

000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 BIOCOREF: BENCHMARKING BIOMEDICAL COREFER- ENCE RESOLUTION WITH LLMs

Anonymous authors

Paper under double-blind review

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 as the literature often contain dense, technical language, frequent use of abbreviations, and complex referential expressions that rely on domain-specific background knowledge Lu & Poesio (2021). 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. (2017) 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 & 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 biomedical coreference remains a demanding

054 task, recent advances suggest that with well-designed prompts and minimal scaffolding, LLMs may
 055 be more capable than previously assumed.
 056

057 In this work, we evaluate coreference resolution in biomedical PubMed articles using two con-
 058 trasting approaches: a span-based model (SpanBERT-Large) Joshi et al. (2020) trained on general-
 059 domain data, and several generative LLMs (LLaMA series) Touvron et al. (2023) prompted to re-
 060 solve coreference without fine-tuning. Our experiments use the CRAFT corpus, a richly annotated
 061 biomedical dataset.
 062

063 The contributions and objectives of this paper are summarized as follows:
 064

- 065 • **Benchmarking LLMs** We compare multiple LLaMA models under different prompting
 066 strategies: local-only, contextual, abbreviation-aware, and entity-aware against a span-
 067 based BERT baseline, reporting performance on the CRAFT corpus.
 068
- 069 • **Domain Analysis:** We identify coreference challenges unique to biomedical text, such as
 070 identical mention strings and abbreviation ambiguity, and analyze how each model type
 071 handles them through qualitative examples and error patterns.
 072

073 2 RELATED WORK

074 Coreference resolution in biomedical text is a particularly challenging task due to complex domain-
 075 specific terminology, high referential ambiguity, and long-range dependencies. Traditional span-
 076 based models such as the end-to-end neural coreference models such as SpanBERT Joshi et al.
 077 (2020) have demonstrated strong performance in general domains. However, their reliance on lim-
 078 ited context windows and the need for supervised fine-tuning limits their applicability in biomedical
 079 settings, where coreference often requires broader semantic grounding.
 080

081 Traditional approaches also include rule-based and statistical systems O’Connor & Heilman (2013);
 082 Ng & Cardie (2002); Soon et al. (2001), followed by neural architectures such as the mention-
 083 ranking model Clark & Manning (2016) and end-to-end span-ranking networks Lee et al. (2017);
 084 Durrett & Klein (2013); Wiseman et al. (2015); Lee et al. (2018). SpanBERT Joshi et al. (2020) fur-
 085 ther improved coreference resolution by introducing span-centric pretraining objectives, achieving
 086 state-of-the-art results on the OntoNotes Dobrovolskii (2021) benchmark. Despite these advances,
 087 such models rely heavily on supervised training and domain-specific tuning, limiting their general-
 088 izability to out-of-domain settings like biomedical text.
 089

090 Large language models (LLMs) like OpenAI GPT Radford et al. (2018) and LLaMA have demon-
 091 strated strong zero-shot capabilities in various NLP tasks Brown et al. (2020); Touvron et al. (2023),
 092 including aspects of coreference. Recent studies evaluated LLMs’ abilities on pronoun resolution
 093 and Winograd schemas Liu et al. (2024); Wu et al. (2020) for other downstream tasks such as
 094 question-answering and query-based span prediction problems. However, few studies have directly
 095 assessed LLMs on span-based or noun phrase coreference, particularly in long or technical docu-
 096 ments. Most relevant to our work are recent prompting frameworks that use generative models for
 097 structured information extraction Xie et al. (2022), though coreference-specific prompting remains
 098 underexplored, especially in specialized domains like biomedical literature.
 099

100 3 TASK OVERVIEW

101 We evaluate four prompt-based strategies for coreference resolution using large language models
 102 over CRAFT-formatted biomedical texts. Each document D is split into paragraphs p_1, \dots, p_N ,
 103 each containing approximately 200 words. Using Stanza parser, we segment the text into sentences
 104 and iteratively append them to each paragraph chunk until the 200-word threshold is reached. If the
 105 last sentence causes the word count to exceed 200, it is deferred to the next paragraph. The goal
 106 is to output the set of detected mentions M_i , their corresponding resolutions A_i , and a “resolved”
 107 version of each paragraph R_i , where each p_i is independently rewritten. Formally:

