BioCoref: Benchmarking Biomedical Coreference Resolution with LLMs

Nourah M. Salem
University of Colorado Anschutz Medical Campus, USA
University of Chicago, USA
nourah(dot)salem@cuanschutz.edu

Elizabeth White and Michael Bada and Lawrence Hunter
University of Chicago, USA

Abstract

Coreference resolution in biomedical texts presents unique challenges due to complex domain-specific terminology, high ambiguity in mention forms, and long-distance dependencies between coreferring expressions. In this work, we present a comprehensive evaluation of generative large language models (LLMs) for coreference resolution in the biomedical domain. Using the CRAFT corpus as our benchmark, we assess the LLMs' performance with four prompting experiments that vary in their use of local, contextual enrichment, and domain-specific cues such as abbreviations and entity dictionaries. We benchmark these approaches against a discriminative span-based encoder, SpanBERT, to compare the efficacy of generative versus discriminative methods. Our results demonstrate that while LLMs exhibit strong surface-level coreference capabilities, especially when supplemented with domain-grounding prompts, their performance remains sensitive to long-range context and mentions ambiguity. Notably, the LLaMA 8B and 17B models show superior precision and F1 scores under entity-augmented prompting, highlighting the potential of lightweight prompt engineering for enhancing LLM utility in biomedical NLP tasks.

1 Introduction

Coreference resolution is the process of identifying entities mentioned in text and grouping all mentions that refer to the same underlying entity (Liu et al., 2023). In the biomedical domain, coreference resolution is a particularly difficult task and has been identified as a key bottleneck in biomedical text mining (Lu and Poesio, 2021). Research articles in this domain often contain dense, technical language, frequent use of abbreviations, and complex referential expressions that rely on domain-specific background knowledge. For instance, resolving a phrase like "the same strain" in a methods section may require linking it back to the "C57BL/6J mice" mentioned several paragraphs earlier, with no intervening repetition or synonyms. Similarly, phrases such as "the compound" may ambiguously refer to any of several chemical entities introduced earlier in experimental descriptions, particularly when multiple drugs or treatments are discussed in parallel. In such cases, surface string similarity offers little guidance; instead, linguistic disambiguation must be informed by contextual and semantic cues.

Adding to the challenge, many biomedical entities share identical surface forms, e.g., a gene and its corresponding protein often have the same name or abbreviation which can confuse automated systems. When clustered by identical surface strings, approximately 65% of the coreference clusters in CRAFT corpus (Cohen et al., 2017b) consist of repeated mentions (Li et al., 2022), emphasizing the need for models that can handle referential ambiguity. Moreover, many coreference links span large textual distances, exceeding the effective context window of conventional models (Lu and Poesio, 2021; Li et al., 2022). These long-range dependencies and requirements for specialized knowledge contribute to the poor generalization of general-domain coreference systems in biomedical contexts.

Given the emergence of increasingly capable large language models (LLMs), a natural question arises: how well can these general purpose models perform coreference resolution in specialized domains like biomedicine, without any task-specific fine-tuning (Gan et al., 2024)? LLMs have demonstrated remarkable abilities in complex reasoning and language understanding via prompt-based zero-shot or few-shot learning, often surpassing traditional models in general-domain tasks. This raises the possibility that their extensive pretraining enables them to handle intricate referential structures, even in domain-specific contexts. While biomedi-

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 519–530

cal coreference remains a demanding task, recent advances suggest that with well-designed prompts and minimal scaffolding, LLMs may be more capable than previously assumed.

In this work, we evaluate coreference resolution in biomedical texts using two contrasting approaches: a span-based model (SpanBERT-Large) (Joshi et al., 2020) trained on general-domain data, and several generative LLMs (LLaMA series) (Touvron et al., 2023) prompted to resolve coreference without fine-tuning. Our experiments use the CRAFT corpus, a richly annotated biomedical dataset.

The contributions and objectives of this paper are summarized as follows:

- **Benchmarking open-weight LLMs** We compare three LLaMA models under different prompting strategies: local-only, contextual, abbreviation-aware, and entity-aware against a span-based baseline, reporting performance on the CRAFT corpus.
- **Domain Analysis:** We identify coreference challenges unique to biomedical text, such as identical mention strings and abbreviation ambiguity, and analyze how each model type handles them through qualitative examples and error patterns.

2 Related Work

Coreference resolution in biomedical text is a particularly challenging task due to complex domain-specific terminology, high referential ambiguity, and long-range dependencies. Traditional span-based models such as the end-to-end neural coreference models such as SpanBERT (Joshi et al., 2020) have demonstrated strong performance in general domains. However, their reliance on limited context windows and the need for supervised fine-tuning limits their applicability in biomedical settings, where coreference often requires broader semantic grounding.

Traditional approaches also include rule-based and statistical systems (O'Connor and Heilman, 2013; Ng and Cardie, 2002; Soon et al., 2001), followed by neural architectures such as the mention-ranking model (Clark and Manning, 2016) and end-to-end span-ranking networks (Lee et al., 2017; Durrett and Klein, 2013; Wiseman et al., 2015; Lee et al., 2018). SpanBERT (Joshi et al., 2020) further improved coreference resolution by introducing span-centric pretraining objectives, achieving state-of-the-art results on the OntoNotes (Dobrovolskii, 2021) benchmark. Despite these advances, such models rely heavily on supervised training and domain-specific tuning, limiting their generalizability to out-of-domain settings like biomedical text.

