Few-Shot Multilingual Coreference Resolution Using Long-Context Large Language Models

Moiz Sajid, Seemab Latif, Zuhair Zafar, Muhammad Moazam Fraz
National University of Science and Technology, Islamabad, Pakistan
msajid.phdai24seecs@seecs.edu.pk; seemab.latif@seecs.edu.pk;
zuhair.zafar@seecs.edu.pk; moazam.fraz@seecs.edu.pk

Abstract

In this work, we present our system, which ranked second in the CRAC 2025 Shared Task on Multilingual Coreference Resolution (LLM Track). For multilingual coreference resolution, our system mainly uses long-context large language models (LLMs) in a few-shot in-context learning setting. Among the various approaches we explored, few-shot prompting proved to be the most effective, particularly due to the complexity of the task and the availability of high-quality data with referential relationships provided as part of the competition. We employed Gemini 2.5 Pro, one of the best available closed-source long-context LLMs at the time of submission. Our system achieved a CoNLL F1 score of 61.74 on the mini-test set, demonstrating that performance improves significantly with the number of few-shot examples provided, thanks to the model's extended context window. While this approach comes with trade-offs in terms of inference cost and response latency, it highlights the potential of long-context LLMs for tackling multilingual coreference without task-specific fine-tuning. Although direct comparisons with traditional supervised systems are not straightforward, our findings provide valuable insights and open avenues for future work, particularly in expanding support for low-resource languages.

1 Introduction

Ever since the work of (Brown et al., 2020) showed that a general-purpose language model, trained on diverse internet-scale data, could perform well across a vast range of NLP tasks, most complex NLP tasks are now tackled first through text generation LLMs. However, there have been only a handful of works on the task of coreference resolution using text generation LLMs. This work presents the second-best approach on the CRAC 2025 Shared Task on Multilingual Coreference Resolution (LLM Track) that utilizes the current state-of-the-art LLM model named Gemini 2.5 Pro (Comanici et al., 2025) from Google. This is the fourth edition of the shared task. Previous editions of the shared task were conducted successfully in 2022 (Žabokrtský et al., 2022), 2023 (Žabokrtský et al., 2023), and 2024 (Novák et al., 2024).

Coreference resolution remains one of the most challenging tasks in natural language processing (NLP), as it requires a comprehensive understanding of language at multiple levels, including semantics, syntax, discourse structure, and pragmatics. The complexity of this task is further amplified in multilingual settings, where variations in linguistic phenomena, grammatical structures, and referential expressions across languages introduce additional challenges. Despite its importance, research on multilingual coreference resolution remains relatively limited, leaving significant gaps in methodologies and resources for addressing this problem effectively.

Our work is, to the best of our knowledge, the first to leverage large language models (LLMs) with extended context lengths for the task of multilingual coreference resolution. Inspired by prior approaches that formulate coreference resolution as a text generation problem (Skachkova, 2024) (Le and Ritter, 2023) (Gan et al., 2024), our method processes raw text as input and directly generates text annotated with coreference clusters as output.

The main contributions of this work are as follows:

- We present the second-best performing system in the CRAC 2025 Shared Task on Multilingual Coreference Resolution (LLM Track), trailing the top submission by a margin of only 1.22.
- We model multilingual coreference resolution as an end-to-end text generation task, enabling the system to learn from few-shot examples with long context spans.

Proceedings of the Eighth Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2025), pages 154–162
Suzhou, China, November 9, 2025. ©2025 Association for Computational Linguistics

- We utilized the current state-of-the-art Gemini 2.5 Pro model for this challenging task, demonstrating its effectiveness in handling large-context, multilingual coreference resolution.

2 Related Work

Early approaches to coreference resolution typically adopted a two-stage framework, first identifying coreferent mentions and then using these mentions to construct coreference clusters. This paradigm was also employed by last year's shared task winner (Straka, 2024). In contrast, several subsequent studies have explored coreference resolution as an end-to-end problem, jointly performing mention detection and coreference clustering. Notable contributions in this direction include the works of (Lee et al., 2017), (Lee et al., 2018), (Joshi et al., 2019), and (Joshi et al., 2020). Most end-to-end methods build upon the foundational work of (Lee et al., 2017), which was the first to train a coreference model in a fully end-to-end manner, unlike prior approaches that relied on external systems for mention detection or clustering.