- 108 • Let $LLM_\phi(\cdot)$ denote the output of the LLM with prompt ϕ .
 109
- 110 • Let M_i be the set of coreferent mentions detected in paragraph p_i .
 111

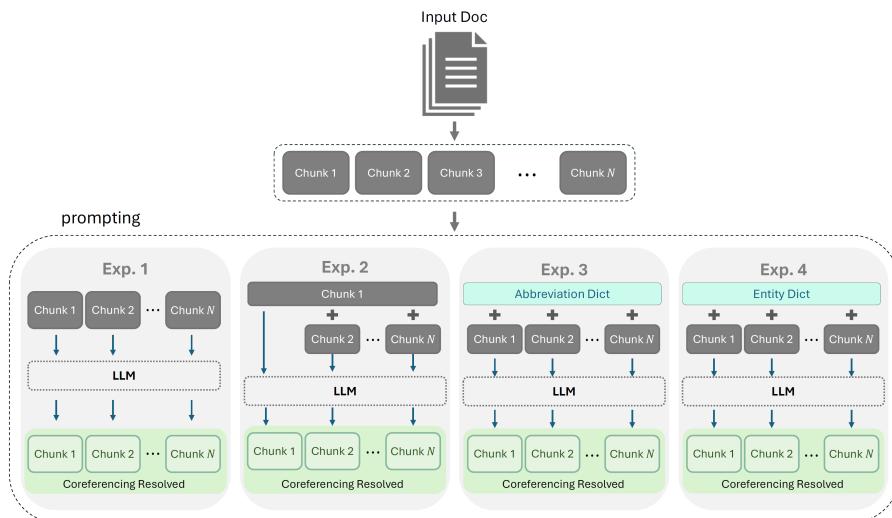
- 108 • Let A_i be the set of antecedent resolutions for M_i .
 109 • Let R_i be the rewritten paragraph p_i with all mentions in M_i resolved using A_i .
 110 • The reconstructed document is $\hat{D} = [R_1, \dots, R_N]$.

113 The coreference resolution task involves resolving 4 categories: pronouns, definite and indefinite
 114 noun phrases, and abbreviations, as illustrated in Table 1. Each is prompted separately for extraction
 115 and resolving.

Coreference Type	Example Expressions
Pronouns	<i>it, they, this, these, those, its, their</i>
Definite noun phrases	<i>the gene, these proteins, such results</i>
Indefinite noun phrases	<i>a protein, some genes, one of the enzymes</i>
Abbreviation coreference	<i>IOP → intraocular pressure</i>

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

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



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

153 3.1 EXPERIMENT 1: LOCAL-ONLY RESOLUTION (BASELINE)

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

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

161 This design reflects a naïve but computationally inexpensive strategy: it minimizes prompt complexity
 162 and token limits, while simulating how local context alone may or may not suffice for resolving

162 **Algorithm 1** Prompt-Based Coreference Resolution

163

164 **Require:** Document D , ExperimentType $\in \{\text{LOCAL, REF_CTX, ABBR, ENTITY}\}$, Model LLM

165 **Ensure:** ResolvedDocument \hat{D} , MentionSets \mathcal{M} , ResolutionSets \mathcal{A}

166 1: Split D into paragraphs: $[p_1, p_2, \dots, p_N]$

167 2: Initialize auxiliary content $C \leftarrow \emptyset$

168 3: **if** ExperimentType = ABBR **or** ExperimentType = ENTITY **then**

169 4: $C \leftarrow \text{EXTRACTCONTEXTINFO}(D, \text{type}=\text{ExperimentType})$

170 5: **end if**

171 6: Initialize $\hat{D} \leftarrow []$, $\mathcal{M} \leftarrow []$, $\mathcal{A} \leftarrow []$