Large language models (LLMs) like OpenAI GPT (Radford et al., 2018) and LLaMA have demonstrated strong zero-shot capabilities in various NLP tasks (Brown et al., 2020; Touvron et al., 2023), including aspects of coreference. Recent studies evaluated LLMs' abilities on pronoun resolution and Winograd schemas (Liu et al., 2024; Wu et al., 2020) for other downstream tasks such as question-answering and query-based span prediction problems. However, few studies have directly assessed LLMs on span-based or noun phrase coreference, particularly in long or technical documents. Most relevant to our work are recent prompting frameworks that use generative models for structured information extraction (Xie et al., 2022), though coreference-specific prompting remains underexplored, especially in specialized domains like biomedical literature.

3 Methods and Materials

3.1 Data

The Colorado Richly Annotated Full-Text (CRAFT) corpus (Cohen et al., 2017a) is a biomedical dataset of 67 full-text, open-access journal articles drawn from PubMed Central, with extensive manual annotations for biomedical concepts, syntactic structure, and coreference identity chains, In Version 2.0[1]. For our experiments, we evaluate language models on 50 articles selected at random, which contained more than 80K entities coreferenced to test the LLMs one.[2].

Three stacking properties make CRAFT uniquely demanding for large language models. First, its coreference chains stretch across *full* journal articles rather than short excerpts: as shown in Figure 1, while 76.14% of coreferential links fall within 500 words, over 23% extend up to 12,000 words, well beyond the effective context window of most conventional neural architectures. Second, roughly 65% of its coreference clusters consist of mentions with identical surface strings, for example, a gene and its cognate protein sharing a name or abbreviation, forcing models to disambiguate

[1] https://github.com/UCDenver-ccp/CRAFT
[2] https://github.com/biocoref

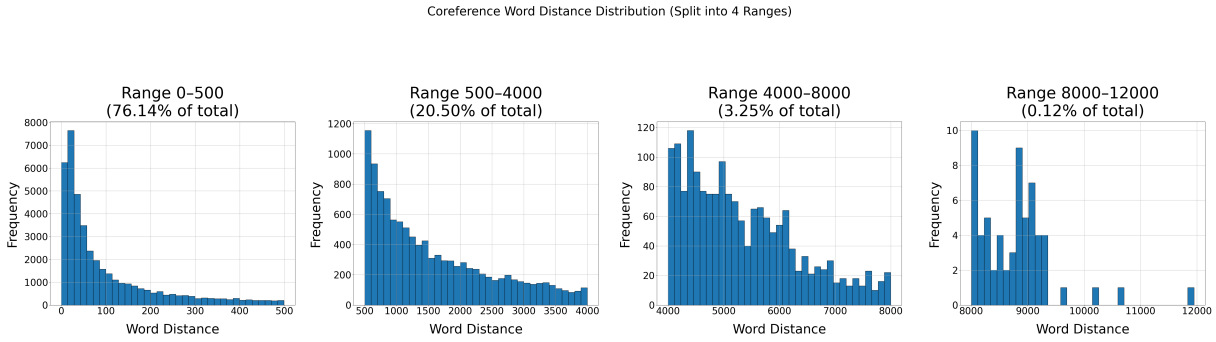

Figure 1: Distribution of word distances between coreferent mentions in biomedical texts, grouped into four ranges.

on semantics rather than lexical cues. Third, CRAFT's dense domain-specific terminology and pervasive abbreviation use cause general-domain systems to collapse: an out-of-the-box coreference resolver achieves only $F_1 = 0.14$ (B^3) on CRAFT, and even a domain-adapted rule-based system reaches only $F_1 = 0.42$ (Cohen et al., 2017a). These properties set CRAFT apart from otherwise strong alternatives. The closest runner-up for hard coreference evaluation is LitBank (Bamman et al., 2020), a challenging long-document complement to OntoNotes (Pradhan et al., 2013) that contains 210,532 tokens across 100 works of English fiction, with an average document length of 2,105 words, roughly four times longer than OntoNotes' 467-token average, and whose narratives form the backbone of recent LLM-oriented coreference benchmarks such as IdentifyMe (Manikantan et al., 2024). However, LitBank truncates each work to its first 2,000 tokens, covers only six ACE-style entity categories, and consists of general-domain prose. CRAFT, by contrast, pairs full-article context with the identical-surface ambiguity and specialized-knowledge demands described above, and couples its coreference layer to eight ontology-grounded semantic classes that yield a 76% increase in named-entity coverage through IDENTITY chains (Cohen et al., 2017a). These factors together establish CRAFT as a strictly harder benchmark than LitBank for evaluating LLM coreference capabilities.

3.2 Models

For our span-based baseline, we evaluate **SpanBERT-large-cased** model (Joshi et al., 2020), a span-optimized transformer pretrained on masked span prediction, on one experiment of coreferencing resolution. Input documents are segmented into 150-word chunks using `Stanza` (Qi et al., 2020), and 500 words in a another experiment setting. We normalized the document chunking sizes using stanza for all the experiments to assure the same chunk indices production for proper evaluation. After chunking the document, noun/pronoun mentions are extracted via `spaCy` (Vasiliev, 2020). Each mention is encoded using SpanBERT's final-layer embeddings and clustered via agglomerative clustering with cosine similarity ($\tau = 0.4$) to group together mentions that semantically refer to the same entity, approximating coreference. For the generative approach, we evaluate three open-weight LLaMA models on each of the 4 coreferencing experiments:

- **LLaMA 3.3 70B-Instruct** (Meta AI, 2024b): a high-capacity model (128k context) released in April 2024.
- **Llama-3.1-8B-Instruct** (Meta AI, 2024a): a compact model optimized for efficient text-only inference.
- **LLaMA 4 Scout 17B** (Meta AI, 2025): a 2025 multimodal model with a 10M-token context window and Mixture-of-Experts architecture.