(Liu et al., 2022) explicitly models the structure of coreference resolution using language models, achieving state-of-the-art results on the OntoNotes benchmark from the CoNLL-12 English shared task dataset (Pradhan et al., 2013). Similarly, (Bohnet et al., 2023) demonstrates strong performance on the same benchmark through a text-to-text generation paradigm, where the text is processed in an autoregressive manner. Their approach employs a Link-Append transition system that encodes previously established coreference links and incrementally predicts new ones.

Our approach leverages end-to-end text generation LLMs for multilingual coreference resolution, enabling the effective application of few-shot prompting strategies on raw input texts paired with their gold annotations from the training split. Furthermore, we extend this methodology by utilizing recent long-context LLMs, which have demonstrated state-of-the-art performance across a wide range of NLP tasks.

3 Experiments

3.1 Dataset

The data utilized in this work is derived from the CRAC 2025 Shared Task, now in its fourth edition, and based on the CorefUD 1.3 collection. This year's dataset encompasses 17 languages, representing an increase of 6, 4, and 1 languages compared to the 2022, 2023, and 2024 editions, respectively. To facilitate the use of large language models (LLMs) such as GPT-4o, LLaMA, and Claude for the coreference resolution task, the organizers released a text-to-text version of the dataset in addition to the standard CoNLL-U format. This alternative representation proved advantageous for our method, as it enabled the effective application of few-shot prompting strategies. Examples of input–output pairs employed in our system for few-shot prompting are provided in Figures 1 and 2. The token counts for the 22 datasets across the training, development, and test splits are reported in Appendix A.1.

LLM	Few-Shot Examples Token Count	en gum	en litbank
Gemini 2.5 Flash	0	32.10	45.65
	100,000	46.68	63.14
	200,000	47.25	51.61
	300,000	50.84	44.16

Table 1: Results on the English-GUM and English-LitBank development sets when varying the token counts of few-shot examples are reported below. In certain cases, Gemini 2.5 Flash returned empty responses, which contributed to the observed performance degradation.

```
How to Prepare Quinoa Quinoa is known as the
little rice of Peru . The Incas treated the crop
as sacred and referred to quinoa as " chisaya
mama " or " mother of all grains . " [ 1 ] By
tradition , the Inca emperor would sow the first
seeds of the season using " golden implements .
" Quinoa is rich in protein and much lighter
than other grains .
```

Figure 1: A sample raw text instance from the English-GUM training split.

3.2 Approach

As mentioned earlier, we adopted an end-to-end text-to-text approach, where the model receives an input text and is required to return the same text with coreference annotations. Initially, we experimented with zero-shot prompting; however, this strategy yielded poor results since our prompt failed to capture all the complexities necessary for effective coreference resolution. We then moved to a 4-shot prompting setup using the same instructions, which produced more reasonable results. These four-shot examples were incorporated into

```
How to Prepare Quinoa|[e1] Quinoa|[e1] is known
as the|[e1 little rice of Peru|[e2],e1] . The|
[e3 Incas|e3] treated the|[e1 crop|e1] as sacred
and referred to quinoa|[e1] as " chisaya|[e1
mama|e1] " or " mother|[e1 of all|[e4 grains|e1],
e4] . " [ 1|[e5] ] By tradition|[e6] , the|[e7
Inca emperor|e7] would sow the|[e8 first seeds
of the|[e9 season|e8],e9] using " golden|[e10
implements|e10] . " Quinoa|[e11] is rich in
protein|[e12] and much lighter than other|[e13
grains|e13] .
```

Figure 2: A gold-annotated version of the same text shown in Figure 1, taken from the English-GUM training split.

the system prompt used for all our experiments and submissions, with the complete prompt provided in Appendix A.2. These early experiments suggested that few-shot prompting, supported by well-curated examples, could be a more effective approach to this problem.