172 7: **for** $i = 1$ to N **do**

173 8: $p \leftarrow p_i$

174 9: **if** ExperimentType = REF_CTX **then**

175 10: reference $\leftarrow p_1$ {Use Paragraph 1 as fixed reference}

176 11: **else**

177 12: reference $\leftarrow \emptyset$

178 13: **end if**

179 14: prompt $\leftarrow \text{BUILDPROMPT}(\text{reference}, p, C, \text{ExperimentType})$

180 15: response $\leftarrow \text{QUERYLLM}(\text{LLM}, \text{prompt})$

181 16: result $\leftarrow \text{PARSEJSON}(\text{response})$

182 17: $\hat{D}.\text{append}(\text{result}[\text{"Rewritten.Paragraph"}])$

183 18: $\mathcal{M}.\text{append}(\text{result}[\text{"Extracted_Expressions"}])$

184 19: $\mathcal{A}.\text{append}(\text{result}[\text{"Resolutions"}])$

185 20: **end for**

186 21: **return** $\hat{D}, \mathcal{M}, \mathcal{A}$

187 biomedical coreference phenomena. We made a separate inference run for the 4 coreferencing categories.

188 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).

193 3.2 EXPERIMENT 2: COREFERENCE RESOLUTION WITH LOCAL AND REFERENCE CONTEXT

194

195

196
$$R_i = \text{LLM}_{\text{ref}}(p_1, p_i)$$

197

- 198
- 199
 - **Prompt:** Provide p_1 and p_i , instructing the LLM to use the former to disambiguate references in the latter.
- 200
- 201
 - **Purpose:** Test the incremental benefit of a reference paragraph for resolving intersentential and cross-paragraph coreferring mentions, without overloading the prompt size.
- 202

203 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.

204 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.

216 3.3 EXPERIMENT 3: ABBREVIATION-AWARE COREFERENCE RESOLUTION USING
217 LLM-EXTRACTED DICTIONARIES
218219 Let $A = \{(a_j, \alpha_j)\}$ be abbreviation-definition pairs extracted from the first 750 words using the
220 GPT-4o.

221
222
$$R_i = \text{LLM}_{\text{abbr}}(A; p_i)$$

223

- 224
-
- 225
- **Prompt:** “Here is a list of abbreviations A . the model is requested to extract all the coref-
226 erencing categories in separate runs, then, rewrite paragraph p_i by expanding ambiguous
227 abbreviations and resolving references.”
 - **Purpose:** Leverage explicit abbreviation knowledge to aid disambiguation of biological
228 mentions.
-
- 229

230 Biomedical texts frequently employ abbreviations for complex names, which can cause substantial
231 ambiguity in coreference resolution. In this experiment, we assess whether providing LLMs with
232 a structured abbreviation dictionary improves coreference resolution compared to unstructured con-
233 text, such as the reference paragraphs used in Experiment 2. To build this dictionary, we parse the
234 first 750 words of each document using Stanza and extract abbreviation-definition pairs (e.g., APP =
235 “amyloid precursor protein”) using the GPT-4o interface. These pairs are then validated against the
236 CRAFT corpus to ensure correctness. The resulting Abbreviation List is passed as auxiliary input
237 during prompting.
238239 3.4 EXPERIMENT 4: ENTITY-AWARE COREFERENCE RESOLUTION USING
240 LLM-EXTRACTED DICTIONARIES
241242 Let $E = \{e_k\}$ be key biomedical entities extracted from the first 750 words using GPT4o.
243

244
245
$$R_i = \text{LLM}_{\text{entity}}(E; p_i)$$

246

- 247
-
- 248
- **Prompt:** “Here is a list of detected biomedical entities E . Extract all the coreferencing
249 mentions, then, rewrite paragraph p_i by expanding ambiguous abbreviations and resolving
250 references”
 - **Purpose:** Provide broader semantic grounding than abbreviations alone, to evaluate
251 whether entity awareness supports coherent coreference resolution.
252

