be placed on the applications of MEI and the adaptations required to cater to different domains.
- In our work on MEI, few-shot prompting evaluations are restricted to long-context models. Future work will focus on optimizing prompts and architecture to enable piecewise aggregation of outputs across document chunks.

- The current `IdentifyMe` benchmark is limited in its domain (literary domain), mention type coverage (only nominal and pronominal mentions), and also entity types (no plural entities). Future work could involve expanding `IdentifyMe` across these dimensions.
- Data contamination is a significant concern when evaluating LLMs trained on internet data. Our initial analysis indicates that the benchmark is not affected by contamination from heavily processed coreference annotations. However, there remains a small chance of contamination. To mitigate this risk, we could annotate documents that are not available on the internet for future benchmark collection.

Related Publications

1. **Kawshik Manikantan**, Shubham Toshniwal, Makarand Tapaswi, and Vineet Gandhi. 2024. Major Entity Identification: A Generalizable Alternative to Coreference Resolution. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 11679–11695, Miami, Florida, USA. Association for Computational Linguistics.

2. **Kawshik Manikantan**, Makarand Tapaswi, Vineet Gandhi, and Shubham Toshniwal. 2025. IdentifyMe: A Challenging Long-Context Mention Resolution Benchmark. In Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics, Albuquerque, New Mexico, USA. Association for Computational Linguistics arXiv preprint arXiv:2411.07466

Bibliography

[1] Bernd Bohnet, Chris Alberti, and Michael Collins. Coreference Resolution through a seq2seq Transition-Based System. *TACL*, 2023.

[2] Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[3] Wenzheng Zhang, Sam Wiseman, and Karl Stratos. Seq2seq is All You Need for Coreference Resolution. In *EMNLP*, 2023.

[4] Nghia T. Le and Alan Ritter. Are Large Language Models Robust Coreference Resolvers?, 2023.

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

[6] Yujian Gan, Massimo Poesio, and Juntao Yu. Assessing the Capabilities of Large Language Models in Coreference: An Evaluation. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1645–1665, Torino, Italia, May 2024. ELRA and ICCL.

[7] Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. Large Language Models are Few-Shot Clinical Information Extractors. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *EMNLP*, 2022.

[8] Ian Porada, Alexandra Olteanu, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung. Investigating Failures to Generalize for Coreference Resolution Models, 2023.

[9] Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. Towards robust linguistic analysis using OntoNotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[10] Nafise Sadat Moosavi, Leo Born, Massimo Poesio, and Michael Strube. Using automatically extracted minimum spans to disentangle coreference evaluation from boundary detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4168–4178, Florence, Italy, July 2019. Association for Computational Linguistics.

[11] Ben Kantor and Amir Globerson. Coreference resolution with entity equalization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 673–677, Florence, Italy, July 2019. Association for Computational Linguistics.

[12] Greg Durrett and Dan Klein. Easy victories and uphill battles in coreference resolution. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1971–1982, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

[13] Kevin Clark and Christopher D. Manning. Entity-centric coreference resolution with model stacking. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1405–1415, Beijing, China, July 2015. Association for Computational Linguistics.

[14] Kevin Clark and Christopher D. Manning. Improving coreference resolution by learning entity-level distributed representations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653, Berlin, Germany, August 2016. Association for Computational Linguistics.

[15] Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel Weld. BERT for coreference resolution: Baselines and analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5803–5808, Hong Kong, China, November 2019. Association for Computational Linguistics.

[16] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77, 2020.

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

[18] Vladimir Dobrovolskii. Word-level coreference resolution. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7670–7675, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[19] Congying Xia, Wenpeng Yin, Yihao Feng, and Philip Yu. Incremental few-shot text classification with multi-round new classes: Formulation, dataset and system. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human*

Language Technologies, pages 1351–1360, Online, June 2021. Association for Computational Linguistics.

[20] Shubham Toshniwal, Sam Wiseman, Allyson Ettinger, Karen Livescu, and Kevin Gimpel. Learning to Ignore: Long Document Coreference with Bounded Memory Neural Networks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8519–8526, Online, November 2020. Association for Computational Linguistics.

[21] Eduard Hovy, Mitchell Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. OntoNotes: The 90% solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York City, USA, June 2006. Association for Computational Linguistics.

