Generative Multilingual Coreference Resolution at CRAC 2026

Jakub Hejman and Ondřej Pražák and Miloslav Konopík

{hejmanj,ondfa,konopik}@fav.zcu.cz

Department of Computer Science and Engineering,
NTIS – New Technologies for the Information Society,
Faculty of Applied Sciences, University of West Bohemia, Technická 8, 306 14 Plzeň
Czech Republic

Abstract

Participating again in this year's edition of the CRAC shared task on coreference resolution, we present our upgraded system with an official uplift of 15.46 percentage points in CoNLL-U score. We incorporated the larger Gemma 3 27B IT model, joint pre-training, headword tagging, more efficient training and inference as well as a sliding window to achieve this result. Our system placed second in the LLM track and third overall with a primary score of 73.83. We reached the highest scores on two datasets. Finally, we compare specialized and general LLM approaches.

1 Introduction

Coreference resolution (CR) is a foundational natural language understanding task involving the identification and clustering of textual expressions that refer to the same real-world entity. While modern Large Language Models (LLMs) routinely demonstrate the ability to implicitly resolve references during broader downstream applications, explicit CR remains valuable for benchmarking. It serves as a probe for evaluating the long-context reasoning, entity-tracking and long-form structured generation of generative models.

This paper describes our submission to the Computational Models of Reference, Anaphora and Coreference (CRAC) 2026 Shared Task on multilingual CR (Novák et al., 2026a). Using the CorefUD 1.4 collection (Novák et al., 2026b) which was extended by five datasets since the last version, the task challenges participants to develop systems capable of identifying mentions, reconstructing zero mentions, and clustering them into coreferential entities. The competition features two tracks: the traditional Unconstrained Track and the Large Language Model (LLM) Track. Our participation in the LLM track builds on our work from the track's inaugural run last year (Hejman et al., 2025).

Our core system remains an end-to-end generative tagger trained via QLoRA (Dettmers et al., 2023). To simplify the task, we predict only the headwords (Dobrovolskii, 2021) of mentions rather than their full spans. This year, our primary strategy is scaling up: we transition to the larger Gemma 3 27B (Gemma Team, 2025) model to predict coreference chains across full documents in a plaintext format. The main logistical challenge of the shared task is managing 27 distinct datasets, each with varying features, and executing training and inference across all of them. Exceptionally long documents with long-distance coreferences add another layer of difficulty. Fitting the full context of these documents along with the much larger LLM into GPU memory while maintaining high hardware utilization is a significant challenge. To improve the efficiency of our system, we introduced a sliding context window with overlap, utilizing the same Llama 3.1 8B foundational model (Grattafiori et al., 2024) as last year. While this sliding window was ultimately deployed on only five datasets, addressing this efficiency bottleneck is a key challenge for processing real-world data.

The results of this year's iteration align with expected scaling gains, tracking with the broader trend of LLMs closing the performance gap against specialized, unconstrained systems. Our submission placed second in the LLM track and third overall. On the primary CoNLL-U metric, our system achieved a score of 73.83. This makes our system highly competitive within our track, only narrowly behind the LLM track winner (74.32, +0.49), and notably close to the long-term dominant specialized system, *corpipe*, which scored 77.11 in the Unconstrained track. These findings suggest that while a slight gap remains on the absolute top of the leaderboard, generative models can offer highly competitive performance across a vast array of typologically diverse languages without the need for complex, task-specific architectures.

Proceedings of the 2nd Joint Workshop on Computational Approaches to Discourse, Context and Document-Level Inferences and Computational Models of Reference, Anaphora and Coreference (CODI-CRAC 2026), pages 162–166

2 System

2.1 Implementation

Our methodology approaches coreference resolution as a generative tagging task. For our primary, full-document system, we introduced two major upgrades over our previous baseline: the utilization of the Unsloth library (Han et al., 2023) for highly optimized QLoRA fine-tuning, and the adoption of instruction-tuned models. Joint pre-training remained a core part of our strategy. We first trained the model on a concatenation of all 27 datasets and then each one separately.

The use of Unsloth made the training of the larger Gemma 3 27B model feasible. Our system design explicitly targeted NVIDIA A40 GPUs due to their broad availability on MetaCentrum, the Czech national infrastructure for scientific and academic grid computing. All QLoRA trainings utilized a rank of $r = 64$.

2.2 Task Formulation

In order to accommodate chat-tuned models, we broke up the input and tagged parts of the text into a user and an assistant message, respectively. To maintain backward compatibility with base models, we set a chat template which results in a simple concatenation of the parts, like in our previous system. Our response-only training approach was updated to work with Unsloth's utilities and chat models.

To simplify the generation target, we employed a headword-only tagging strategy rather than tagging full mention spans.