4 Experiments

We evaluate four prompt-based strategies for coreference resolution using large language models over CRAFT-formatted biomedical texts. Each document D is split into paragraphs $p_1, \ldots, p_N$, each containing approximately 200 words. Using Stanza, we segment the text into sentences and iteratively append them to each paragraph chunk until the 200-word threshold is reached. If the last sentence causes the word count to exceed 200, it is deferred to the next paragraph. The goal is to

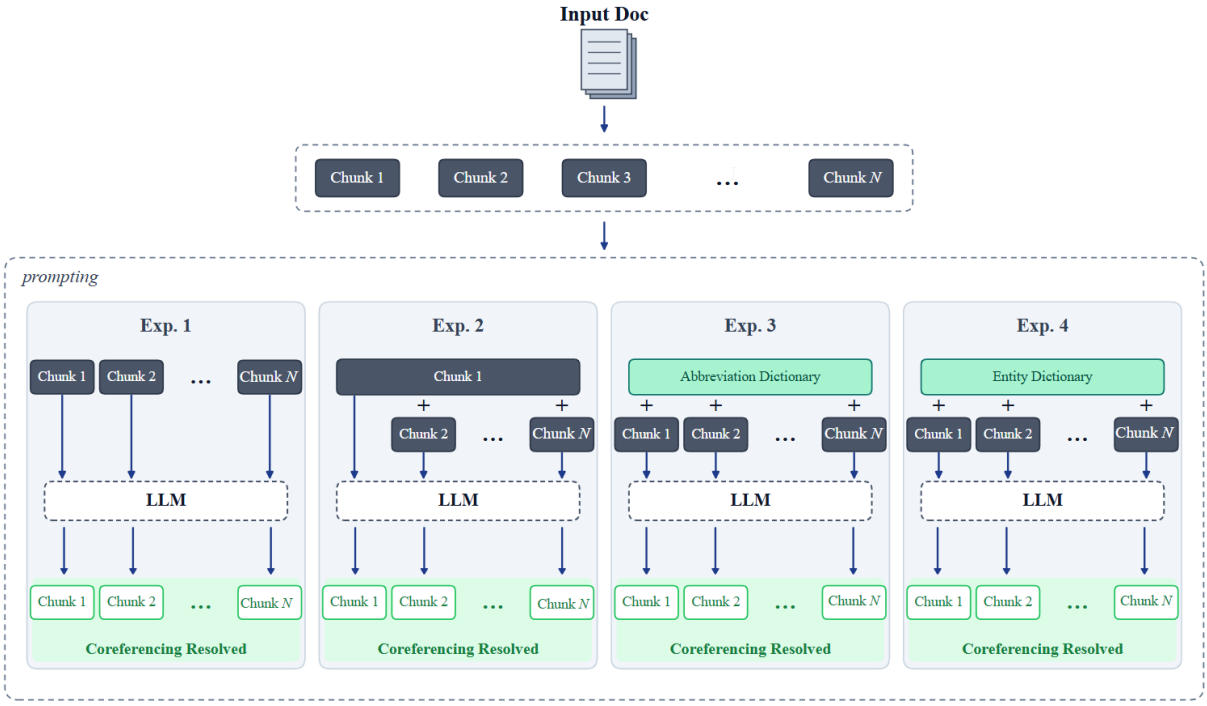

Figure 2: Overview of the coreference resolution pipeline under four prompting strategies. Each chunk is processed by an LLM independently (Exp. 1), with prior context (Exp. 2), or with auxiliary inputs such as abbreviation (Exp. 3) or entity dictionaries (Exp. 4).

output the set of detected mentions M_i, their corresponding resolutions A_i, and a "resolved" version of each paragraph R_i, where each p_i is independently rewritten. Formally:

- Let $\text{LLM}_\phi(\cdot)$ denote the output of the LLM with prompt ϕ.
- Let M_i be the set of coreferent mentions detected in paragraph p_i.
- Let A_i be the set of antecedent resolutions for M_i.
- Let R_i be the rewritten paragraph p_i with all mentions in M_i resolved using A_i.
- The reconstructed document is $\hat{D} = [R_1, \ldots, R_N]$.

The coreference resolution task involves resolving 4 categories: pronouns, definite and indefinite noun phrases, and abbreviations, as illustrated in Table 1.

To evaluate how different categories of auxiliary information affect coreference resolution, we design four prompting configurations: (1) a local-only setup with no external context, (2) a reference-based setup that incorporates the first paragraph as a fixed disambiguation source, (3) an abbreviation-aware setup using a dictionary of extracted abbreviation-definition pairs, and (4) an entity-aware setup using a list of biomedical entities extracted from the document. Algorithm 1 summarizes these 4 styles of the prompting experiments.

Coreference Type	Example Expressions
Pronouns	*it*, *these*, *those*, *its*, *their*
Definite noun phrases	*the gene*, *such results*
Indefinite noun phrases	*a protein*, *some genes*, *one of the enzymes*
Abbreviations	*IOP* $\rightarrow$ *intraocular pressure*

Table 1: Coreference categories and example expressions used in our experiments.

Figure 2 provides an overview of the experimental pipeline used to assess the effectiveness of LLMs across these different coreference categories.

4.1 Experiment 1: Local-Only Resolution (Baseline)

$$R_i = \text{LLM}_{\text{local}}(p_i)$$

In this initial experiment, we investigate the effectiveness of local coreference resolution by prompting LLMs to resolve coreference chains within short, isolated 200-word segments of a biomedical article. Each chunk is independently passed to the LLM. The goal is to assess how well the model performs coreference resolution without any cross-paragraph or global context.

This design reflects a naïve but computationally inexpensive strategy: it minimizes prompt complexity and token limits, while simulating how local context alone may or may not suffice for resolving biomedical coreference phenomena. We made a separate inference run for the 4 coreferencing categories.