To validate our hypothesis that incorporating a large number of well-curated few-shot examples enhances model performance, we conducted a small ablation study on the development splits of the English-GUM and English-LitBank datasets from the shared task. The token count distributions for these datasets, along with others used in the task, are provided in Appendix A.1. For this experiment, we employed Gemini 2.5 Flash, as it offers a more cost-efficient alternative to Gemini 2.5 Pro, particularly when handling long context lengths. The results, shown in Table 1, indicate that increasing the number of high-quality examples generally improves performance, although occasional instances where Gemini 2.5 Flash produced empty outputs adversely impacted the overall outcomes of the ablation study.

To build on this insight, we employed a dynamic few-shot learning strategy using Google's Gemini 2.5 Pro model for coreference resolution tasks. For each test instance, the system dynamically constructs a context window by selecting language-specific training examples and their corresponding gold-standard annotations, then shuffling them randomly to avoid ordering bias. For instance, we combined multiple training datasets of the same language before selecting them as few-shot examples. The approach leverages adaptive context management, progressively adding training examples as human-AI dialogue pairs until reaching a 300,000-token limit, ensuring optimal use of the model's context window while maintaining computational efficiency. For certain datasets, such as Czech-PCEDT, English-GUM, English-LitBank, and Hungarian-KorKor, we utilized up to 500,000 tokens for few-shot examples, while for the remaining datasets, we limited the few-shot examples to 300,000 tokens in our submission. It is important to note that we did not use a fixed number of shots across all datasets.

Each test query is processed within a structured prompt framework that includes a system prompt, a few randomized few-shot examples, and the target input, enabling the model to learn task-specific patterns in-context without parameter updates. This methodology supports language-adaptive processing by automatically selecting relevant examples for the target language and provides a scalable, multilingual framework for evaluating coreference resolution across diverse linguistic settings. We set the temperature parameter to zero across all experiments to ensure deterministic outputs and suppress stochastic or creative variations in the LLM's responses.

The complete system prompt used in our approach is provided in Appendix A.2. Our prompt design did not capture all the intricacies required for effectively solving the coreference resolution task. We included only a limited set of instructions and omitted explicit guidance for handling zero mentions. Nevertheless, through few-shot prompting strategies, our system was able to implicitly learn and annotate zero mentions successfully.

4 Shared Task Results

The official results of the shared task are summarized in Table 2. Our system, NUST-FewShot, ranked second among four participating submissions, achieving an average CoNLL F1 score of 61.74, the primary evaluation metric of the task. The CoNLL F1 score is computed as the unweighted average of the F1 scores from MUC, B-cubed, and CEAFe. Given the multilingual nature of the datasets, the final score is reported as the macro-average of the individual CoNLL F1 scores across all languages.

In addition to the primary CoNLL F1 metric, three alternative evaluation metrics are reported in Table 2: partial matching, exact matching, and head matching with singletons included. Under partial matching without singletons, our system performs nearly on par with the top-ranked system. However, the performance gap becomes slightly

System	head match	partial match	exact match	head match (with singletons)
GLaRef-CRAC25	**62.96**	**61.66**	**58.98**	**65.61**
NUST-FewShot	61.74	61.14	56.34	63.44
PUXCRAC2025	60.09	59.68	55.22	54.77
UWB	59.84	59.55	38.81	62.77

Table 2: The table presents the results of all systems participating in the CRAC 2025 Shared Task on Multilingual Coreference Resolution (LLM Track). The primary evaluation metric is the CoNLL F1 score, reported in the second column labeled head-match. Our system, NUST-FewShot, achieved the second-best overall performance among the submitted systems.

more pronounced under the exact matching without singletons and head matching with singletons metrics.

Table 3 reports our system's official test results on the shared-task language-specific datasets, whereas Table 5 (A.3) reports those of the three best overall systems. Our system achieved the best performance on 10 out of the 22 datasets. Nonetheless, a notable performance gap remains between our system and the top-performing language-specific models for several datasets, particularly Catalan-AnCora, Czech-PCEDT, and Czech-PDT. We hypothesize that this discrepancy may be due to our model's limited ability to capture fine-grained linguistic nuances unique to these languages, despite the availability of a substantial number of gold-annotated examples in their training sets. Overall, while our approach demonstrates competitive results on nearly half of the datasets, further ablation studies are necessary to better understand its strengths and weaknesses and to explore strategies for improving cross-linguistic adaptability.