Because the target sequence requires the model to reproduce the input text interleaved with dense annotation tags, the output length can often exceed twice the input length[1]. Any truncation of the output severely penalizes the final evaluation score, as the system fails to output the remainder of the document. To prevent this in our primary non-sliding system, we utilized vLLM (Kwon et al., 2023) for high-throughput inference and set `max_new_tokens` to the minimum of the model's maximum context limit (131,072) and the expected number of tokens in the output with a pessimistic annotation factor.

Figure 1 illustrates the schema of our input and output formatting, highlighting how the context is structured during inference.

[1]This analysis was presented in our last year's system paper.

2.3 Sliding Window

Our standard sliding configuration uses a window of 2,000 tokens with a 700-token overlap.

We ensure that window breaks occur strictly at sentence boundaries. Because mentions cannot cross sentence boundaries in the CorefUD format, cutting sentences would lead to incomplete annotations. We only resort to splitting within a sentence in unavoidable cases.

In our generative sliding window approach, the overlap is a cleaned[2] version of the last part of the previous iteration's generation. This way we sidestep the problem of reconciling two annotations of the same text, while aligning the generation closely to the model's training distribution. Additionally, cleaning the overlap should lead to better results due to the accurate representation of the current state being presented to the model and avoidance of any potential invalid markup being present in the context.

We evaluated the end-to-end formulation on full documents in all datasets where possible, and supplemented with the sliding context formulation on problematic sets. The sliding window achieved better development set scores on five of the datasets than our best single-pass attempt. Our final submission combined the best-performing approach for each respective dataset.

3 Results & Discussion

3.1 Result Analysis

The official shared task results, summarized in Table 1, display several interesting trends regarding the competitive landscape of coreference resolution. The top group of submissions is tightly clustered in the 72–77 average score range. This top tier consists of three variants of *corpipe* (Straka, 2025), a highly optimized, specialized architecture with a long-running winning streak in this competition, and the two leading LLM-based submissions (Antoine Bourgois and our system, LLM-UWB). Following this leading group, there is a distinct drop-off to a middle tier of systems scoring around 67–68, with the baseline and the remaining participants trailing further behind.

One of the most notable takeaways from this year's results is the strong showing of generative models. While *corpipe-ensemb* retained the highest average score, LLM track participants man-

[2]Processed using the official shared task output cleaner implementation modified to handle our headword-only format.

```
<start_of_turn>user
Rozděluje se to na šach praktický a potom korespondenční nebo taky , řekněme , internetový .
Praktickému šachu jsem se věnoval doma studiem z knížek a pak , když byly turnaje , třeba turnaj
družstev , jsme hráli o víkendech , sobota , neděle .

### LABEL<end_of_turn>
<start_of_turn>model

Rozděluje se to na šach|[e1] praktický|[e2] a potom korespondenční nebo taky , řekněme ,
internetový . Praktickému šachu|[e2] jsem se věnoval ##já|[e3] doma studiem z knížek a pak
, když byly turnaje , třeba turnaj družstev , jsme hráli ##my|[e4] ##někoho|[e2] o víkendech ,
sobota , neděle .<end_of_turn>
```

Figure 1: An illustration of the formatting during a sliding window iteration. The prompt consists of three main parts: the input (light blue highlight), output separator (brown highlight) and the expected output (gray highlight). The inserted zero mentions (pro-drops) are colored green, headword tags orange. The text overlapping with the previous window within the output field is dark blue. Tokens corresponding to the Gemma 3 chat template are bold.

aged to secure the best score on 7 out of the 27 datasets. Our system, LLM-UWB, achieved the highest score on two datasets: `no_nynorsknarc` (79.45) and `fr_ancor` (77.51). Furthermore, on the majority of the datasets, our scores traced closely behind the state-of-the-art.

However, our average score was heavily impacted by significant performance drops on a specific subset of datasets. As shown in the table, our system lagged noticeably behind the best scores on `fr_litbankfr` (-21.99), `nl_openboek` (-11.28), `cu_proiel` (-9.17), and `fr_democrat` (-7.14). This can be attributed to the sliding window fallback and the smaller Llama 3.1 8B model that were used for these corpora with the exception of `cu_proiel`. While this significantly reduced inference times and allowed us to avoid zero scores as the worst case, it was not our best configuration.

3.2 Comparing Across Years

To properly contextualize these results, it is useful to contrast them with the previous year's CRAC shared task. However, several factors complicate a direct comparison. First, the introduction of new datasets alters the overall average. We resolve this by recalculating the macro-average strictly across the intersection of datasets shared between both years; all scores reported in this subsection reflect this adjusted subset.