This framework allows us to isolate and quantify the limitations of local-only resolution in biomedical texts. It also establishes a baseline against which we can measure subsequent experiments incorporating other resources, such as abbreviation expansion (Experiment 3).

4.2 Experiment 2: Coreference Resolution with Local and Reference Context

$$R_i = \text{LLM}_{\text{ref}}(p_1, p_i)$$

- **Prompt:** Provide p_1 and p_i, instructing the LLM to use the former to disambiguate references in the latter.
- **Purpose:** Test the incremental benefit of a reference paragraph for resolving inter-sentential and cross-paragraph coreferring mentions, without overloading the prompt size.

Building upon the limitations identified in Experiment 1, where coreference resolution was performed in isolation within 200-word chunks, we introduce an additional layer of local context to guide the LLM. In this experiment, each prompt to the model includes not only the target paragraph, but also the first 200-word paragraph in the paper, which carries most of the referential information introduced in the paper and can therefore act as an answer key for the unresolved references in the target paragraph.

Each prompt is structured with two segments: a reference block (Paragraph 1) and a focus block (Paragraph n), with explicit instructions for the LLM to resolve all ambiguous mentions in the focus block using context from the reference. This experiment assesses the impact of lightweight contextual bridging on coreference resolution quality. Compared to the purely local setting in Experiment 1, this approach tests whether even a single paragraph of surrounding context can significantly improve the coherence and referential clarity of LLM-generated outputs, without exceeding typical token limits or requiring full-document inputs.

Algorithm 1 Prompt-Based Coreference Resolution

Require: Document D, ExperimentType $\in$ {LOCAL, REF_CTX, ABBR, ENTITY}, Model LLM
Ensure: ResolvedDocument $\hat{D}$, MentionSets $\mathcal{M}$, ResolutionSets $\mathcal{A}$

Split D into paragraphs: $[p_1, p_2, \ldots, p_N]$
Initialize auxiliary content $C \leftarrow \emptyset$
if ExperimentType = ABBR **or** ExperimentType = ENTITY **then**
  $C \leftarrow$ EXTRACTCONTEXTINFO(D, type=ExperimentType)
end if
Initialize $\hat{D} \leftarrow$ [ ], $\mathcal{M} \leftarrow$ [ ], $\mathcal{A} \leftarrow$ [ ]
for $i = 1$ to N **do**
  $p \leftarrow p_i$
  if ExperimentType = REF_CTX **then**
    reference $\leftarrow p_1$ {Use Paragraph 1 as fixed reference}
  else
    reference $\leftarrow \emptyset$
  end if
  prompt $\leftarrow$ BUILDPROMPT(reference, p, C, ExperimentType)
  response $\leftarrow$ QUERYLLM(LLM, prompt)
  result $\leftarrow$ PARSEJSON(response)
  $\hat{D}$.append(result["Rewritten_Paragraph"])
  $\mathcal{M}$.append(result["Extracted_Expressions"])
  $\mathcal{A}$.append(result["Resolutions"])
end for
return $\hat{D}$, $\mathcal{M}$, $\mathcal{A}$

4.3 Experiment 3: Abbreviation-Aware Coreference Resolution Using LLM-Extracted Dictionaries

Let $A = \{(a_j, \alpha_j)\}$ be abbreviation-definition pairs extracted from the first 750 words using the GPT-4o.

$$R_i = \text{LLM}_{\text{abbr}}(A; p_i)$$

- **Prompt:** "Here is a list of abbreviations A.

the model is requested to extract all the coreferencing categories in separate runs, then, rewrite paragraph p_i by expanding ambiguous abbreviations and resolving references."

- **Purpose:** Leverage explicit abbreviation knowledge to aid disambiguation of biological mentions.

Biomedical texts frequently employ abbreviations for complex names, which can cause substantial ambiguity in coreference resolution. In this experiment, we assess whether providing LLMs with a structured abbreviation dictionary improves coreference resolution compared to unstructured context, such as the reference paragraphs used in Experiment 2. To build this dictionary, we parse the first 750 wordsof each document using Stanza and extract abbreviation-definition pairs (e.g., APP = "amyloid precursor protein") using the GPT-4o interface. These pairs are then validated against the CRAFT corpus. The resulting Abbreviation List is passed as auxiliary input during prompting.

4.4 Experiment 4: Entity-Aware Coreference Resolution Using LLM-Extracted Dictionaries

Let $E = \{e_k\}$ be key biomedical entities extracted from the first 750 words using GPT4o.

$$R_i = \text{LLM}_{\text{entity}}(E; p_i)$$

- **Prompt:** "Here is a list of detected biomedical entities E. Extract all the coreferencing mentions, then, rewrite paragraph pi by expanding ambiguous abbreviations and resolving references"

- **Purpose:** Provide broader semantic grounding than abbreviations alone, to evaluate whether entity awareness supports coherent coreference resolution.

In this experiment, we examine whether incorporating explicit biomedical entity information into the prompting process can enhance the performance of large language models on coreference resolution. Instead of only relying on implicit context or abbreviation mappings, we provide the LLM with a curated Entity List; a list of biomedical terms extracted using GPT-4o interface from the first 750 words of each document and validated against the CRAFT corpus to ensure correctness.

Table 2: LLaMA models performance metrics for LOCAL and REF_CTX tasks

Size	Task	P	R	F1
70B	LOCAL	0.800	0.458	0.583
	REF_CTX	0.805	0.390	0.525
17B	LOCAL	0.825	0.613	0.704
	REF_CTX	0.850	**0.573**	**0.685**
8B	LOCAL	**0.874**	**0.723**	**0.791**
	REF_CTX	**0.906**	0.539	0.676

This entity list serves as a form of semantic grounding. For each paragraph in the input article, the LLM is prompted with both the paragraph and the corresponding entity list. The model is then asked to resolve any ambiguous mentions by aligning them with the most probable entry in the entity list and rewriting the paragraph accordingly.