5 Conclusion

This paper presented the second-best performing system in the CRAC 2025 Shared Task (LLM Track). Our approach achieved top performance on 10 out of the 22 datasets in the competition. We employed an end-to-end text generation framework leveraging few-shot learning with Gemini 2.5 Pro, a state-of-the-art long-context LLM, which processes raw text as input and produces coreference-annotated text as output. Coreference resolution with LLMs remains a nascent area of research, with only a handful of recent studies addressing this challenge. We hope that our work not only demonstrates the potential of LLM-based approaches for this task but also paves the way for future research exploring this promising direction.

Dataset	CoNLL score
ca anc	60.87
cs pce	51.36
cs pdt	54.30
cu pro	58.48
de pot	48.74
en gum	69.78
en lit	70.38
es anc	61.75
fr anc	71.94
fr dem	57.59
grc pro	57.85
hbo ptn	80.15
hi hdt	71.32
hu kor	43.49
hu sze	52.27
ko ecm	66.05
lt lcc	59.16
no bok	72.76
no nyn	68.86
pl pcc	70.83
ru ruc	71.40
tr itc	39.00

Table 3: The table shows CoNLL scores for our system across the 22 test datasets from the CRAC 2025 Shared Task.

Limitations

The primary limitations of our work are the reliance on multiple LLM calls and the associated computational cost, which can become significant. Our approach utilizes Gemini 2.5 Pro with a large context window via an API interface, leading to high costs due to the extensive token usage from few-shot examples, lengthy input texts, and generated outputs with predictions. Furthermore, we did not investigate other advanced long-context models, such as OpenAI's GPT-4.1 and Meta's LLaMA 4 Scout, which support context lengths of up to 1 million and 10 million tokens, respectively.

References

Bernd Bohnet, Chris Alberti, and Michael Collins. 2023. Coreference resolution through a seq2seq transition-based system. *Transactions of the Association for Computational Linguistics*, 11:212–226.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3290 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

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

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.

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel Weld. 2019. 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. Association for Computational Linguistics.

Nghia T. Le and Alan Ritter. 2023. Are large language models robust coreference resolvers? *Preprint*, arXiv:2305.14489.

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

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. 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 687–692, New Orleans, Louisiana. Association for Computational Linguistics.

Tianyu Liu, Yuchen Eleanor Jiang, Nicholas Monath, Ryan Cotterell, and Mrinmaya Sachan. 2022. Autoregressive structured prediction with language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 993–1005, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Michal Novák, Barbora Dohnalová, Miloslav Konopik, Anna Nedoluzhko, Martin Popel, Ondrej Prazak, Jakub Sido, Milan Straka, Zdeněk Žabokrtský, and Daniel Zeman. 2024. Findings of the third shared task on multilingual coreference resolution. In *Proceedings of the Seventh Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 78–96, Miami. Association for Computational Linguistics.

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. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152, Sofia, Bulgaria. Association for Computational Linguistics.

Natalia Skachkova. 2024. Multilingual coreference resolution as text generation. In *Proceedings of the Seventh Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 114–122, Miami. Association for Computational Linguistics.

Milan Straka. 2024. CorPipe at CRAC 2024: Predicting zero mentions from raw text. In *Proceedings of the Seventh Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 97–106, Miami. Association for Computational Linguistics.

Zdeněk Žabokrtský, Miloslav Konopik, Anna Nedoluzhko, Michal Novák, Maciej Ogrodniczuk,

Martin Popel, Ondrej Prazak, Jakub Sido, and Daniel Zeman. 2023. Findings of the second shared task on multilingual coreference resolution. In *Proceedings of the CRAC 2023 Shared Task on Multilingual Coreference Resolution*, pages 1–18, Singapore. Association for Computational Linguistics.

Zdeněk Žabokrtský, Miloslav Konopík, Anna Nedoluzhko, Michal Novák, Maciej Ogrodniczuk, Martin Popel, Ondřej Pražák, Jakub Sido, Daniel Zeman, and Yilun Zhu. 2022. Findings of the shared task on multilingual coreference resolution. In *Proceedings of the CRAC 2022 Shared Task on Multilingual Coreference Resolution*, pages 1–17, Gyeongju, Republic of Korea. Association for Computational Linguistics.