Second, there have been minor modifications to the CoNLL-U $\leftrightarrow$ Text converter. Our empirical testing indicates that these changes introduce minimal variance (typically yielding a score difference of less than 1, most often 0). This impact is negligible compared to the inherent information loss of the conversion process itself, which averages a 0.46 penalty across the 22 shared minidev splits, primarily due to headword approximation.

Finally, the underlying CorefUD datasets have undergone revisions, including updated coreference annotations and dependency trees. Fortunately, the vast majority of datasets remain functionally unchanged from the perspective of our text-based models. By cross-evaluating the raw CoNLL-U files between the two CorefUD versions, we estimate the upper bound of the score discrepancy caused by dataset changes to be 0.56. This figure conflates changes in both annotations and dependency trees.

Because dependency trees are completely opaque to our text-in/text-out LLM approach and are utilized only by the converter, we attempt to isolate their impact. We perform a similar cross-evaluation by first converting the organizer-provided text files back to CoNLL-U format before scoring. If this process eliminates the score gap for a given dataset, it confirms that the underlying changes were strictly confined to the dependency trees, meaning the theoretical performance ceiling for text-based LLMs remains unchanged. Applying this adjustment reduces the average cross-version discrepancy to a mere 0.16 points. This negligible difference strongly suggests that the intrinsic variance of our pipeline (e.g., model sampling, training run randomness, and converter lossiness) likely has a greater impact on score fluctuations than the cross-year dataset updates.

When filtering for only the shared datasets, our adjusted CoNLL-U score for this year is 75.30. Compared to our official score of 59.84 from last year, this represents a substantial +15.46 point uplift (and a +5.27 improvement over our unofficial post-deadline score of 70.03, which included late optimizations). By contrast, the leading specialized

Participant	avg.	ca_ancora	cs_pcedt	cs_pdt	de_potsdamcc	en_gum	es_ancora	fr_democrat	hu_szegedkoref	lt_lcc	pl_pcc	ru_rucor	hu_korkor	no_bokmaalnarc
corpipe-ensemb	**77.11**	**84.42**	**79.34**	**81.81**	**75.0**	**77.42**	**85.28**	73.38	**72.61**	**76.13**	**82.07**	**86.21**	**68.72**	78.9
corpipe-single	76.18	83.04	78.83	81.59	74.79	76.73	84.23	73.8	71.42	75.18	81.54	84.51	68.23	77.44
Antoine Bourgois	74.32	77.02	72.79	76.31	71.33	76.9	78.3	**74.46**	66.6	65.35	78.35	82.62	65.47	**81.19**
LLM-UWB	73.83	82.67	75.83	80.03	73.35	75.94	83.07	66.66	69.0	73.47	80.06	84.65	65.29	80.44
corpipe-sing-lg	72.32	79.65	73.85	77.93	69.92	73.48	82.3	70.98	67.71	75.71	78.04	82.15	64.56	74.42
portnlp	68.69	73.68	71.55	74.1	67.94	70.6	75.36	53.4	62.53	68.08	76.85	81.24	59.15	72.13
thmorton	67.56	76.77	70.76	74.61	71.92	68.63	77.39	68.99	60.96	66.49	73.72	79.12	60.52	72.88
stanza-coref	67.0	77.73	74.0	76.4	70.36	72.69	80.14	57.1	66.87	73.0	73.68	80.38	59.94	72.84
crac_baseline	54.54	67.91	63.36	66.21	52.36	61.88	70.26	55.6	54.29	62.42	67.46	68.23	42.24	61.35
pavlk-mm	46.19	41.32	36.66	40.31	49.26	55.3	47.27	20.72	37.33	52.75	43.17	52.29	41.56	54.55
AU-KBC	35.24	38.8	25.77	36.77	29.77	44.23	37.35	35.5	31.65	27.97	36.25	33.72	29.51	42.93