5 Results Analysis

The span-based baseline SpanBERT-large achieves an F1 of only 0.1322, highlighting the difficulty of biomedical coreference resolution for traditional models constrained by limited context windows, domain-specific terminology, and a mismatch between general-domain fine-tuning and specialized biomedical discourse.

To ensure accurate evaluation, we removed 9,335 gold-standard annotations from the selected CRAFT articles that lacked relation entries as the ones that don't connect to original source cannot be validates for coreferencing resolution. We then matched predicted resolutions against the remaining ∼83,608 annotated spans using case-insensitive partial character overlap ($\geq$2 characters). Predictions were extracted from structured JSON when available, or via a fallback regex parser. Precision, recall, and F1 were computed at the mention level, with unmatched predictions treated as false positives and missed gold spans as false negatives.

Our results reveal consistent patterns in how generative LLMs perform on biomedical coreference resolution:

Model Scale vs. Effectiveness.
Surprisingly, the 8B and 17B LLaMA models outperform the 70B variant across all experiments, indicating that scale alone does not drive coreference performance in domain-specific, fine-grained

Table 3: LLaMA models performance metrics for `abb_dictionary` and `entity_dictionary` tasks

Size	Task	P	R	F1
70B	ABBR	0.844	0.395	0.538
	ENTITY	0.826	0.379	0.519
17B	ABBR	**0.919**	0.400	0.558
	ENTITY	**0.891**	**0.633**	**0.740**
8B	ABBR	0.868	**0.653**	**0.745**
	ENTITY	0.882	0.551	0.678

Table 4: LLaMA models performance metrics for distance-aware local coreference resolution

Size	P	R	F1
70B	0.794	0.430	0.557
17B	0.841	0.554	0.668
8B	**0.892**	**0.601**	**0.718**

reasoning tasks. A likely explanation is that smaller models generalize more conservatively and commit fewer overconfident errors, whereas larger models, despite stronger generative capacity, appear more susceptible to prompt misalignment and semantic overreach.

To further probe local-only performance, we ran an additional variant of Experiment 1 in which paragraphs were selected not by a sequential 200-word window but by the most frequent reference distance observed in CRAFT (500 words), testing whether LLMs benefit from input chunks that align with natural coreference distances. Results appear in Table 4.

Impact of Coreference Distance.
The distance-aware local variant yields only marginal precision gains and a noticeable drop in recall and F1 for LLaMA 17B and 8B, indicating that proximity alone is insufficient for robust coreference; many biomedical entities require contextual cues beyond sentence-local information. LLaMA 8B still achieves the highest F1, confirming its strong span-level sensitivity, yet the degradation with larger context corroborates what we observe in Experiment 2 (`REF_CTX`).

Reference Context Has Mixed Effects.
In Experiment 2, which incorporates a fixed reference paragraph to aid disambiguation, recall often drops compared to the purely local setup (Experiment 1), particularly for LLaMA 8B and 70B. This suggests that, without fine-tuning or explicit multi-span integration mechanisms, LLMs may not reliably incorporate reference paragraphs into their reasoning. Prior work has observed that transformer-based models tend to prioritize recent tokens or local context unless explicitly guided otherwise, which could explain the degraded performance here.

Structured Dictionaries Improve Recall.
The abbreviation-aware settings (Experiments 3) yield measurable recall and F1 gains, most visibly for the 8B model, which reaches recall of 0.653 and 0.745 in `ABBR`. When supplied with structured input; abbreviation definitions, even smaller models like the 8B one can more reliably identify correct antecedents. These findings align with prior evidence that structured, grounded prompting improves information extraction, particularly when domain-specific context is required.

Overall, these results highlight both the promise and the current limitations of generative LLMs for biomedical coreference: while auxiliary signals such as abbreviation and entity dictionaries meaningfully boost performance, LLMs still struggle to integrate multi-paragraph context and resolve less explicit coreferences, underscoring the need for domain-specific adaptation or hybrid approaches that combine generative models with symbolic or retrieval-based components.

Coreference Type Sensitivity and Model Behavior.

To better understand how LLMs handle different forms of coreference, we evaluated each model across four categories; pronouns, indefinite noun phrases, abbreviations, and definite noun phrases, under the `LOCAL`, `REF_CTX`, `ABBR`, and `ENTITY` setups. Because CRAFT does not label coreference types, we built a post-processing pipeline that classifies predicted and gold mentions via lexical heuristics: pronouns, indefinite expressions, and abbreviations are identified by exact matches against dictionaries we manually compiled (available on our GitHub), and the remaining mentions are treated as definite noun phrases. Per-type precision, recall, and F1 are then computed from the model's original prediction labels.