[22] David Bamman, Sejal Popat, and Sheng Shen. An annotated dataset of literary entities. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2138–2144, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[23] Sooyoun Han, Sumin Seo, Minji Kang, Jongin Kim, Nayoung Choi, Min Song, and Jinho D. Choi. FantasyCoref: Coreference resolution on fantasy literature through omniscient writer's point of view. In *Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 24–35, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[24] Nafise Sadat Moosavi and Michael Strube. Lexical features in coreference resolution: To be used with caution. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 14–19, Vancouver, Canada, July 2017. Association for Computational Linguistics.

[25] Yilun Zhu, Sameer Pradhan, and Amir Zeldes. OntoGUM: Evaluating contextualized SOTA coreference resolution on 12 more genres. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 461–467, Online, August 2021. Association for Computational Linguistics.

[26] Marc Vilain, John Burger, John Aberdeen, Dennis Connolly, and Lynette Hirschman. A model-theoretic coreference scoring scheme. In *Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*, pages 45–52, 1995.

[27] Amit Bagga and Breck Baldwin. Algorithms for scoring coreference chains. In *The first international conference on language resources and evaluation workshop on linguistics coreference*, volume 1, pages 563–566, 1998.

[28] Xiaoqiang Luo. On coreference resolution performance metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 25–32, Vancouver, British Columbia, Canada, October 2005. Association for Computational Linguistics.

[29] M. Recasens and E. Hovy. Blanc: Implementing the rand index for coreference evaluation. *Nat. Lang. Eng.*, 17(4):485–510, oct 2011.

[30] Nafise Moosavi and Michael Strube. Which coreference evaluation metric do you trust? a proposal for a link-based entity aware metric. pages 632–642, 08 2016.

[31] Rahul Aralikatte, Heather Lent, Ana Valeria Gonzalez, Daniel Herschcovich, Chen Qiu, Anders Sandholm, Michael Ringaard, and Anders Søgaard. Rewarding coreference resolvers for being consistent with world knowledge. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1229–1235, Hong Kong, China, November 2019. Association for Computational Linguistics.

[32] Patrick Xia and Benjamin Van Durme. Moving on from OntoNotes: Coreference resolution model transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5241–5256, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[33] Shubham Toshniwal, Patrick Xia, Sam Wiseman, Karen Livescu, and Kevin Gimpel. On generalization in coreference resolution. In *Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 111–120, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[34] Shanheng Zhao and Hwee Tou Ng. Domain adaptation with active learning for coreference resolution. In *Proceedings of the 5th International Workshop on Health Text Mining and Information Analysis (Louhi)*, pages 21–29, Gothenburg, Sweden, April 2014. Association for Computational Linguistics.

[35] Michelle Yuan, Patrick Xia, Chandler May, Benjamin Van Durme, and Jordan Boyd-Graber. Adapting coreference resolution models through active learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7533–7549, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[36] Nupoor Gandhi, Anjalie Field, and Emma Strubell. Annotating Mentions Alone Enables Efficient Domain Adaptation for Coreference Resolution. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *ACL*, 2023.

[37] 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. Language Models are Few-Shot Learners. In *NeurIPS*, 2020.

[38] Xiaohan Yang, Eduardo Peynetti, Vasco Meerman, and Chris Tanner. What GPT knows about who is who. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, pages 75–81, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[39] Sabyasachee Baruah and Shrikanth Narayanan. Character Coreference Resolution in Movie Screenplays. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of ACL*, 2023.

[40] Yu-Hsin Chen and Jinho D. Choi. Character identification on multiparty conversation: Identifying mentions of characters in TV shows. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 90–100, Los Angeles, September 2016. Association for Computational Linguistics.

[41] Sayyed M. Zahiri and Jinho D. Choi. Emotion Detection on TV Show Transcripts with Sequence-based Convolutional Neural Networks, 2017.

[42] Heng Ji, Joel Nothman, Ben Hachey, and Radu Florian. Overview of tac-kbp2015 tri-lingual entity discovery and linking. *Theory and Applications of Categories*, 2015.