253 In this experiment, we examine whether incorporating explicit biomedical entity information into the
254 prompting process can enhance the performance of large language models on coreference resolution.
255 Instead of only relying on implicit context or abbreviation mappings, we provide the LLM with a
256 curated Entity List; a list of biomedical terms extracted using GPT-4o interface from the first 750
257 words of each document and validated against the CRAFT corpus to ensure correctness.
258259 This entity list serves as a form of semantic grounding. For each paragraph in the input article, the
260 LLM is prompted with both the paragraph and the corresponding entity list. The model is then asked
261 to resolve any ambiguous mentions by aligning them with the most probable entry in the entity list
262 and rewriting the paragraph accordingly.
263264 4 DATASET
265266 The Colorado Richly Annotated Full-Text (CRAFT) corpus is an annotated biomedical dataset con-
267 sisting of 67 full-text, open-access journal articles from PubMed Central. The dataset includes
268 extensive manual annotations for biomedical concepts, syntactic structure, and coreference iden-
269 tity chains. In Version 2.0, CRAFT introduced comprehensive coreference annotations, comprising
nearly 30,000 coreference relations that span both identity and appositive links across base noun

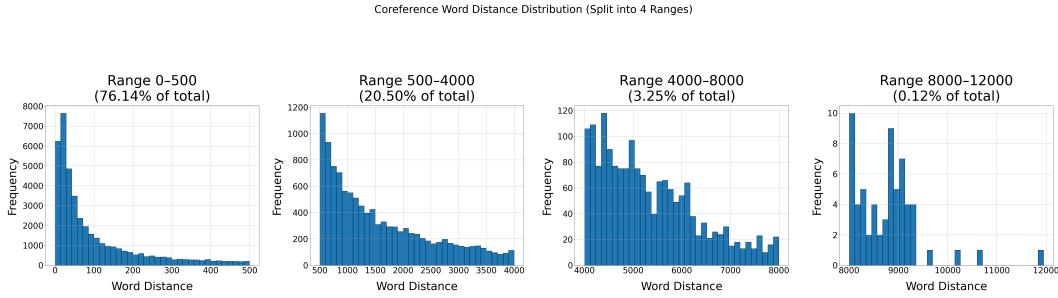


Figure 2: Distribution of word distances between coreferent mentions in biomedical texts, grouped into four ranges.

phrases¹. We tested the language models on 50 articles selected at random, which are available on the paper’s GitHub².

Figure 2 shows the distribution of word-level distances between entities and their coreferential mentions in the CRAFT dataset, segmented into four ranges. While most coreference links (76.14%) fall within 0–500 words, a substantial portion (over 23%) spans distances up to 12,000 words. These long-range dependencies present a significant challenge for coreference models, which must retain and retrieve relevant context across spans exceeding the effective window of standard neural architectures.

In addition to these distance-based challenges, the dataset itself contains numerous subtle forms of annotations variability that could be challenging for LLMs. For example, mentions of age: “3 months of age”; statistical and symbolic variables may be abbreviated as single letters like “n”; temporal references: “1997”; dosage information often mixes numbers and units, as in “9 mg/kg xylazine”; and biological sex descriptors appear inconsistently as “Male,” “M,” “Female,” or “F.” These heterogeneous expression patterns increase the complexity of accurate entity recognition and coreference resolution, making the dataset a very helpful benchmark for the current LLMs.

5 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), as the model can handle larger context inputs. 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 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 (2024a): a high-capacity model (128k context) released in April 2024.
- **Llama-3.1-8B-Instruct** Meta AI (2024b): 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.

¹<https://github.com/UCDenver-ccp/CRAFT>

²<https://github.com/XXXX/BioCoref>

324 6 RESULTS ANALYSIS

326 To ensure accurate evaluation, we removed 9335 gold-standard annotations from the CRAFT se-
 327 lected article that were not connected via relation entries. We then matched predicted resolutions
 328 against the remaining 83,608 annotated spans using partial character overlap (≥ 2 characters), en-
 329 suring case-insensitive alignment. Predictions were extracted from structured JSON when available,
 330 or via a fallback regex parser. Precision, recall, and F1 scores were computed at the mention level,
 331 with unmatched predictions treated as false positives and missed gold spans as false negatives.