As shown in Figure 3, pronoun coreference consistently yields the highest F1 across all LLaMA

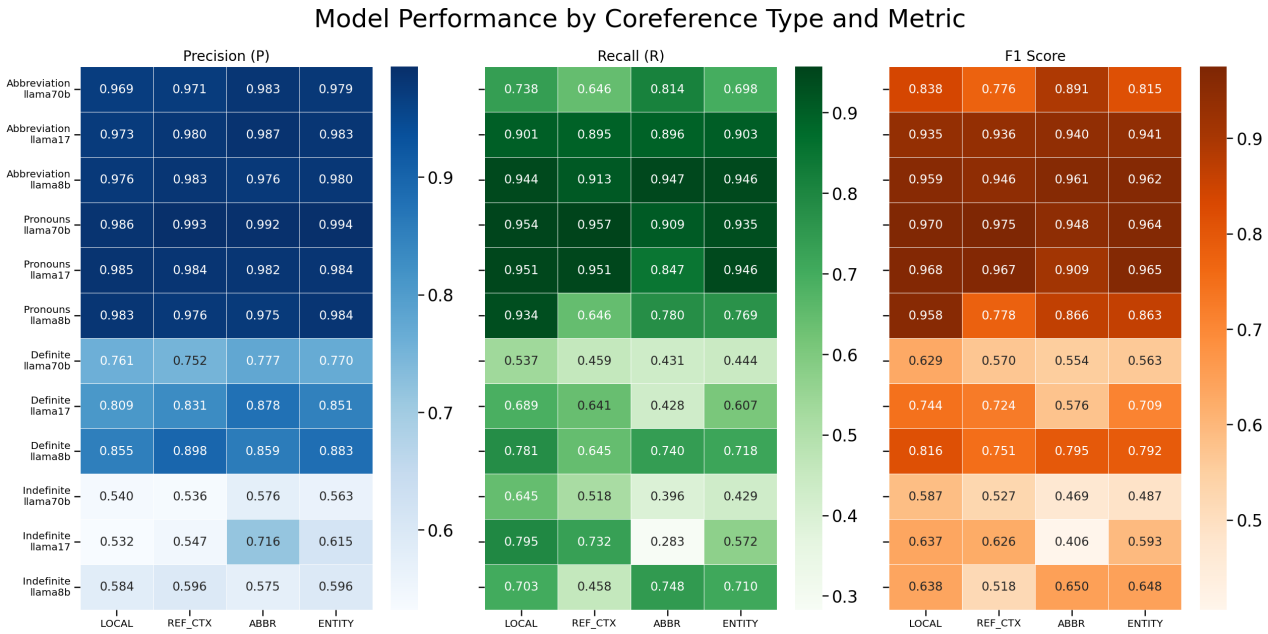

Figure 3: Heatmap of precision, recall, and F1 scores for LLaMA models (70B, 17B, 8B) across four experimental setups (LOCAL, REF CTX, ABBR, ENTITY) and coreference categories (pronouns, indefinite NPs, abbreviations, definite NPs).

models, peaking at 0.975 for LLaMA 70B under REF_CTX, likely because pronouns are heavily represented in pretraining corpora and rely on short-range syntactic cues; complementary evidence from Figure 4 in Appendix A shows pronouns are also resolved in high absolute counts, especially by LLaMA 17B under minimal context.

Abbreviation coreference is similarly strong, particularly under ABBR and ENTITY: injecting abbreviation dictionaries produces noticeable gains in both F1 (e.g., 0.961 for LLaMA 8B in ABBR) and resolved-mention counts, confirming that domain-specific cues substantially aid biomedical abbreviation understanding.

Definite noun phrase resolution reaches moderate F1, with LLaMA 8B highest in LOCAL (F1: 0.816), but both metrics and counts decline slightly under ABBR and ENTITY, likely due to contextual noise from input augmentation, suggesting that local syntactic proximity matters more for definite noun phrases than external cues. Indefinite noun phrases pose the greatest challenge: F1 is lowest across all models and experiments, with LLaMA 17B marginally stronger in REF_CTX and LOCAL. Notably, although LLaMA 70B leads in pronoun F1, it consistently underperforms elsewhere in both metrics and counts, indicating that larger models may be more prone to distraction or overfitting in domain-specific settings.

6 Conclusion

Our study presented a systematic evaluation of generative large language models (LLMs) for coreference resolution in the biomedical domain. We benchmarked three LLaMA models across four prompt-based settings and compared them to a span-based baseline, using the richly annotated CRAFT corpus for evaluation.

Our findings show that LLMs can resolve coreference with reasonable precision, but often suffer from low recall without explicit cues. Notably, smaller models like LLaMA 8B frequently outperformed larger ones, especially when provided with structured input such as abbreviation. In contrast, injecting additional free-text context offered limited benefit and sometimes degraded performance.

Overall, these results highlight the nuanced relationship between model size, coreference category, and the design of contextual input. They emphasize that targeted domain-specific augmentation, such as structured dictionaries, can have a greater impact on performance than model scale alone. Notably, smaller models can match or even exceed the performance of larger ones when paired with carefully designed prompts. Future directions should explore fine-tuning strategies, integration of external biomedical knowledge, and hybrid generative extractive systems to further enhance recall and robustness.

References

David Bamman, Olivia Lewke, and Anya Mansoor. 2020. An Annotated Dataset of Coreference in English Literature. https://arxiv.org/abs/1912.01140. Proceedings of the 12th Language Resources and Evaluation Conference (LREC), pages 44–54, Marseille, France.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Kevin Clark and Christopher D Manning. 2016. Deep reinforcement learning for mention-ranking coreference models. *arXiv preprint arXiv:1609.08667*.

K. Bretonnel Cohen, Arrick Lanfranchi, Miji Joo-young Choi, Michael Bada, William A. Baumgartner, Natalya Panteleyeva, Karin Verspoor, Martha Palmer, and Lawrence E. Hunter. 2017a. Coreference annotation and resolution in the Colorado Richly Annotated Full Text (CRAFT) corpus of biomedical journal articles. https://doi.org/10.1186/s12859-017-1775-9. BMC Bioinformatics 18(1):372.

K Bretonnel Cohen, Arrick Lanfranchi, Miji Joo-young Choi, Michael Bada, William A Baumgartner Jr, Natalya Panteleyeva, Karin Verspoor, Martha Palmer, and Lawrence E Hunter. 2017b. Coreference annotation and resolution in the colorado richly annotated full text (craft) corpus of biomedical journal articles. *BMC bioinformatics*, 18(1):372.