[43] Bhuwan Dhingra, Qiao Jin, Zhilin Yang, William Cohen, and Ruslan Salakhutdinov. Neural models for reasoning over multiple mentions using coreference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 42–48, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[44] Rik Koncel-Kedziorski, Dhanush Bekal, Yi Luan, Mirella Lapata, and Hannaneh Hajishirzi. Text Generation from Knowledge Graphs with Graph Transformers. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2284–2293, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[45] Eva Sharma, Luyang Huang, Zhe Hu, and Lu Wang. An entity-driven framework for abstractive summarization. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3280–3291, Hong Kong, China, November 2019. Association for Computational Linguistics.

[46] David Bamman, Olivia Lewke, and Anya Mansoor. An annotated dataset of coreference in English literature. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 44–54, Marseille, France, May 2020. European Language Resources Association.

[47] Nafise Sadat Moosavi and Michael Strube. Which coreference evaluation metric do you trust? a proposal for a link-based entity aware metric. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 632–642, Berlin, Germany, August 2016. Association for Computational Linguistics.

[48] M. Recasens and E. Hovy. Blanc: Implementing the rand index for coreference evaluation. *Nat. Lang. Eng.*, 17(4):485–510, oct 2011.

[49] Jing Lu and Vincent Ng. Conundrums in entity coreference resolution: Making sense of the state of the art. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6620–6631, Online, November 2020. Association for Computational Linguistics.

[50] Jonathan K. Kummerfeld and Dan Klein. Error-driven analysis of challenges in coreference resolution. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 265–277, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

[51] Naoya Inoue, Charuta Pethe, Allen Kim, and Steven Skiena. Learning and evaluating character representations in novels. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1008–1019, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[52] Maria Antoniak, Anjalie Field, Jimin Mun, Melanie Walsh, Lauren Klein, and Maarten Sap. Riveter: Measuring Power and Social Dynamics Between Entities. In *ACL (Volume 3: System Demonstrations)*, 2023.

[53] Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The Long-Document Transformer. *arXiv preprint arXiv:2004.05150*, 2020.

[54] Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength Natural Language Processing in Python, 2020.

[55] Saul B Needleman and Christian D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 1970.

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

[57] James Munkres. Algorithms for the assignment and transportation problems. *Journal of the society for industrial and applied mathematics*, 1957.

[58] OpenAI. GPT-4 Technical Report, 2024.

[59] Baptiste Rozière, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, Jérémy Rapin, Artyom Kozhevnikov, Ivan Evtimov, Joanna Bitton, Manish Bhatt, Cristian Canton Ferrer, Aaron Grattafiori, Wenhan Xiong, Alexandre Défossez, Jade Copet, Faisal Azhar, Hugo Touvron, Louis Martin, Nicolas Usunier, Thomas Scialom, and Gabriel Synnaeve. Code Llama: Open Foundation Models for Code, 2024.

[60] 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. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

[61] Najoung Kim, Sebastian Schuster, and Shubham Toshniwal. Code Pretraining Improves Entity Tracking Abilities of Language Models, 2024.

[62] Hong Chen, Zhenhua Fan, Hao Lu, Alan Yuille, and Shu Rong. PreCo: A large-scale dataset in preschool vocabulary for coreference resolution. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 172–181, Brussels, Belgium, October-November 2018. Association for Computational Linguistics.

[63] Jessica Lin and Amir Zeldes. GUMsley: Evaluating entity salience in summarization for 12 English genres. In Yvette Graham and Matthew Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2575–2588, St. Julian's, Malta, March 2024. Association for Computational Linguistics.

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

[65] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding. In *International Conference on Learning Representations*, 2021.

[66] GuoDong Zhou and Jian Su. A high-performance coreference resolution system using a constraint-based multi-agent strategy. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 522–528, Geneva, Switzerland, aug 23–aug 27 2004. COLING.

[67] Heeyoung Lee, Angel Chang, Yves Peirsman, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. Deterministic coreference resolution based on entity-centric, precision-ranked rules. *Computational Linguistics*, 39(4):885–916, December 2013.

[68] Aria Haghighi and Dan Klein. Simple coreference resolution with rich syntactic and semantic features. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1152–1161, Singapore, August 2009. Association for Computational Linguistics.

[69] A. Copeland. A reasonable social welfare function. 1951.

[70] OpenAI. Gpt-4o-mini: Advancing cost-efficient intelligence, 2024. Accessed: 2024-10-15.

[71] Meta-AI. The Llama 3 Herd of Models, 2024.

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