A Appendix

A.1 Data Distribution

Table 4 presents the token count distribution across the various datasets included in the shared task. The token counts are computed specifically using the Gemini 2.5 Pro tokenizer[1], as our submission was based on this model. Token counts may vary slightly when calculated with alternative tokenizers due to differences in their tokenization strategies.

A.2 Prompt

We used the following system prompt in all our experiments and results:

"""

You are a coreference resolution annotator.

Your job is to read a multilingual passage and annotate all mentions that refer to the same underlying entity (could be a word or many words) using a unique identifier with a bracketed pipe-based format. Understand proper context of the text before making the annotations. Prioritize resolving pronouns based on proximity and grammatical role, but consider the semantic context to avoid incorrect annotations.

Format Rules:

1. Surround every span referring to a shared entity with the format: ‘mention text|[eID]‘ where eID is a unique entity ID (e1, e2, ...).

2. If an entity contain multiple words, start the annotation with a single pipe as ‘|[eID‘ and close it with a single pipe as ‘|eID]‘ for example ‘|[eID - entity comes here|eID]‘

3. If a span refers to **multiple entities**, use: ‘text|[eA,[eB] tail|eA],eB]‘ (notice)

4. Use the **same ID** consistently for all mentions of the same entity, even across paragraphs.

5. Do **not annotate singleton mentions** (those that appear only once in the text).

6. Annotate **all types of coreference**: full noun phrases, pronouns, nested noun mentions, and even abstract or generic references like "such outcomes", "it", etc.

7. If there is **nested structure**, use proper nesting with comma-separated closing IDs.

8. Do not resolve ’it’ if it refers to an implied or abstract concept (e.g., ’It is widely believed...’).

Example 1:

Original:
Alice went to the park . She brought her dog .

Annotated:
Alice|[e1] went to the park|[e2] . She|[e1] brought her|[e1] dog|[e3] to the park|[e2] .

Example 2:

Original:
Education and early loves Alina gained her early formal education at Aberdeen Grammar School , and in August 1799 entered the school of Dr. William Glennie , in Dulwich . [17] Placed under the care of a Dr. Bailey , she was encouraged to exercise in moderation but not restrain herself from " violent " bouts in an attempt to overcompensate for her deformed foot .

Annotated:
Education|[e1] and early|[e2 loves|e2] Alina|[e3] gained her|[e1,[e3] early formal education at Aberdeen|[e4],[e5 Grammar School|e1],e5] , and in August|[e6 1799|[e7],e6] entered the|[e8 school of Dr.|[e9 William Glennie|e9] , in Dulwich|[e10],e8] . [17|[e11]] Placed under the|[e12 care of a|[e13 Dr. Bailey|e12],e13] , she|[e3] was encouraged to

[1] https://ai.google.dev/gemini-api/docs/tokens

Language	Train	Val	Test	Train (including gold annotations)	Val (including gold annotations)
Catalan	483,179	36,371	36,363	795,896	60,446
Czech	2,947,128	92,391	91,706	4,785,013	156,544
Old Church Slavonic	150,355	24,262	18,980	272,045	41,028
German	38,405	4,841	4,623	68,321	8,613
English	375,074	49,229	49,547	864,829	113,484
Spanish	448,164	30,796	30,909	806,930	56,228
French	728,738	61,596	61,100	1,972,972	159,321
Ancient Greek	170,245	10,787	13,288	311,317	18,477
Ancient Hebrew	35,346	48,014	46,851	46,473	63,500
Hindi	48,633	12,482	28,728	88,993	21,456
Hungarian	234,845	28,304	27,543	333,437	40,497
Korean	957,191	61,049	60,973	1,557,657	99,991
Lithuanian	63,638	7,353	7,500	82,708	9,231
Norwegian	595,703	61,901	59,054	1,399,765	143,566
Polish	747,247	43,515	43,105	1,739,784	102,615
Russian	194,110	33,118	18,434	262,653	45,792
Turkish	81,379	8,757	9,523	196,850	20,726
Total	8,299,380	614,766	608,227	15,585,643	1,161,515

Table 4: The token counts for all languages in the shared task, after merging datasets for languages with multiple sources, are reported on a split-wise basis. The last two columns additionally account for tokens from the gold annotations.

exercise in moderation but not restrain herself|[e3] from "|[e14 violent " bouts|e14] in an|[e15 attempt to overcompensate for her|[e16,[e3] deformed foot|e15],e16] .