332 First, we report the performance of the span-based baseline **SpanBERT-large**, which achieves an
 333 F1 score of only 0.1322. This highlights the difficulty of biomedical coreference resolution for
 334 traditional models, due to limited context windows, domain-specific terminology, and the mismatch
 335 between general-domain fine-tuning and specialized biomedical discourse.

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

342 Table 2: Performance metrics for LOCAL and REF_CTX tasks

344 Model	345 LOCAL			346 REF_CTX		
	347 Precision	348 Recall	349 F1 Score	350 Precision	351 Recall	352 F1 Score
353 LLaMA 70B	0.800	0.458	0.583	0.805	0.390	0.525
354 LLaMA 17B	0.825	0.613	0.704	0.850	0.573	0.685
355 LLaMA 8B	0.874	0.723	0.791	0.906	0.539	0.676

356 Table 3: Performance metrics for abb_dictionary and entity_dictionary tasks

357 Model	358 ABBR			359 ENTITY		
	360 Precision	361 Recall	362 F1 Score	363 Precision	364 Recall	365 F1 Score
366 LLaMA 70B	0.844	0.395	0.538	0.826	0.379	0.519
367 LLaMA 17B	0.919	0.400	0.558	0.891	0.633	0.740
368 LLaMA 8B	0.868	0.653	0.745	0.882	0.551	0.678

370 Our results reveal consistent patterns in how generative LLMs perform on coreference resolution in
 371 the biomedical domain:

372 **Model Scale vs. Effectiveness.** LLaMA 8B and the 17B models outperformed the 70B variant
 373 across all experiments. This suggests that model scale alone is not indicative of better coreference
 374 performance, especially in domain-specific tasks. One likely explanation is that smaller models
 375 generalize more conservatively and make fewer overconfident errors, whereas larger models de-
 376 spite stronger generative capacity may be more susceptible to prompt misalignment and semantic
 377 overreach.

378 To further probe local-only performance, we ran an additional variant of Experiment 1 where para-
 379 graphs were selected not sequentially by size of 200-word window, but based on the most frequent
 380 reference distance observed in the CRAFT corpus, which is 500 words. The goal of this experi-
 381 ment is to test whether LLMs perform better when the input chunk maximally aligns with natural
 382 coreference distances, rather than strict 200-word segmentation. Results are shown in Table 4.

383 **Impact of Coreference Distance.** A noticeable drop in recall and F1 for LLaMA 17B and 8B is
 384 shown in this distance-aware local experiment variant. This suggests that proximity alone is insuffi-
 385 cient for robust coreference; many biomedical entities require contextual cues beyond sentence-local

378
379
380 Table 4: Performance metrics for distance-aware local coreference resolution
381
382
383
384

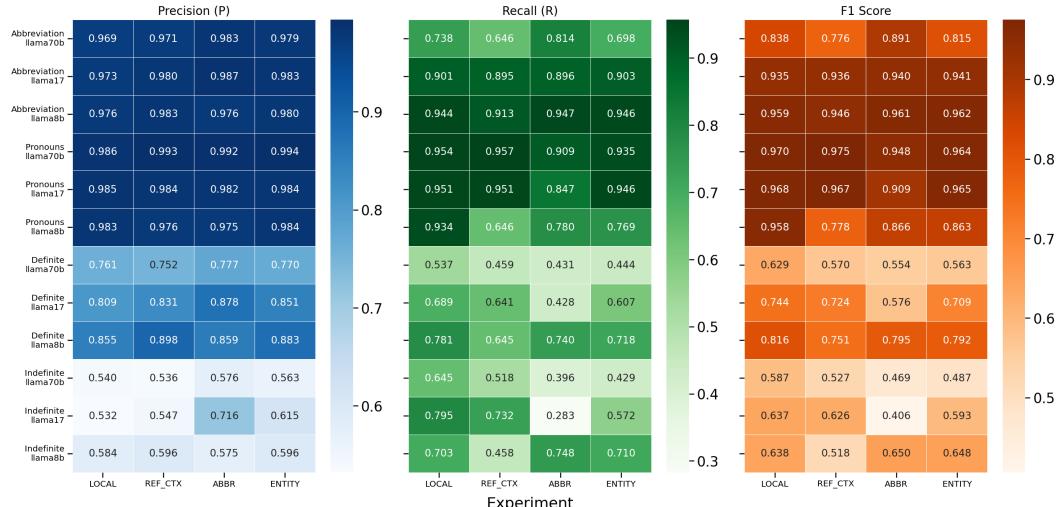
Model	Precision	Recall	F1 Score
LLaMA 70B	0.794	0.430	0.557
LLaMA 17B	0.841	0.554	0.668
LLaMA 8B	0.892	0.601	0.718