Vladimir Dobrovolskii. 2021. Word-level coreference resolution. *arXiv preprint arXiv:2109.04127*.

Greg Durrett and Dan Klein. 2013. Easy victories and uphill battles in coreference resolution. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1971–1982.

Yujian Gan, Juntao Yu, and Massimo Poesio. 2024. Assessing the capabilities of large language models in coreference: An evaluation. In *Joint 30th International Conference on Computational Linguistics and 14th International Conference on Language Resources and Evaluation, LREC-COLING 2024*, pages 1645–1665. European Language Resources Association (ELRA).

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the association for computational linguistics*, 8:64–77.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. *arXiv preprint arXiv:1707.07045*.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. *arXiv preprint arXiv:1804.05392*.

Yufei Li, Xiangyu Zhou, Jie Ma, Xiaoyong Ma, Pengzhen Cheng, Tieliang Gong, and Chen Li. 2022. Distinguished representation of identical mentions in bio-entity coreference resolution. *BMC Medical Informatics and Decision Making*, 22(1):116.

Ruicheng Liu, Rui Mao, Anh Tuan Luu, and Erik Cambria. 2023. A brief survey on recent advances in coreference resolution. *Artificial Intelligence Review*, 56(12):14439–14481.

Yanming Liu, Xinyue Peng, Jiannan Cao, Shi Bo, Yanxin Shen, Tianyu Du, Sheng Cheng, Xun Wang, Jianwei Yin, and Xuhong Zhang. 2024. Bridging context gaps: Leveraging coreference resolution for long contextual understanding. *arXiv preprint arXiv:2410.01671*.

Pengcheng Lu and Massimo Poesio. 2021. Coreference resolution for the biomedical domain: A survey. *arXiv preprint arXiv:2109.12424*.

Kawshik Manikantan, Makarand Tapaswi, Vineet Gandhi, and Shubham Toshniwal. 2024. IdentifyMe: A Challenging Long-Context Mention Resolution Benchmark for LLMs. https://arxiv.org/abs/2411.07466. ArXiv preprint arXiv:2411.07466.

Meta AI. 2024a. Meta Llama 3.1 8B Instruct. https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct. Released July 23, 2024, under the Llama 3.1 Community License.

Meta AI. 2024b. Meta Llama 3.3 70B Instruct. https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct. Released July 23, 2024, under the Llama 3.1 Community License.

Meta AI. 2025. Meta Llama 4 Scout 17B-16E Instruct. https://huggingface.co/meta-llama/Llama-4-Scout-17B-16E-Instruct. Released April 5, 2025, under the Llama 4 Community License.

Vincent Ng and Claire Cardie. 2002. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 104–111.

Brendan O'Connor and Michael Heilman. 2013. Arkref: A rule-based coreference resolution system. *arXiv preprint arXiv:1310.1975*.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards Robust Linguistic Analysis using OntoNotes. https://aclanthology.org/W13-3516/. Proceedings of the Seventeenth Conference on Computational Natural Language Learning (CoNLL), pages 143–152.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, and 1 others. 2018. Improving language understanding by generative pre-training.

Nourah M. Salem and 1 others. 2025. BioCoref: Benchmarking Biomedical Coreference Resolution with LLMs. https://arxiv.org/abs/2510.25087. ArXiv preprint arXiv:2510.25087.

Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. 2001. A machine learning approach to coreference resolution of noun phrases. *Computational linguistics*, 27(4):521–544.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Yuli Vasiliev. 2020. *Natural language processing with Python and spaCy: A practical introduction*. No Starch Press.

Sam Joshua Wiseman, Alexander Sasha Rush, and Jason Weston. 2015. Learning anaphoricity and antecedent ranking features for coreference resolution. Association for Computational Linguistics.

Wei Wu, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li. 2020. Corefqa: Coreference resolution as query-based span prediction. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 6953–6963.

Tianbao Xie, Chen Henry Wu, Peng Shi, Ruiqi Zhong, Torsten Scholak, Michihiro Yasunaga, Chien-Sheng Wu, Ming Zhong, Pengcheng Yin, Sida I Wang, and 1 others. 2022. Unifiedskg: Unifying and multi-tasking structured knowledge grounding with text-to-text language models. *arXiv preprint arXiv:2201.05966*.

7 Appendix A

7.1 Additional Experiment Results

Figure 4 reports the absolute number of coreference mentions extracted by each LLaMA model (70B, 17B, 8B) under the four experimental setups (LOCAL, REF_CTX, ABBR, ENTITY), broken out by coreference category: pronouns, indefinite noun phrases, abbreviations, and definite noun phrases. The panel-wise y-axis ranges differ substantially across categories, reflecting the relative frequency with which each coreference type occurs in the CRAFT articles; definite noun phrases and abbreviations are by far the most numerous classes, while pronouns and indefinite noun phrases appear at an order of magnitude smaller scale.

Two patterns stand out. First, pronoun and definite-noun-phrase extraction are dominated by the larger models under minimal context, with LLaMA 70B and LLaMA 17B producing the highest counts in the LOCAL and REF_CTX conditions and then tapering off under the dictionary-augmented prompts, consistent with their high pronoun F1 scores in Figure 3 and the short-range syntactic nature of these mentions. Second, abbreviation extraction shows the opposite trend: LLaMA 8B produces the largest number of resolved abbreviations, and its count grows sharply under the ABBR and ENTITY prompts, mirroring the 0.961 F1 it achieves on abbreviations in the ABBR setup. Indefinite noun phrases follow a middle pattern, with LLaMA 17B leading in the lower-context conditions (LOCAL, REF_CTX) and all three models converging once structured dictionaries are supplied. Together these counts reinforce the quantitative findings from Tables 2 and 3: structured domain cues disproportionately benefit the smaller 8B model, particularly on the abbreviation-heavy categories where biomedical grounding matters most.