Example 3:

Original:
Los jugadores de el Espanyol aseguraron hoy que prefieren enfrentar se a el Barcelona en la final de la Copa de el Rey en lugar de en las semifinales , tras clasificar se ayer ambos equipos catalanes para esta ronda . La mayoría de los jugadores españolistas expresaron su opinión de que sería más fácil vencer a su máximo rival en un solo partido que tener que enfrentar se a el conjunto de Louis Van Gaal en las semifinales , donde tendrían que disputar una eliminatoria de ida y vuelta .

Annotated:
Los|[e1 jugadores de el Espanyol|[e2],e1] aseguraron hoy que prefieren ##|[e1] enfrentar se a el Barcelona|[e3] en la|[e4 final de la|[e5 Copa de el Rey|e4],e5] en lugar de en las|[e6 semifinales|e6] , tras clasificar se ayer ambos|[e7 equipos catalanes|e7] para esta|[e6 ronda|e6] . La|[e1 mayoría de los jugadores españolistas|e1] expresaron su|[e1] opinión de que sería más fácil vencer a su|[e2],[e3 máximo rival|e3] en un solo partido que tener que enfrentar se a el conjunto|[e3 de Louis Van Gaal|e3] en las|[e6 semifinales , donde|[e6] tendrían ##|[e7] que disputar una|[e8 eliminatoria de ida y vuelta|e6],e8] .

Example 4:

Original:
Zèbre Zèbre est un nom vernaculaire , ambigu en français , pouvant désigner plusieurs espèces différentes d’ herbivores de la famille de les équidés , et de le genre Equus , vivant en Afrique . Ils se trouvent principalement en Afrique centrale et australe . Ces animaux se caractérisent par des bandes de rayures verticales noires et blanches .

Annotated:
Zèbre|[e1] Zèbre|[e1] est un|[e2 nom vernaculaire , ambigu|e2] en français|[e3] , pouvant désigner plusieurs|[e4 espèces différentes d’ herbivores|[e5] de la|[e6 famille de les|[e7 équidés|e6],e7] , et de le|[e8 genre Equus|e8] , vivant en Afrique|[e9],e4] . Ils|[e4] se trouvent principalement en Afrique|[e10 centrale et australe|e10] .
"""

A.3 Top Systems Results

Table 5 presents the CoNLL scores for the top three overall best-performing systems across the 22 test datasets from the CRAC 2025 Shared Task.

System	ca anc	cs pce	cs pdt	cu pro	de pot	en gum	en lit	es anc	fr anc
GLaRef-CRAC25	**73.45**	**65.12**	**71.33**	58.25	**59.60**	58.73	69.01	**74.43**	66.74
NUST-FewShot	60.87	51.36	54.30	**58.48**	48.74	**69.78**	**70.38**	61.75	**71.94**
PUXCRAC2025	68.01	56.94	62.96	43.74	57.41	61.71	69.12	70.52	63.77

System	fr dem	grc pro	hbo ptn	hi hdt	hu kor	hu sze	ko ecm	lt lcc	no bok
GLaRef-CRAC25	60.43	**65.75**	43.96	56.36	**52.53**	59.82	63.04	62.55	64.74
NUST-FewShot	57.59	57.85	**80.15**	**71.32**	43.49	52.27	**66.05**	59.16	**72.76**
PUXCRAC2025	**61.54**	47.86	45.31	66.85	50.58	**61.61**	50.32	**65.35**	65.18

System	no nyn	pl pcc	ru ruc	tr itc
GLaRef-CRAC25	61.63	**72.55**	68.79	**56.23**
NUST-FewShot	**68.86**	70.83	**71.40**	39.00
PUXCRAC2025	63.00	66.55	67.59	56.06

Table 5: The table shows the CoNLL scores for the top three overall best-performing systems across the 22 test datasets from the CRAC 2025 Shared Task. Our system, NUST-FewShot, achieved the best performance on 10 of the 22 datasets, surpassing the overall top-ranked system, GLaRef-CRAC25, which led on 9 of the 22 datasets.