385
386
387 information. The LLaMA 8B model still achieved the highest F1 score. However, the decline in the
388 performance, as the context size increases is now proven by this experiment as well as (experiment
389 2: REF_CTX).
390

391 **Reference Context Has Mixed Effects.** In Experiment 2, which incorporated a fixed reference
392 paragraph to aid disambiguation, recall often dropped compared to the purely local setup (Experi-
393 ment 1), particularly for LLaMA 8B and 70B. This result suggests that unstructured introduction to
394 the entities does not necessarily guide the model to better locate the referencing in later paragraphs
395 in the paper.
396

397 **Structured Dictionaries Improve Recall.** Both the abbreviation-aware and entity-aware settings
398 (Experiments 3 and 4) demonstrated measurable gains in recall and F1, especially for the 8B model.
399 When supplied with structured input, such as abbreviation definitions or pre-extracted entity lists,
400 the models were better able to identify correct antecedents. This effect was most pronounced in
401 LLaMA 8B. These findings are consistent with evidence that structured, grounded prompting im-
402 proves performance in information extraction tasks.
403

404 Overall, these findings highlight both the promise and current limitations of generative LLMs for
405 biomedical coreference. While auxiliary signals such as abbreviation and entity dictionaries can
406 meaningfully boost performance, LLMs still struggle to integrate multi-paragraph context and re-
407 solve less explicit coreferences.
408

409 Model Performance by Coreference Type and Metric
410421 Figure 3: Heatmap of precision, recall, and F1 scores for LLaMA models (70B, 17B, 8B) across four
422 experimental setups (LOCAL, REF_CTX, ABBR, ENTITY) and coreference categories (pronouns,
423 indefinite NPs, abbreviations, definite NPs).
424

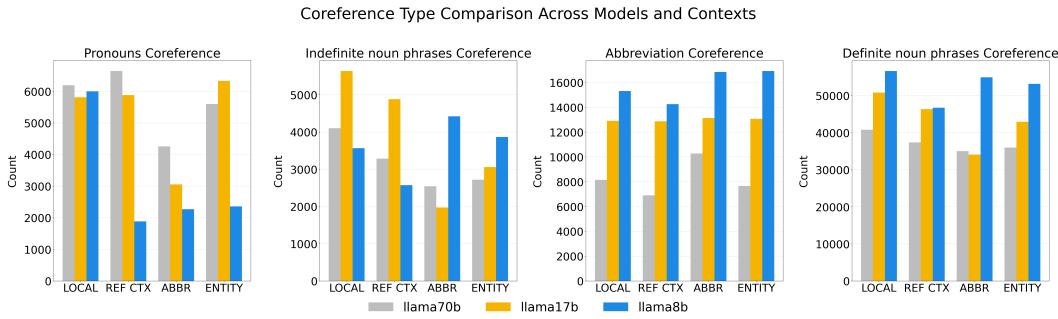
425 **Coreference Type Sensitivity and Model Behavior.** To better understand how the LLMs handle
426 different forms of coreference, we evaluated their performance across the four categories: pronouns,
427

432 indefinite noun phrases, abbreviations, and definite noun phrases, under the four contextual setups:
 433 LOCAL, REF_CTX, ABBR, and ENTITY.