Participant	no_nynorsknarc	tr_itcc	cu_proiel	en_litbank	grc_proiel	hbo_ptnk	fr_ancor	hi_hdtb	ko_ecmt	fr_litbankfr	cs_pdtsc	en_fantasycoref	la_coreflat	nl_openboek
corpipe-ensemb	76.83	73.55	**68.75**	85.33	**79.01**	74.46	76.49	**78.36**	**70.29**	**82.46**	**76.7**	**81.12**	**62.63**	74.72
corpipe-single	76.44	**74.01**	67.86	83.68	77.87	72.0	75.45	77.83	69.72	81.48	76.66	80.75	58.69	73.1
Antoine Bourgois	77.59	73.09	60.09	**85.4**	74.87	**79.83**	77.33	76.84	68.84	80.21	70.89	78.79	58.66	**77.42**
LLM-UWB	**79.45**	69.09	59.58	84.58	75.28	76.8	**77.51**	77.05	66.89	60.47	73.85	80.2	56.19	66.14
corpipe-sing-lg	73.96	69.52	56.27	79.38	69.05	66.39	71.51	76.31	68.58	76.52	72.3	74.92	57.46	69.88
portnlp	72.07	62.97	57.94	78.37	71.08	72.72	70.55	75.61	69.48	54.94	69.77	74.8	44.79	72.93
thmorton	72.02	42.92	57.2	73.62	65.2	57.9	70.18	72.82	61.07	66.92	68.22	72.99	55.95	64.45
stanza-coref	70.81	60.93	38.92	73.32	53.86	61.24	69.55	75.45	67.13	64.52	71.22	69.9	36.86	60.03
crac_baseline	61.09	46.76	24.68	66.17	30.63	31.7	61.82	66.6	64.97	46.07	66.06	65.14	06.8	40.57
pavlk-mm	53.49	39.77	29.35	63.0	43.43	61.4	46.42	60.6	59.53	46.66	40.93	58.07	32.77	39.26
AU-KBC	39.65	37.21	34.8	34.41	43.36	48.19	37.38	46.19	21.62	28.55	37.91	41.35	16.17	34.51

Table 1: The official results taken from the shared task evaluation set leaderboard. All numeric values are primary CoNLL-U score (Novák et al., 2024; Žabokrtský et al., 2022) (without singletons). Rows are sorted by average score. Our entry is named LLM-UWB. The best score for each dataset is bold, and participants of the LLM track are in bold.

system, *corpipe-ensemb*, saw a marginal gain of only +0.96, moving from 76.51 to 77.47. This underscores the rapid, ongoing performance scaling of generalized generative models on this task.

4 Conclusion

In this paper, we presented our submission to the CRAC 2026 Shared Task. By fine-tuning instruction-tuned language models (primarily Gemma 3 27B) via QLoRA, we achieved highly competitive results, placing second in the LLM track and third overall (73.83 CoNLL-U score), and securing the best performance on two datasets out of the 27 evaluated. These results align with a broader trend: generative models are rapidly approaching the capabilities of long-dominant, specialized architectures like *corpipe* across a diverse array of languages and annotation styles, although at a higher cost.

Our findings indicate that the primary barrier to surpassing specialized systems is no longer the raw linguistic capability of the models or their inability to generate coherent long-form annotations, but rather the computation required, increased size, and the logistical challenge of handling particularly long documents. While our primary full-document approach yielded excellent results on most corpora, extreme sequence lengths forced us to fallback on a 2,000-token sliding context window powered by a significantly smaller Llama 3.1 8B model for five specific datasets. Although this sliding window successfully bypassed memory limitations and elegantly solved the overlap reconciliation problem by reusing cleaned past generations, the reduced ability of the 8B model noticeably penalized our overall average score.

Acknowledgments

This work has been supported by the Grant no. SGS-2025-022 - New Data Processing Methods in

Current Areas of Computer Science.

Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

Google Gemini was used as assistance with the literature search and presentation of this paper. All aspects were checked and thoroughly edited by hand.

References

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *Preprint*, arXiv:2305.14314.

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

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Daniel Han, Michael Han, and Unsloth team. 2023. Unsloth.

Jakub Hejman, Ondrej Prazak, and Miloslav Konopík. 2025. Fine-tuned llama for multilingual text-to-text coreference resolution. In *Proceedings of the Eighth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 140–148, Suzhou, China. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

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.

Michal Novák, Miloslav Konopik, Anna Nedoluzhko, Martin Popel, Ondrej Prazak, Jakub Sido, Milan Straka, Zdeněk Žabokrtský, and Daniel Zeman. 2026a. Findings of the fifth shared task on multilingual coreference resolution: Expanding datasets for long-range entities. In *Proceedings of the 2nd Joint Workshop on Computational Approaches to Discourse, Context and Document-Level Inferences and Computational Models of Reference, Anaphora and Coreference (CODI-CRAC 2026)*, San Diego, California, USA. Association for Computational Linguistics.

Michal Novák, Martin Popel, Daniel Zeman, Zdeněk Žabokrtský, Anna Nedoluzhko, Kutay Acar, David Bamman, Antoine Bourgois, Peter Bourgonje, Silvie Cinková, Eleonora Delfino, Hanne Eckhoff, Gülşen Cebiroğlu Eryiğit, Jan Hajič, Sooyoun Han, Christian Hardmeier, Dag Haug, Tollef Jørgensen, Andre Kåsen, and 35 others. 2026b. Coreference in universal dependencies 1.4 (CorefUD 1.4). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Milan Straka. 2025. CorPipe at CRAC 2025: Evaluating multilingual encoders for multilingual coreference resolution. In *Proceedings of the Eighth Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 130–139, Suzhou, China. 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.