7.2 Computational Resouces

All open-weight LLM experiments were conducted on a high-performance Google Cloud VM instance of type `a2-highgpu-8g`, equipped with 96 vCPUs, 680 GB of RAM, and 8 NVIDIA A100 GPUs (40 GB each). The environment was provisioned with the c0-deeplearning-common-cu118-v20241118-debian-11-py310 boot disk image, ensuring compatibility with CUDA 11.8 and PyTorch 2.x frameworks.

8 Appendix B

8.1 Prompt Template

To guide the language model's behavior consistently across experiments, we employ a structured system prompt for each coreference type, and this prompt instructs the model to identify and resolve only a targeted subset of coreference expressions. In this example, the focus is on definite noun phrase coreferences within a paragraph, while explicitly excluding pronouns, indefinite expressions, and abbreviations.

The same system-prompt skeleton is reused across all four experimental setups, with only the

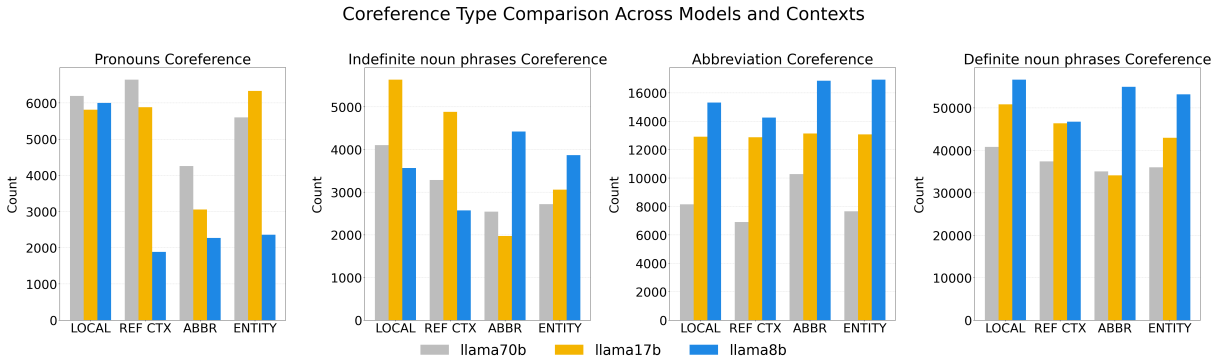

Figure 4: Extracted coreference type counts by model and context.

auxiliary-context block changing: LOCAL uses the skeleton verbatim, REF_CTX prepends the first paragraph of the article as a fixed reference block, ABBR injects the extracted abbreviation–definition pairs $\mathcal{A}$, and ENTITY injects the extracted biomedical entity list $\mathcal{E}$. The abbreviation and entity dictionaries used in Experiments 3 and 4 are themselves produced by a separate GPT-4o extraction prompt applied to the first 750 words of each document and validated against the CRAFT gold annotations. The four coreference categories (pronouns, definite noun phrases, indefinite noun phrases, and abbreviations) are resolved in independent inference runs; only the "ONLY X" clause and the accompanying skip list change between runs, while the JSON output schema remains identical. All LLaMA inferences are run with temperature 0 and top-$p = 1$ to maximize determinism, with `max_new_tokens` set to accommodate the full rewritten paragraph plus its JSON envelope. Model responses are first parsed as strict JSON; when parsing fails, a fallback regex extractor recovers the `Extracted_Expressions`, `Resolutions`, and `Rewritten_Paragraph` fields and the example is retained only if all three are present. Malformed outputs that cannot be recovered are discarded and counted as empty predictions for evaluation.

System Prompt

You are a scientific language model with expert-level understanding of coreference resolution.
Your task is to extract and resolve ONLY **definite noun phrase coreferences** (e.g., "the gene", "these proteins", "such results") within the paragraph.
Skip the following:

- Pronouns (e.g., "it", "they")
- Indefinite noun phrases (e.g., "a result", "some proteins")
- Abbreviations (e.g., "IOP")

Follow these steps:

1. Extract coreference expressions that appear *verbatim* in the paragraph. **Do NOT invent or rephrase them**.
2. For each expression, resolve it to its correct antecedent using context from the same paragraph.
3. Rewrite the paragraph by substituting each extracted expression with its resolved referent.

DO NOT paraphrase, summarize, add, remove, or reorder any content. **Preserve the original wording and sentence structure, except for the substitutions.**

Expected JSON Output Schema:

```
{
        "Extracted_Expressions": [
        "[expression_1]",
        "[expression_2]"
        ],
        "Resolutions": {
                "[expression_1]": "[detailed
    explanation describing the antecedent]",
                "[expression_2]": "[detailed
    explanation describing the antecedent]"
        },
        "Rewritten_Paragraph": "[the rewritten
    paragraph, identical except for
    substitutions]"
}
```

Example:

> *Input:* "These results were unexpected. They indicate a new trend."
> *Rewritten:* "The XYZ experiment results were unexpected. The results indicate a new trend."

Example Output:

```
{
        "Coreference_Resolution": {
                "Extracted_Expressions": [
                "IOP",
                "IOPs",
                "They"
                ],
                "Resolutions": {
                        "IOP": "intraocular
    pressure",
                        "IOPs": "intraocular
    pressures",
                        "They": "Genetically
    distinct mouse strains"
                },
                "Rewritten_Paragraph": "
    Intraocular pressure in genetically distinct
     mice: an update and strain survey..."
           }
}
```