434 To support this analysis, we developed a post-processing pipeline to classify the predicted and the
 435 CRAFT ground truth mentions into four types using lexical heuristics, as the CRAFT dataset does
 436 not label coreference types. Pronoun, indefinite, and abbreviations dictionaries we prepared from the
 437 source articles and are available on our GitHub. Remaining mentions were treated as definite noun
 438 phrases. Each was then evaluated using the model’s original prediction labels to compute precision,
 439 recall, and F1 scores per type.

440 As shown in Figure 3, pronoun coreference consistently achieved the highest F1 scores across all
 441 LLaMA models, with LLaMA 70B reaching 0.975 under the REF_CTX setup. This strong perfor-
 442 mance is likely due to the frequent occurrence of pronouns in pretraining corpora and their reliance
 443 on short-range syntactic cues. Complementary evidence from Figure 4 confirms that pronouns were
 444 also resolved in high absolute counts across experiments, especially by LLaMA 17B under minimal
 445 context.

446 Abbreviation coreference also exhibited strong performance, particularly under the ABBR and
 447 ENTITY experiments. Injecting abbreviation dictionaries yielded a noticeable increase in both F1
 448 scores (e.g., LLaMA 8B achieving 0.961 in ABBR) and mention resolution counts. These results
 449 affirm that domain-specific cues can significantly enhance model understanding of biomedical ab-
 450 breviation references.



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

466 7 CONCLUSION

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

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

481 ACKNOWLEDGMENTS

483 All Open-weight LLMs discussed in the paper were used for the purpose of inference and evalua-
 484 tion. OpenAI Models were used to refine approximately half of the manuscript, focusing solely on
 485 enhancing readability and grammatical accuracy. Great care was taken to ensure that no new content
 was introduced or that existing ideas were altered during this process.

486 REFERENCES
487

- 488 Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal,
489 Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are
490 few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- 491 Kevin Clark and Christopher D Manning. Deep reinforcement learning for mention-ranking coref-
492 erence models. *arXiv preprint arXiv:1609.08667*, 2016.
- 493 K Bretonnel Cohen, Arrick Lanfranchi, Miji Joo-young Choi, Michael Bada, William A Baumgart-
494 ner Jr, Natalya Panteleyeva, Karin Verspoor, Martha Palmer, and Lawrence E Hunter. Coreference
495 annotation and resolution in the colorado richly annotated full text (craft) corpus of biomedical
496 journal articles. *BMC bioinformatics*, 18(1):372, 2017.
- 497 Vladimir Dobrovolskii. Word-level coreference resolution. *arXiv preprint arXiv:2109.04127*, 2021.
- 498 Greg Durrett and Dan Klein. Easy victories and uphill battles in coreference resolution. In *Proce-
500 dings of the 2013 conference on empirical methods in natural language processing*, pp. 1971–1982,
501 2013.
- 502 Yujian Gan, Juntao Yu, and Massimo Poesio. Assessing the capabilities of large language models
503 in coreference: An evaluation. In *Joint 30th International Conference on Computational Linguis-
504 tics and 14th International Conference on Language Resources and Evaluation, LREC-COLING
505 2024*, pp. 1645–1665. European Language Resources Association (ELRA), 2024.
- 506 Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. Span-
507 bert: Improving pre-training by representing and predicting spans. *Transactions of the association
508 for computational linguistics*, 8:64–77, 2020.
- 509 Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. End-to-end neural coreference reso-
510 lution. *arXiv preprint arXiv:1707.07045*, 2017.
- 511 Kenton Lee, Luheng He, and Luke Zettlemoyer. Higher-order coreference resolution with coarse-
512 to-fine inference. *arXiv preprint arXiv:1804.05392*, 2018.
- 513 Yufei Li, Xiangyu Zhou, Jie Ma, Xiaoyong Ma, Pengzhen Cheng, Teliang Gong, and Chen Li.
514 Distinguished representation of identical mentions in bio-entity coreference resolution. *BMC
515 Medical Informatics and Decision Making*, 22(1):116, 2022.
- 516 Ruicheng Liu, Rui Mao, Anh Tuan Luu, and Erik Cambria. A brief survey on recent advances in
517 coreference resolution. *Artificial Intelligence Review*, 56(12):14439–14481, 2023.
- 518 Yanming Liu, Xinyue Peng, Jiannan Cao, Shi Bo, Yanxin Shen, Tianyu Du, Sheng Cheng, Xun
519 Wang, Jianwei Yin, and Xuhong Zhang. Bridging context gaps: Leveraging coreference resolu-
520 tion for long contextual understanding. *arXiv preprint arXiv:2410.01671*, 2024.
- 521 Pengcheng Lu and Massimo Poesio. Coreference resolution for the biomedical domain: A survey.
522 *arXiv preprint arXiv:2109.12424*, 2021.
- 523 Meta AI. Meta Llama 3.3 70B Instruct. [https://huggingface.co/meta-llama/
524 Llama-3.3-70B-Instruct](https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct), 2024a. Released July 23, 2024, under the Llama 3.1 Com-
525 munity License.
- 526 Meta AI. Meta Llama 3.1 8B Instruct. [https://huggingface.co/meta-llama/
527 Llama-3.1-8B-Instruct](https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct), 2024b. Released July 23, 2024, under the Llama 3.1 Community
528 License.
- 529 Meta AI. Meta Llama 4 Scout 17B-16E Instruct. [https://huggingface.co/meta-llama/
530 Llama-4-Scout-17B-16E-Instruct](https://huggingface.co/meta-llama/Llama-4-Scout-17B-16E-Instruct), 2025. Released April 5, 2025, under the Llama 4
531 Community License.
- 532 Vincent Ng and Claire Cardie. Improving machine learning approaches to coreference resolution.
533 In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pp.
534 104–111, 2002.

540 Brendan O’Connor and Michael Heilman. Arkref: A rule-based coreference resolution system.
 541 *arXiv preprint arXiv:1310.1975*, 2013.
 542

543
 544 Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. Stanza: A python
 545 natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*,
 546 2020.
 547

548
 549 Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language under-
 550 standing by generative pre-training. 2018.
 551

552
 553 Wee Meng Soon, Hwee Tou Ng, and Daniel Chung Yong Lim. A machine learning approach to
 554 coreference resolution of noun phrases. *Computational linguistics*, 27(4):521–544, 2001.
 555

556 Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée
 557 Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and
 558 efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
 559

560
 561 Yuli Vasiliev. *Natural language processing with Python and spaCy: A practical introduction*. No
 562 Starch Press, 2020.
 563

564 Sam Joshua Wiseman, Alexander Sasha Rush, and Jason Weston. Learning anaphoricity and an-
 565 tecendent ranking features for coreference resolution. Association for Computational Linguistics,
 566 2015.
 567

568 Wei Wu, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li. Corefqa: Coreference resolution as
 569 query-based span prediction. In *Proceedings of the 58th annual meeting of the association for*
 570 *computational linguistics*, pp. 6953–6963, 2020.
 571

572 Tianbao Xie, Chen Henry Wu, Peng Shi, Ruiqi Zhong, Torsten Scholak, Michihiro Yasunaga,
 573 Chien-Sheng Wu, Ming Zhong, Pengcheng Yin, Sida I Wang, et al. Unifiedskg: Unifying and
 574 multi-tasking structured knowledge grounding with text-to-text language models. *arXiv preprint*
 575 *arXiv:2201.05966*, 2022.
 576

582
 583
 584
 585 **A APPENDIX**
 586

587
 588 **A.1 PROMPT TEMPLATE**
 589

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

594

System Prompt

595

596

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.

599

Skip the following:

600

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

603

Follow these steps:

604

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.

605

606

607

608

609

610

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

611

612

613

614

Expected JSON Output Schema:

615

```
{
  "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]"
}
```

625

626

Example:

627

628

Input: “These results were unexpected. They indicate a new trend.”

629

Rewritten: “The results were unexpected. The results indicate a new trend.”

630

631

632

Example Output:

633

```
{
  "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..."
  }
}
```

645

646

647